Pitch of Voiced Speech in the Short-Time Fourier Transform

Algorithms, Ground Truths, and Evaluation Methods

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titel eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
eingereichte Dissertation

von
Bastian Bechtold
geboren am 8. November 1985
in München, Deutschland

Erstbetreuer:
Prof. Dr. Steven van de Par
Department of Medical Physics and Acoustics
Carl von Ossietzky Universität Oldenburg

Summary

Speech is fundamental to humanity as our primary means of communication and expression. Speech sounds are produced by a vibration of our vocal chords or a constriction in the vocal tract, and are shaped by resonances in the vocal cavities, and expelled through our nose and mouth. By rapidly changing the configuration of our vocal organs, we can mold such sounds into language, to be transmitted through the air, and heard by human ears.

When this sound is produced by a vibration of the vocal chords, its waveform becomes periodic, and its spectrum harmonic. We perceive such a "voiced" sound as having a *pitch* that corresponds to the frequency of the vocal chord vibration and the harmonic spacing of the spectrum. This quantity can be estimated by computer algorithms, and is then typically referred to as a *fundamental frequency*.

Fundamental frequency estimation algorithms are a key ingredient for various speech analysis tasks such as speech recognition, speaker identification, and speech compression. This dissertation is about the construction of such algorithms, how to evaluate their performance, and a large comparison of common implementations.

The first major contribution of this dissertation is a new algorithm for estimating the fundamental frequency of speech. The algorithm combines features from multiple domains into a probabilistic pitch confidence measure that evaluates the probability of a short audio segment having a certain fundamental frequency. The measure is unusual in that it is a true probability that can both accept and reject each candidate frequency instead of merely finding the most probable one. Its estimates are thus more sparse than similar algorithms', but also more robust. These characteristics are validated in a large evaluation with speech and noise recordings and in comparison with a number of notable reference algorithms.

However, the evaluation brought to light an idiosyncrasy of speech analysis, that the only truth of speech properties is often human perception. These being unavailable to computer programs, evaluations of their accuracy must rely on some other form of truth, which is necessarily flawed. To investigate this, we conducted a study of numerous speech databases and their fundamental frequency ground truths, and found them unsatisfyingly variant and inconsistent.

Based on this notion, the second major contribution is then a new ground truth measure for fundamental frequency, constructed from the consensus of a number of existing fundamental frequency estimation algorithms. In contrast to existing truths, ours does not rely on the estimates and biases of a single algorithm, nor on laryngograph recordings, as these were found to be problematic for evaluating fundamental frequency algorithms. Our ground truth was validated to be very similar to existing ground truths, but more suitable to the task of evaluating the accuracy of algorithms in difficult edge cases.

Thirdly, a comparison study of unprecedented depth was conducted of not only fundamental frequency estimation algorithms, but also speech and noise corpora, as well as ground truths. In preparation for this comparison study, a uniquely large and reproducible dataset of algorithms, signals, truths, and performance measures was constructed, which can be of independent value to future researchers and is available on this dissertation's companion website.

The comparison itself investigated the characteristics of 25 fundamental frequency estimation algorithms from the last 30 years of digital signal processing history in unique detail. This comparison revealed a number of previously unknown properties of all included algorithms, particularly in their biases towards certain speech corpora and performance measures.

In summary, this dissertation examined the algorithmic estimation of the fundamental frequency of speech. Analogous to human perception, pitch estimates are often inherently ambiguous and since therefore a definition of "true" pitch does not exist, a consensus approach was proposed. Consequently, this topic of research has a rich history, and its algorithms are now as intricate and interesting as speech itself.

Zusammenfassung

Sprache ist die Grundlage aller zwischenmenschlicher Kommunikation und Ausdrucks. Wir erzeugen Sprachlaute durch Vibration unserer Stimmbänder oder eine Verengung des Luftstroms im Rachenraum, welche im Vokaltrakt zur Resonanz gebracht und durch Mund und Nase abstrahlt werden. Indem wir unsere Sprachorgane schnell und präzise bewegen, erzeugen wir so Sprache, die durch die Luft in die Ohren anderer Menschen transportiert wird.

Wenn diese Laute durch eine Vibration der Stimmbänder erzeugt werden, wird das Signal periodisch, und sein Spektrum harmonisch. Einen solchen „stimmhaften“ Laut empfinden wir mit einer spezifischen *Tonhöhe*, welche der Frequenz der Stimmbandschwingung und dem Abstand der Harmonischen entspricht. Diese Frequenz lässt sich auch algorithmisch mit Computerprogrammen bestimmen, und wird dann *Grundfrequenz* genannt.

Grundfrequenzschätzungs-Algorithmen sind ein wichtiger Bestandteil vieler Sprachanalyse-Werkzeuge, wie der Spracherkennung, der Sprechererkennung, oder der Sprachkompression. Diese Dissertation handelt von eben solchen Algorithmen, Auswertungsmethoden ihrer Genauigkeit, und einer Vergleichsstudie verschiedener Implementierungen.

Der erste wichtige Beitrag dieser Dissertation ist ein neuer Grundfrequenzschätzungs-Algorithms. Er kombiniert Merkmale verschiedener Signaldarstellungen in einer Wahrscheinlichkeit, ob ein kurzer Signalausschnitt an einer bestimmten Frequenz stimmhaft ist. Dieses Maß ist ungewöhnlich, da es als echte Wahrscheinlichkeit die Stimmhaftigkeit sowohl bestätigen als auch ablehnen kann, anstatt lediglich die wahrscheinlichste Frequenz anzugeben. Unsere Frequenz-Schätzungen sind dementsprechend konservativer als die anderer Algorithmen, aber auch deutlich robuster. Diese Eigenschaften wurden in einer großen Studie mit Sprach- und Störgeräuschaufnahmen validiert, und mit bekannten Referenzalgorithmen verglichen.

Die Studie legte allerdings ein Grundproblem der Sprachanalyse offen: dass die einzige *Wahrheit* der Sprache nur in der Wahrnehmung der Menschen zu finden ist, und diese leider für Computerprogramme nicht verfügbar ist. Statt dessen müssen Vergleiche zwangsläufig auf eine andere Art der Wahrheit ausweichen, und deren Kompromisse in Kauf nehmen. Um dies zu untersuchen, haben wir eine Vergleichsstudie verschiedener Sprachdatenbanken und deren Grundfrequenz-Wahrheiten durchgeführt, die relevante Inkonsistenzen und Unterschiede zu Tage brachte.

Der zweite wichtige Beitrag dieser Dissertation ist eine neue Grundfrequenz-Wahrheit, die wir aus einer mehrheits-Wahl verschiedener Grundfrequenzschätzungs-Algorithmen erzeugten. Im Gegensatz zu bestehenden Wahrheiten muss unsere Wahrheit weder auf die Schätzungen und Eigenheiten einzelner Algorithmen zurückgreifen, noch auch auf Laryngograph-Messungen, da diese sich als problematisch für die Evaluation von Grunfrequenzschätzungs-Algorithmen herausstellten. Unsere mehrheits-Wahrheit ist bestehenden Wahrheiten sehr ähnlich, allerdings für die Bewertung von Grundfrequenzschätzungs-Algorithmen besonders in schwierigen Randbereichen besser geeignet.

Als drittes bereitet diese Dissertation eine einmalig große Vergleichsstudie vor, von Grundfrequenzschätzungs-Algorithmen, aber auch von Sprach- und Störgeräuschdatenbanken und Wahrheiten. In der Vorbereitung dieser Studie entstand eine einzigartig große Datenbank von Algorithmen, Signalen, Wahrheiten und Bewertungsmaßen, die auch jenseits des eigentlichen Vergleichs für zukünftige Wissenschaftler auf unserer Webseite zur Verfügung gestellt wird.

Die eigentliche Vergleichsstudie umspannt 25 Grundfrequenzschätzungs-Algorithmen der letzten dreißig Jahre in nie dagewesener Detailtiefe. Dieser große Vergleich zeigt bei jedem der Algorithmen neue Eigenschaften, und ganz besonders neue Biases für bestimmte Signalzustände, Sprachdatenbanken, und Bewertungsmaße.

Zusammenfassend untersucht diese Dissertation die algorithmische Schätzung der Grundfrequenz von Sprache. Und wie in unserer menschlichen Wahrnehmung, so ist auch die Schätzungen oft fundamental mehrdeutig. Da damit keine eindeutige „Wahrheit“ existieren kann, wurde eine mehrheits-

Entscheidung vorgestellt. Dennoch hat dieses Fachgebiet eine lange Geschichte, und seine Algorithmen sind inzwischen genauso spannend und komplex wie die Sprache selbst.

Contents

List of Figures

List of Tables

Glossary

ASR Automatic Speech Recognition

CQT Constant-Q Transform

DNN Deep Neural Network

F0 Fundamental Frequency, or Pitch

FFE Full Frame Error

FFT Fast Fourier Transform

FPE Fine Pitch Error

FRB Fine Remaining Bias

GPE Gross Pitch Error

GRE Gross Remaining Error

IF Instantaneous Frequency

LPC Linear Predictive Coding

MAD Mean Absolute Deviation of the Mean

OPE High/Low Octave Pitch Error

PDA Pitch Determination Algorithm

RMS Root Mean Square

RNN Recursive Neural Network

SNR Signal-to-Noise Ratio

STFT Short-Time Fourier Transform

VAD Voice Activity Determination

VDE Voicing Decision Error

Part I

Introduction

Where we introduce the topic of speech analysis in general, and fundamental frequency estimation in particular, and define the scope of this dissertation.

Chapter 1 introduces speech analysis as a general topic, and the fundamental frequency of speech as one of its foundations, along with the major concepts necessary for understanding their purpose and applications.

Chapter 2 goes into deeper detail on how voiced speech is produced, and how it gives rise to a perception of a pitch. Depending on these points of view, different definitions for the pitch of speech are brought forth, which form the various bases for the fundamental frequency estimation algorithms and ground truths in the remainder of this dissertation.

Chapter 3 reins in the scope of this dissertation from the infinite varieties and intricacies of speech in general to a more compact subset of utterances available in published speech databases and to algorithmic evaluation.

The conclusion of the introduction in Chapter 4 explicitly raises the main questions this dissertation is addressing, and summarizes its contributions.

If there is one capability that is uniquely human, it is that of language. Language allows humanity to cooperate at otherwise impossible scales, to communicate knowledge over generations and across continents. It is the foundation of all human achievement. As such, it is also of intrinsic interest to scientists, both to better understand ourselves, and to enable machines to understand us.

If language is the medium of information, then speech and writing are its encoding for transmission. From just a few weeks after birth, humans long to communicate by making sounds [171, ch. 5]. First, to connect with their parents and siblings, then to communicate emotions and desires, and later to learn and teach others about the world. As young humans mature, so does their command of language, speech, and writing. Yet for most people, *speech* remains the richest encoding possible, with writing only a pale imitation, and seemingly lacking in nuance and expression.

Speech is produced from air streaming out of the lungs, which is disturbed by a constriction in the vocal cavities, or by a periodic opening and closing of the vocal chords. The ensuing sound waves resonate in the vocal cavities, and are radiated through the nose and mouth. Each of these parts, the flow rate of the air, the place and type of constriction, the frequency of the vocal chord vibration, the shape and resonant frequencies of the cavities, and the shape of the orifice, can be modulated several times per second to produce the varied sounds of speech [131, 35, 90].

On the perception side, an equally complex array of auditory transducers and neural processing stages translate the resulting audio signal back into language. There is some evidence that we perceive speech differently than other sounds, with our perception guided by an intimate knowledge of speech production [98]. Technological analysis of speech can make use of such knowledge as well, and model speech not just as an arbitrary acoustic phenomenon, but as a physiologically constrained process with a limited number of variables.

One of these variables, and the topic of this dissertation, is the *fundamental frequency* of voiced speech, where a periodic vibration of the vocal chords gives rise to a harmonic and periodic signal that we perceive as having a single *pitch*. While the pitch of voiced speech does not carry much vocabulary information in western languages, it is an important side-channel of prosodic information, such as for communicating emotional state, sentence boundaries, and emphasis [44].

Beyond the linguistic meaning of pitch, however, it gives voiced speech a particular comb-like spectral shape that is recognizable even when severely distorted by background noise [90, ch. 4.2]. This shape aids humans in auditory scene analysis, and machines in a number of different applications, such as voice activity detection, speaker identification, speech recognition, separation, enhancement, compression, and modification [21].

The analysis of speech and its pitch are thus a prerequisite for using speech to interact with machines. If language is what defines humanity, it is machines that enable us to step beyond the limitations of our bodies: Machines enable humans to communicate instantly across large distances, to manipulate our surroundings beyond our own physical strength, and to process and analyze data too complex for single human brains. But robust use of synthetic speech generation and speech recognition has long been unavailable for machine interaction.

Thus, voiced speech and its pitch are the essence of human phonation, and its analysis is at a unique crossroads between the technological and the humane. How to bridge this gap between algorithms and perception is what makes speech fascinating, and is the motivation for this dissertation.

The remainder of the Introduction is structured as follows: Chapter 1 briefly introduces the history and current applications of scientific speech analysis. Chapter 2 introduces the required nomenclature of speech production and perception, as well as a glimpse into the worlds of psychoacoustics and audiology, each with their own definitions of pitch. After that, Chapter 3 defines the scope and properties of the specific kinds of speech signals to be discussed in this dissertation. Finally, the conclusion in Chapter 4 will end the introduction with a few notes on applications of speech analysis and a juxtaposition of the technological *fundamental frequency* and perceptual *pitch*.

Chapter 1

Speech Analysis and Pitch Analysis

For millennia, humans have successfully communicated without a concrete theory of how speech signals work. The notion of a *signal* as a modulated sound wave traveling through the air is claimed to originate from the ancient Greeks, with Aristotle explaining sound as:

> Sound takes place when bodies strike the air, [...] by its being moved in a corresponding manner; the air being contracted and expanded and overtaken, and again struck by the impulses of the breath and the strings, for when air falls upon and strikes the air which is next to it, the air is carried forward with an impetus, and that which is contiguous to the first is carried onward; so that the same voice spreads every way as far as the motion of the air takes place.
>
> –Aristotle (384-322 BC), *Treatise on Sound and Hearing*

Yet it would take two millennia until the invention of the microphone brought a new, electric representation of a signal, not as intangible vibrations of the air, but as measurable voltages on a wire [62]. And what can be measured, can be analyzed. Electronic circuits were quickly invented to modify sound recordings, and transmit them over long distances. The invention of magnetic tape and vinyl records could make electronic recordings permanent, and replay them (almost) infinitely without any degradation in quality [48].

Only a few years later, during the second half of the twentieth century, did the invention of digital computers again reinterpret these voltages as digital series of numbers, and started the field of digital signal processing as it is known today [68]. Suddenly, sound recordings were no longer bound to their electric origin, but a subject of intense mathematical study and manipulation.

The first, and most pressing application of this new understanding, was to increase the efficiency of transmitting human speech over wires and radio. As early as 1850, undersea cables were used to transmit telegraph messages across first the Rhine river, then the English Channel, and in 1858, the Atlantic Ocean. However, text telegraphs are a poor substitute for human speech, and in 1927, a transatlantic commercial radio telephone service opened, costing an astounding £9 per three minutes (equivalent to $550 in 2010). In 1956, this was replaced with the first transatlantic undersea cable, capable of 36 simultaneous telephone conversations at a time. This cable carried more phone calls in its first few days of service than the previous radio telephone had had in a year [62, 55]. Improving the throughput of such long-distance telephone wires, and reducing the number of cables in the budding national telephone networks were the first challenges for digital speech processing in the advent of the information age.

According to *Speech Analysis and Synthesis and Perception* by Flanagan from 1965 [35], the telephone spread quickly, with 150 million telephones in use by 1965. But human speech signals on

a telephone circuit required a communication channel of 3 kHz, or 30 kBit/s, whereas the linguistic content was thought to be no more than 50 bit/s [35, ch. 1]. Thus a massive reduction in bandwidth was assumed possible, so to ease the burden on connecting all of humanity through telephones.

Early speech compression techniques were based on the idea that speech recordings could be disentangled into an excitation signal analogous to the parameters of the human glottis, and filter parameters that could recreate the effect of the human vocal tract. Transmitting these parameters instead of the entire waveform would save significant bandwidth at the cost of being engineered for speech signals only.

On the receiving end, these parameters would be fed into a matching speech synthesizer, which would recreate the original speech. At the time, these analysis-synthesis systems were still in their infancy [118] and not yet in widespread use.

While these systems would remain somewhat of a curiosity for civilian use, the military realized their potential early on as a means for encrypting signal transmissions digitally as early as 1943 [7]. In this use case, the purpose of the analysis-synthesis system was not to save bandwidth, but to use the existing telephone bandwidth securely.

One of the defining parameters of these systems was the fundamental frequency of voiced sounds. This was used either directly to characterize the excitation signal, or as an intermediary step for finding spectral peaks and encoding prosodic information. In 1983, these developments were described in the book *Pitch Determination of Speech Signals* [59] by Hess, where the author detailed the advances in fundamental frequency estimation in terms of digital signal processing, as well as its analog prehistory. The developments in digital pitch determination since then are investigated in great detail in Chapter 11. Suffice it to say that approaches are as varied as speech itself, and provide a fascinating cross-section of speech analysis as a whole.

By 1992, speech coding technologies had advanced far enough to be no longer of academic interest only, but a reality in the design and implementation of the integrated services digital network (ISDN), with its myriad civil and military applications. These applications and their challenges are described in *Digital Speech Processing, Speech Coding, Synthesis and Recognition* by Ince in 1992 [63].

The digitization of speech also paved the way for speech analysis applications beyond the purposes of human-to-human communication. Their use expanded significantly into interactions between humans and machines, in the form of speech recognition and speech synthesis. Similarly, the demands on speech coding increased from mere transmission to providing increased security and reliability and memory-efficient voice storage. These issues only grew in significance with the shared transmission channel between voice and data that would be common for ISDN and beyond.

Ince illustrated applications of these technologies in a combat aircraft environment: voice recognition systems in the cockpit help control tactical systems, efficient voice coding and speech enhancement systems are used for communicating between aircraft, and synthetic-speech aural warning systems provide feedback from sensors. Speech-driven man-machine interactions proved preferable to visual displays in this case, as they were found to be perceived independently from visual input or motoric tasks. Indeed, these systems were deemed absolutely necessary for maintaining combat readiness in the face of the enormity of data available in modern cockpits, a lesson that would take a few decades to be incorporated into civilians' lives as well.

Nowadays, these technologies have trickled down into smart phones, allowing for sending and receiving of text messages hands-free while driving a car, digital assistants for controlling media hubs, automatic transcription and translation of online videos, and low-bandwidth digital voice transmissions. In fact, today's speech coding methods such as MELP [96] (300 bit/s) or Codec2 [133] (700 bit/s) are approaching the bandwidth of 50 bit/s Flanagan predicted in 1965 as mentioned at the beginning of this chapter.

Later publications on speech analysis added additional applications such as speaker recognition and language recognition [43], computational auditory scene analysis [160], speech modification and music

information retrieval [21], and deeper speech analysis such as linguistic analysis, voice transformation and speech enhancement [6], which have not yet found their way into everyday life.

All of these technologies rely on models and parametrizations of speech signals to conduct their higher-level analyses. It thus comes as no surprise that the estimation of these parameters is still an active area of research as well. Fundamental frequency estimation, in particular, has seen constant developments from the dawn of digital technologies until today.

Algorithms to estimate the fundamental frequency of speech have evolved with the technological possibilities of their time, from the computationally constrained event detection systems of the 1960s and 1970s, to short-time periodicity and harmonicity estimators in the 1980s and 1990s, to today's machine learning tools. Every step of their evolution enabled new capabilities, beginning with basic accuracy for single recordings, to near-human levels of speaker-agnostic noise resistance in today's most powerful algorithms, such as the one introduced later in this dissertation.

These improvements in fundamental frequency estimation algorithms required a similar improvement in evaluation methods and ground truths to validate their claims of accuracy. This has been partly realized in ever larger speech databases with associated fundamental frequency ground truths for evaluating algorithm accuracy.

Thus, there is still room for improvement in fundamental frequency estimation, as well as speech analysis in general. Research has shown without a doubt that speech remains the most convenient and most effective means of communication between humans [107]. The academic interest in speech and its parameters has persisted, and will no doubt continue to do so for the foreseeable future, until machines can truly understand us as well as we can each other.

Chapter 2

Speech Production, Perception, and Nomenclature

Speech is humanity's primary means of encoding language. It is produced by the human vocal organs, transmitted through the air, and received by a human ear and auditory system. While other methods of producing or consuming speech-like signals exist, these are mere facsimiles of human speech and must adhere to the same rules to be considered speech. Thus, both the production and the consumption of speech is a deeply human affair, and can only truly be interpreted by humans.

Yet, we build technological systems that aim to produce speech-like signals and consume human speech, "just like humans". For these applications, we need to understand the structure of speech signals, both in their physical and technical specifications, and also in the meaning they encode for humans. This work is primarily concerned with one important aspect of the former, the fundamental frequency of speech from a signal processing standpoint, in order to allow for a richer machine interpretation of human speech.

To gain a technical understanding of speech and its pitch, we must understand how it is produced and how it is perceived. Thus, the next sections examine speech production, its auditory perception, and finally its signal properties. These sections only define terms with respect to speech analysis and offer a brief venture into the worlds of linguistics, phonetics, audiology, and psychoacoustics where necessary.

2.1 Speech Production

Speech is produced in the human vocal organs, a simplified drawing of which is shown in Figure 2.1. Starting from the bottom to the top, the lungs supply the vocal tract with air and energy, which is excited by a constriction of the airflow or a rhythmic opening and closing of vocal folds, and made to resonate in the pharynx, nasal, and oral cavities, and exit through the nose and/or mouth.

The first obstacle for the lung's airflow are the vocal folds, two membranes stretched across the larynx. The vocal folds can be controlled both in the size of their opening and in their tension, all of which interact in various ways with the air flow: if completely open, such as in breathing or with certain consonants, they do not disturb the air flow at all and play no role in the resulting speech signal. If partly closed, they vibrate in the airflow, and produce a *voice*. Depending on the size of the opening, the tension of the vocal folds, and the air pressure from the lungs, this vibration might periodically stop the airflow, or merely constrict it. This constitutes the difference between the various kinds of voice, such as whispering, breathy voice, normal voice, or shouting voice. At the other extreme, the vocal folds can close completely, which then produces glottal stop sounds such as a /g/ or /k/.

In the normal voice, the vocal folds open and close to produce a roughly triangular air flow over

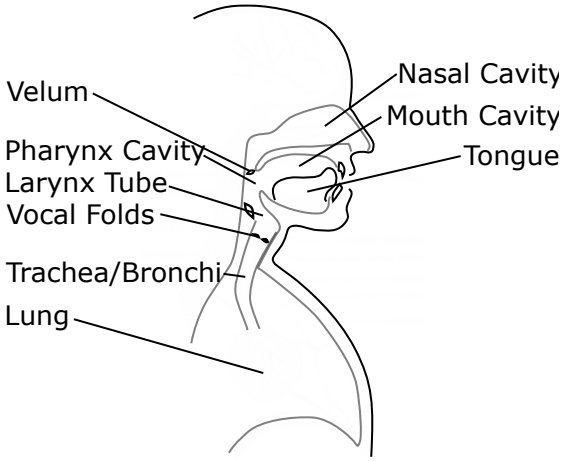


Figure 2.1: Drawing of the human vocal organs, as they pertain to speech production [35].

time, called a *glottis pulse*, with complete closures in between pulses [35]. The frequency and length of these pulses can be controlled to produce a "buzzing" sound at various frequencies, with a harmonic spectrum that diminishes at about 12 dB/octave [131, 59]. The precise shape of the glottis pulses is part of the unique difference between different people's voices and only partly controllable by the speaker.

A periodic opening and closing of the vocal folds thus introduces a first definition of speech pitch. From a speech production viewpoint, the pitch of voiced speech is the frequency of the glottis pulses. These can be measured by a device called a laryngograph or electroglottograph, which estimates the vocal fold contact area by measuring the electrical impedance across the larynx with two electrodes [59, ch. 5.2.3]. These electrical measurements are easily disturbed by neck movements, however, and can exhibit anomalies during phoneme transitions and mixed phonation such as /w/ or /z/, which are produced both by vocal fold vibrations and an unvoiced constriction in the airflow [132].

Not all speech sounds originate in the vocal folds, however. Some sounds are produced in part or entirely by turbulent air flow caused by constrictions in another part of the vocal tract. This includes constrictions with the lips or tongue for consonants such as /f/ or /s/, but also sudden releases of air pressure in the mouth, such as /p/ or /t/. These sounds have a non-harmonic spectrum that is much broader than the spectrum of glottis pulses [59].

The glottis pulses or noisy excitations then pass through the various cavities of the vocal tract, such as the larynx, pharynx, mouth, and nasal cavities. As these are volumes of varying size and shape, they excite resonances, whose characteristic spectral peaks are called *formants* that transform the "buzzing" or "whooshing" excitation to speech sounds [59, 6, 131]. Figure 2.2 shows an example of how a vowel sound is produced by shaping the spectrum of a glottis-pulse train by vocal tract formants.

During speech production, the resonances are varied by moving the tongue, lips, cheeks, jaw, and larynx, and by opening or closing the velum to connect or disconnect the nasal cavity. Other resonances are characteristic to each person, and cannot be varied voluntarily, such as the size and shape of the larynx, pharynx, and nasal cavity [6].

Finally, the openings of the mouth and nose radiate the speech sounds into the environment, with an additional modification of the sound depending on the mouth opening and head shape.

The resulting waveform for voiced sounds remains approximately periodic, with the same period as the glottis pulses. A second definition of pitch is thus the periodicity of the speech signal. This is different from the aforementioned laryngograph measurements in that the vocal tract resonances take

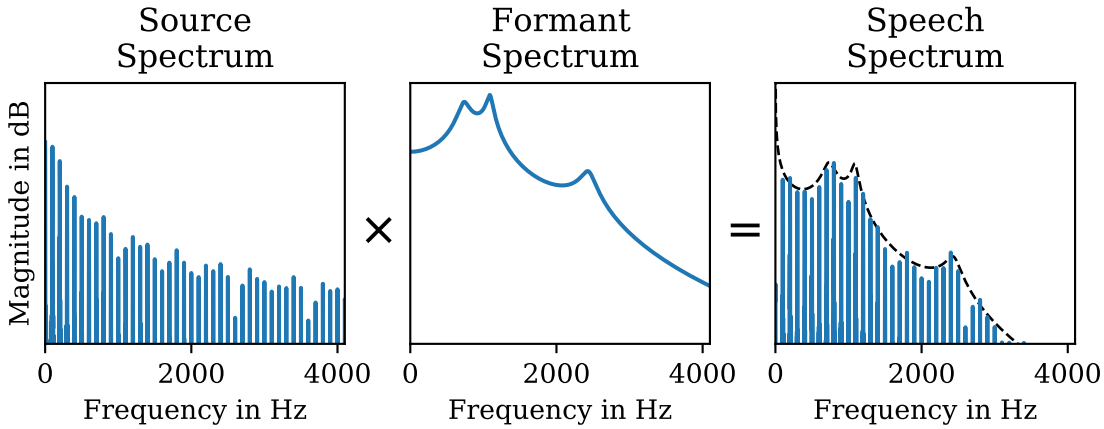


Figure 2.2: Illustration of the creation of voiced speech: A series of glottis pulses with a roughly 12 dB/octave spectrum is filtered by the vocal tract with formants at 730 Hz, 1090 Hz, and 2440 Hz (a male vowel /a/) to form a speech spectrum. Dashed line is the spectral envelope.

a short time to fully build up, which can obscure periodicity changes, particularly during onsets and offsets. Additionally, mixed excitation with both glottis pulses and an unvoiced constriction markedly reduces periodicity of voiced speech and can complicate its algorithmic estimation.

2.2 Speech Perception

Speech is produced and perceived by humans. Its function is to transmit information, and it is specifically adapted to humans' vocal organs and auditory system as speech is undoubtedly constrained by the limitations of human perception. We cannot perceive what we cannot hear, and we are unlikely to say anything that cannot be heard. Understanding the characteristics of human auditory perception is thus important for understanding speech signals.

Speech signals enter our ears as sound waves: rapid oscillations in the ambient air pressure that were produced by a human speaker (or a facsimile of one). These pressure waves are guided by the external ear into the ear canal, where they excite the eardrum, a membrane that separates the ear canal from the middle ear. This membrane converts the air's vibrations into physical vibrations of the ossicles, which are connected to the fluids of the inner ear. In the inner ear, the vibrations travel along the helicoid tube of the cochlea and resonate at a frequency-dependent point on the basilar membrane, where they excite hair cells to produce electrical signals on the auditory nerve. A drawing of these organs is shown in Figure 2.3.

During this process, various passive and active systems amplify the intensity of the vibrations and sharpen the frequency selectivity of the cochlea, so as to present as clearly separated frequencies to the auditory nerve as possible. These nerve signals are produced in a phase-locked manner with the vibration itself, always firing around a fixed moment in the full cycle of oscillation, although not necessarily for every wave period. The signals are only synchronized up to roughly 4000 Hz, which preserves some amount of phase information for later processing stages. Intensity information is likewise preserved in the density of nerve firings. Information about the frequency content of the vibration is encoded spatially, as each hair cell and nerve fiber is tuned only to a small bandwidth of frequencies [98, 6].

The inner ear implements a physical filter bank, where each frequency excites not just a single place, but a range of places on the basilar membrane. The resulting operational frequency resolution

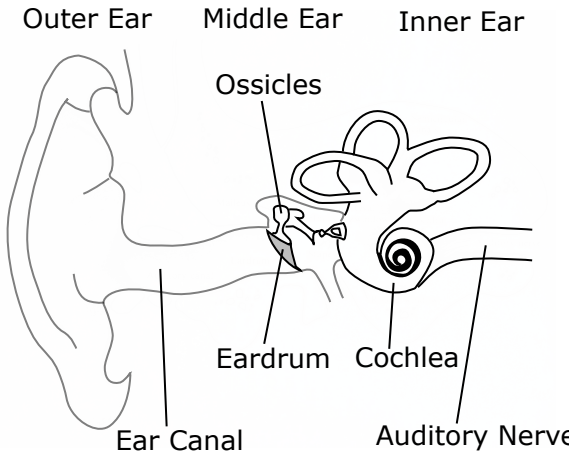


Figure 2.3: Drawing of the human hearing organs [98].

is referred to as a *critical band* [98]. Thus, two very similar frequencies presented simultaneously may not be distinguishable, as they excite the very same place on the basilar membrane, even though we can hear their difference in pitch if presented separately. Similar masking effects can be seen in other places as well, where dominant stimuli can overpower and suppress weaker signal components if presented in close proximity in time or frequency.

At this point in the auditory system, a voiced speech signal is represented by nerve firings at multiple harmonically related places on the basilar membrane, mostly phase-locked to one another. This provides the first perception-oriented definition of pitch as the common fundamental of multiple harmonics. Since harmonics only develop over multiple periods, this measure has a lower time resolution than the production-based definitions, and can thus be ambiguous during rapid transitions.

From the inner ear, nerve signals travel through various neural structures, where the signals from both ears are combined and pre-processed, until they finally enter the auditory cortex for language processing and integration with the rest of the cognitive functions of the brain. Along the way, the raw audio information is integrated into higher-level features such as onsets, sweeping, duration, repetition, timbre, pitch, loudness, and localization [35, 98].

In general, most percepts scale approximately logarithmically with physical stimulus parameters such as sound intensity and fundamental frequency[1]. For example, signal intensity needs to square in order for loudness to double. Similarly, an equal-interval progression of pitches is achieved when each pitch is multiplied by a fixed factor. Thus, perception-related measures for signal intensity are logarithmic decibels instead of linear pressures, and octaves for pitch instead of Hertz [98].

While the pitch perception of pure tones is easily explained by the excitation of a particular place on the basilar membrane, it is curious that we perceive a similar single pitch for tone complexes as well. Many a psychoacoustical experiment has been conducted to ascertain the exact mechanism of pitch perception. According to *An introduction to the Psychology of Hearing* by Moore [98], pitch is evoked by periodic or harmonic signals and is caused most strongly by near-harmonic spacing of partials in the center of the audible range, between 300 and 1000 Hz. Yet, clever experiments can induce a perception of pitch in many alternative ways, for example from single harmonics, binaural cues, phase changes only, or timing only. Furthermore, pitch perception shifts with loudness, envelope fluctuations, and interfering tones, even though partial spacing remains unchanged. Perhaps most tellingly, we can consciously switch between integrating partials into a tone complex and singling

[1]Also known as *Weber's Law*, or its "near miss" where it scales only approximately logarithmically [98].

them out as separate pitches [98, ch. 5]. Thus there are clearly multiple mechanisms for evoking a perception of pitch, some of which are basic signal patterns, others high-level interpretive actions. It is therefore perhaps prudent to speak of *pitch* only in human contexts, and relegate algorithmic discussions to the more rigorously defined concept of *fundamental frequency*[2].

The human perception of pitch is therefore its most complex definition. It can be measured in human experiments by comparing sinusoidal reference signals with a known fundamental frequency to isolated speech sounds, but such experiments are time-consuming and impractical for large datasets. As such, human pitch estimates are rarely available for designing and evaluating the kinds of fundamental frequency estimation algorithms discussed in this dissertation. They are highly relevant to other scientific fields, however, such as auditory modelling or prosodic research.

Corollary to loudness and pitch, there is *timbre*, which encompasses any difference between tones of equal loudness, duration, and pitch. Timbre can be described by crude categories such as roughness, clarity, or warmth, which encompass differences in spectral makeup and timing envelope. For example, timbre differentiates between otherwise similar sounds from a clarinet, a human voice, or the various differences between vowels. Speech signals, in particular, encode considerable information in timbre, which will be of great interest in the design of a fundamental frequency estimation algorithm later [98, ch. 7].

In the last processing stage of speech in the brain, all of this information is finally integrated into a holistic speech model. Research clearly shows that human brains process speech differently from other sounds, making use of contextual information such as grammatical and syntactic structure, language, physiological limitations, spectral changes, envelope changes, spectral content, and also visual information and knowledge about the speaker. All of this has been shown to be used for resolving ambiguities in speech signals [98, ch. 8]. Perhaps this immense knowledge base can explain how humans can identify and understand speech with astounding acuity, even if severely distorted or obscured.

2.3 Speech Properties

Without access to human auditory systems, algorithms need to rely on signal processing for making sense of speech signals. Some aspects of human perception are readily translated into algorithmic terms, while others remain elusive. On the other hand, algorithms are not constrained to the same limitations as human ears. They can easily trade time resolution against frequency resolution, and be supplied with high-resolution recordings with frequency and dynamic ranges beyond the capabilities of the human ear. For speech signals, however, these advantages are likely irrelevant, as there is no evolutionary incentive to develop intricacies of speech that cannot be perceived.

On a linguistic level, speech is made of sentences[3], which are composed of words, which contain syllables, which are constructed from phonemes. Phonemes are the atomic units of speech, from which any utterance can be constructed. English and most European languages use on the order of 40 phonemes, and normal speech can produce phonemes at a rate of about 10 phonemes per second [35].

Phonemes can be characterized either as relatively long and steady signals, like vowels, which are voiced signals with a distinct spectral signature of a fundamental frequency and strong formant peaks. Or they can be characterized by transitions, like many consonants. For example, some phonemes such as /p/, /b/, /t/, /d/, /k/, and /g/ block the air flow for a short while, and then open up rapidly for an onset of voiced or unvoiced sound. Consonants in particular are often strongly co-articulated, and exhibit different spectral shapes depending on preceding and succeeding phonemes [175].

[2]This will be discussed more thoroughly later.

[3]technically *clauses*, since we rarely use complete sentences in spoken speech.

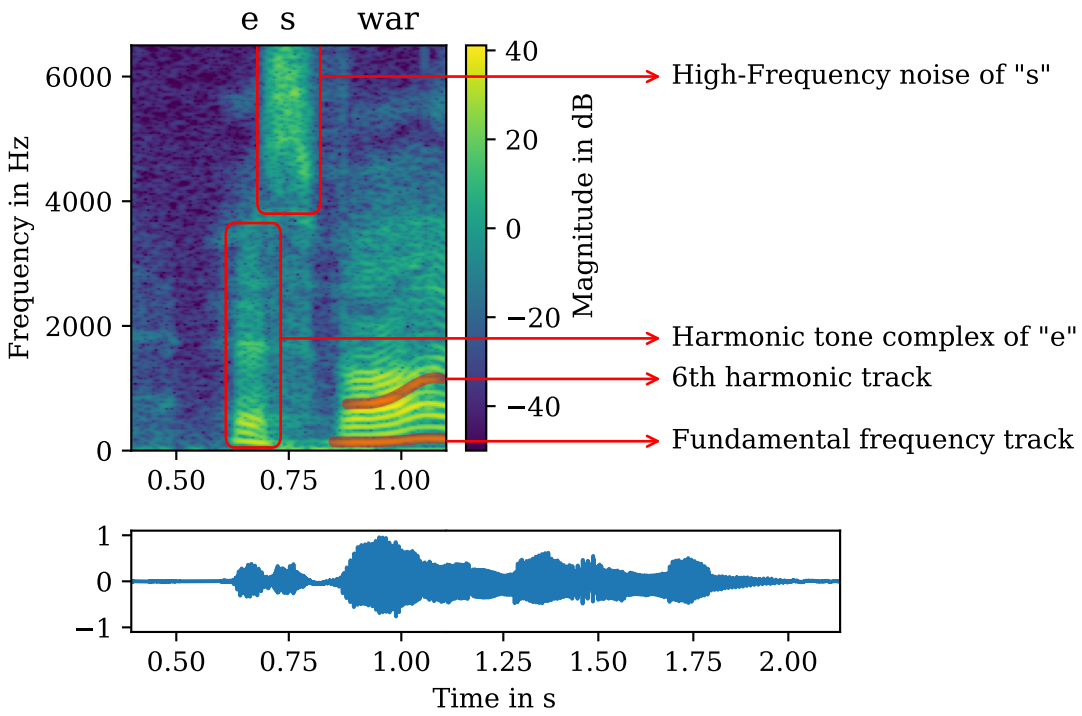


Figure 2.4: Spectrogram of a speech signal, with voiced and unvoiced parts. Voiced parts consist of low-frequency parallel tonal tracks that form a harmonic tone complex. Unvoiced parts are noisy with a broader spectrum. All parts have a time-varying spectral shape.

In signal processing terms, speech can be described as a noisy or tonal excitation, filtered by an adjustable peak filter bank, and shaped by variable air flow and openings. Atonal parts of speech are shaped noises, while tonal parts exhibit a harmonic structure.

Figure 2.4 shows an example recording of voiced and unvoiced speech. The unvoiced part of the speech is a noisy segment above 4000 Hz around 0.75 s. The voiced speech directly before 0.75 s and after 0.8 s has a fundamental frequency track around 150 Hz, with harmonics at integer multiples of the fundamental. We perceive each entire tone complex as a tone with a pitch of the fundamental around 150 Hz. Formants are visible in the above example as broad magnitude maxima around 500 Hz, 1800 Hz, and 2500 Hz for the /e/, and 700 Hz, 1000 Hz, and 2500 Hz for the /a/. These spectral shapes give rise to the /e/ and /a/ timbre.

The different phonemes in Figure 2.4 clearly overlap (“es”), and morph into one another (“war”). Each phoneme in this example spans roughly 100 ms, including its onset and offset, and the spectrum constantly changes without ever coming to rest in a steady state. Just before the /w/ is a *stop*, where the air flow is momentarily interrupted. This indicates a break between words in this case, or it might be considered part of the succeeding phoneme, depending on context. Thus, phonemes should not be seen as discrete and unique entities, but merely as transitory states that the vocal tract expresses as it produces speech.

Chapter 3

Speech Signals for Pitch Analysis

Humans use speech not only in calm conversation, but also as a means for expression, far beyond the limits of language: melodically and rhythmically in music, shouting or screaming in anger or fear, whispering, or even with pathological defects. Speech happens in all situations of life, from a quiet conversation with a friend to shouted exultations during a rock concert. This can happen in all of humanity's languages, with voices ranging from the chirping of small children to creaky old men.

This incredibly wide scope of use cases is too broad to be studied in its entirety. For the purposes of analyzing the fundamental frequency of speech in this dissertation, a practical subset of speech has to be chosen:

> *In this work,* speech *shall refer exclusively to non-reverberant, single-channel, digital recordings of a single normal*[1] *adult voice at comfortable loudness and additive background noise, speaking plain English.*

This limitation allows the use of multiple, large speech and background noise databases. This is a critical advantage as a scientific work, as these databases are widely used in other publications, making results comparable to others; databases are freely available, making results reproducible; and the databases are annotated with various linguistic and physiological metadata such as fundamental frequency ground truths or laryngograph recordings.

In the rest of this chapter, we shall look at each of these limitations, and clearly delineate their specific purpose and what this means for the applicability of our findings:

Digital Recordings

All of the analysis done in this work is implemented as computer programs, using digital signal processing algorithms. Speech signals are represented as sampled, quantized series of numbers. Depending on the database, sample rates of 16 kHz, 20 kHz, and 48 kHz and bit depths of 16 bits were used.

The conversion from sound waves to numbers limits the frequency content of each speech recording to half the sample rate and introduces a small measure of quantization noise. In practice, the quantization noise is largely irrelevant, as most investigations in the present dissertation introduce additional background noise far higher in level than that. Similarly, since speech power generally decreases with frequency, the high-frequency components discarded by sampling would likely be obscured by background noise anyway.

Digital signal processing is done in computer programs using 64-bit floating-point numbers. This data format far exceeds the resolution of the original recordings, and its errors should be small in comparison to the quantization noise discussed above.

[1] *Normal voice* here refers to the technical term for a mode of phonation, not the colloquial synonym of "ordinary".

More importantly, many of the third-party algorithms used in this dissertation were programmed by inexpert programmers, and sometimes produce errors in unexpected signal conditions or due to numerical instabilities. If these errors caused the program to malfunction or stop, it was re-run with a different signal. However, some errors occurred silently and led to incorrect results. While precautions were taken to be as error-tolerant as possible in our evaluations, the fact remains that many of our evaluations are unprecedented in scope, and therefore are bound to find error cases that the original authors were not aware of. Such programming errors might show as diminished accuracy in our evaluations.

Single Channel Recordings

Humans perceive speech with both ears and can make use of the comparative information of these two streams, both in frequency content and in timing differences. These cues are highly useful to humans for distinguishing between different simultaneous speakers and ignoring background noise. In signal processing terms, multi-microphone recordings allow for the use of beamforming[2] to reduce background noise and reverberation [154, 158].

However, using these features in this dissertation would also require additional considerations of sound source locations and room acoustics. While interesting in their own right, these issues are ancillary to understanding the fundamental frequency of voiced speech. Hence, in the interest of focusing more readily on the time-frequency structure of speech, issues of multi-microphone recordings, room acoustics, and head-related transfer functions, are excluded in this evaluation.

This simplification has the downside of ignoring a whole host of pertinent aspects of speech perception and might indeed put our algorithms at a disadvantage in comparison to human listeners. On the other hand, it removes a few variables from the equation as well, and greatly simplifies experimental aspects of the evaluations.

Non-reverberant Speech Recordings

In the real world, most speech signals happen in noisy, reverberant environments. In contrast, all the speech databases used in this dissertation were recorded in sound-proofed recording environments, with as little distractions as possible. Such clean recordings can be mixed with noise recordings for a plausible simulation of speech recordings in noisy environments, and allows for the evaluation of various signal conditions without having to re-record audio material for every new condition.

On the other hand, there is some fidelity that is lost in this simplistic process. Re-creating each evaluation scenario in the real world, with real reverberations and real environmental noises, would certainly be more life-like. As these recordings are time-consuming to set up, however, including them would necessarily limit the scope of what could possibly be evaluated in the given time. The points on *additive noise* and *background noise* below shed more light onto the limitations of this methodology.

Single Speaker

Speech exhibits an intricate structure in time and frequency. When two speakers are speaking at the same time, these structures overlap and mix, making the details of each speaker that much harder to discern.

In the real world, such mixing occurs regularly and human listeners are very good at distinguishing between multiple speakers. Humans have various methods at their disposal for separating simultaneous speakers, some of which are based on inter-aural differences, while others are applicable to single-channel recordings as well [98, 160].

[2] A technique that filters signals by direction of arrival. Also known as *spatial filtering*.

However, in most real-world conversational scenarios, humans tend to focus on one dominant speaker at a time. As this work includes background noises with speakers in them as well as explicit babble noise[3], we will still deal with the disturbances caused by interjecting speakers, but will not indulge in trying to separate the speakers or estimating their different voice patterns.

Normal Voice

There are at least four common modes of phonation: Normal speaking, breathy voice, falsetto, and hoarse speaking, which differ in the way the vocal cords vibrate [6]: In normal speaking, the vocal cords open and close repeatedly to admit short pulses of air into the vocal tract. For the breathy voice, such as when whispering, the vocal cords never fully close, and a small continuous stream of air remains at all times. When strained during a workout or vocal stress[4], we use a hoarse voice, where the pulses are shorter and more abrupt. Finally, falsetto or head voice is mostly used when singing and reaches higher frequencies by closing the vocal cords only very briefly each iteration.

All four of these modes can be used to produce speech, but sound distinct and are utilized for different purposes. The speech databases employed in this work exclusively contain normal mode phonations, which we use for calm, conversational speech.

Using only the normal mode of phonation is a reasonable limitation for the purposes of studying fundamental frequency estimation of normal speech. However, while Western European languages certainly have little use for other modes of phonation, other languages might. Furthermore, this equally eliminates any study of pathological voices and might make this work's findings less applicable to emotionally charged voices.

Adult Speakers

The speech databases used in this study are limited to the voices of adults. As humans grow from children to adults, all of their vocal organs grow with them, which gradually lowers resonant frequencies both in the vocal folds and the vocal cavities, while vastly increasing the volume of the lungs. These changes vary between sexes, and can be rather abrupt, particularly during puberty. Male adolescents additionally switch from using their head voice to their chest voice in the late stages of puberty, which changes their voice yet again [52, 164].

Due to these changes, the voices of children, adolescents, and adults are quite distinct, and valued differently for their particular qualities. In ancient times, the voice of castrated males, with adult vocal power but a child-like larynx, was deemed the most beautiful of all. While this has (thankfully) fallen out of favor during the last centuries, pubertal development is sometimes intentionally delayed to preserve singers' voices slightly longer [52].

As these variations and changes to the voice are sweeping, their study would unduly widen the scope of the present work. This is particularly tragic, as it seems to happen all over the signal processing community, a fact easily observed when children try to use any kind of voice-controlled technology and invariably fail to be understood [123, 165].

Thankfully, most speech databases are at least somewhat balanced with regard to sexes and pitches, which provides at least some measure of the variabilities of human voices. A further investigation of the diversity of speech databases is presented in Chapter 9.

Additive Noise

Background noise is everywhere, from the faint whir of a computer fan, to birds chirping, to

[3] A mixture of many simultaneous speakers.

[4] or in an involuntary response to loud background noises, due to the *Lombard Effect* [15].

the din of construction work, and the hustle and bustle in a busy cafeteria. There is almost no situation in life where we are free from the sounds of nature or civilization.

Consequently, speech is a signal that is necessarily resilient against noise and can still transport information even if severely degraded. Typical background noises have a wider frequency content than speech. Thus, high-frequency components of speech are often the first to be drowned out in noise, with the rest following shortly thereafter, which is probably why speech is particularly tolerant to high-frequency degradations. Still, most speech remains intelligible up to signal-to-noise ratios (SNR) of roughly 5 dB [90].

This dissertation acknowledges this resiliency by mixing speech signals with background noise recordings at various SNRs. We employ multiple background noise databases, which contain noises from all walks of life, including the aforementioned traffic noise, construction work, and cafeterias, as well as artificial and more uniform noises such as white noise.

However, this only covers a fraction of the disturbances possible in the real world. One particular omission is recording degradations such as jitter or frequency distortions. These kinds of noises are non-additive and can range from minor annoyances to rendering the recording entirely unrecognizable, but cannot be recreated through additive noises from common noise databases. On the other hand, these kinds of distortions are usually artifacts of the recording medium and do not occur acoustically. As such, this dissertation does not consider them an aspect of human speech and will not include them in its evaluation.

Another common issue that does occur acoustically is reverberations. Reverberation happens when speech or noise is reflected off several surfaces, and therefore reaches the microphone from more than one path simultaneously. Reverberations are heard particularly strongly in enclosed spaces with hard surfaces, such as churches or caves. Normal living spaces often include dampening elements such as lowered ceilings in classrooms or furniture in homes, which reduce reverberations. In moderation, reverberation is a normal part of speech and easily tolerated. When too strong, however, speech can become hard to understand quite quickly. For the speech databases, recordings are typically from acoustically insulated recording booths, but noise databases often include all naturally occurring reverberation.

In the context of the present work, reverberations pose a serious problem, as they introduce a number of difficult-to-factor variables, such as the amount of early and late reflections to apply to the speech signal and to the background noise. These factors are closely related to the problems of multi-channel recordings discussed earlier and would introduce a number of new environmental variables into the evaluations. Furthermore, all recordings in our databases already include a small measure of reverberations on their own, which might produce unnatural results if combined with simulated reverberations.

In light of these difficulties, this work will ignore reverberations, beyond what is present in the databases already.

Plain Language

There is a wide range of emotional stances possible in the production of speech. We use rhythm, pitch, and stress to convey meta-information about emphasis or emotions without changing the phonemes themselves. These speech characteristics are called prosodic features, and greatly enhance our ability to communicate through speech beyond the actual words.

However, the databases used in this work mostly contain simple sentences read in a level voice, without much emotion attached. This is partly because they are being read out-of-context by the study participants, and partly to keep the database entries as interchangeable as possible.

This robs the speech material used in this dissertation of a natural means of expression that is an integral part of conversational speech. Additionally, it makes our results less applicable to other languages that require the use of prosodic features. On the other hand, prosodic features do not change the actual phonemes. Their impact on the time-frequency structures of speech is therefore likely to be small, and an acceptable price to pay for using well-regarded speech databases.

English

Lastly, this dissertation uses recordings of the English language only, spoken by European and American speakers.

This limitation is mostly due to the language of science (in the West) being English, and therefore the higher availability of datasets in that language. After all, a work in German, French or Italian is not as likely to be read widely as a work in and about English. Consequently, even non-English scientists often prefer to create speech databases in English rather than in their native tongues, on account of being more widely citable [119]. In fact, this very study does not even include the author's native language of German at all, despite that native material obviously being understandable to the author. Even the English language itself could provide more varied dialects than are represented in our databases.

Most techniques discussed in this work should be applicable to other languages as well, since this work is mostly concerned with the basic capabilities of the human vocal tract, rather than their language content. In particular, European languages are relatively similar to one another phonetically, and are unlikely to differ much in the pitch of their speech [13]. Thus, there is little value in repeating each experiment in multiple European languages for the purposes of fundamental frequency estimation. Still, the more different a language is from central European languages and English in particular, the less applicable our findings are likely to become.

In summary, speech is humanity's primary means of communication, and is consequently used in a vast array of circumstances and modalities, far beyond what could reasonably be studied in one dissertation. We therefore limit the scope of this work to a small subset of "typical" speech in noise that we hope is sufficiently representative in terms of its fundamental frequencies, while minimizing the number of variables.

Chapter 4

Conclusions

The usefulness of speech and language to humans does not require explanation; it is so integral to our existence and society that we could hardly call ourselves "human" without it. Yet, our understanding of how speech works, precisely, remains an active area of research[1].

This introduction covered speech and its fundamental frequency, their production in the vocal system, and their perception in the auditory system. Without question, there is an enormous depth to these topics that remained unmentioned. Indeed, entire books have been written about each one of them [35, 98, 131].

Many descriptions of speech have been given, in terms of a linguistic stream of information, as a physiological process of phonation, an acoustic sound wave, and as an auditory perception. In recent years, these interpretations have been expanded in a new direction: speech as a digital signal for analysis and synthesis in computer programs.

This new interpretation of speech for the purposes of digital signal processing allows speech to be understood and produced by non-human, technological devices, albeit (presumably) not with the same fidelity as human listeners and speakers.

In this context, the investigation into voiced speech's pitch, and thereby into the detailed structure inherent to speech, needs to be defined both in technological terms and in physiological ones. Historically, this distinction was made clear in the use of the technical term *fundamental frequency*, and on the other hand *pitch*, its human perception.

However, both terms leave something to be desired: While *pitch* is simply a human percept, it is by no means an unambiguous one. That "quality that is high or low" can in fact be both, simultaneously, which makes the term a bit too loose to be estimated in a computer algorithm.

The technical *fundamental frequency* is similarly vague, in that its definition is only clear-cut for simple sinusoids, but not speech. Multiple definitions of a "fundamental frequency" of speech were brought forth in this introduction, all relating to some aspect of our production or perception of pitch. After all, speech is only relevant in a human context, and so must a well-defined fundamental frequency of it correspond to our perception.

In this dissertation, we will therefore leave both of these definitions open to interpretation. *Pitch* will be used both as a human percept, and as an algorithmic approximation thereof, and *fundamental frequency* will mean both "the frequency of the lowest harmonic", as well as more perceptual interpretations.

With such leeway in the definition of pitch or fundamental frequency of speech, estimation methods are similarly diverse. Algorithms estimate pitch from periodicity measures, as in speech production, or from harmonicity measures, as in perception, or directly from a ground truth and machine learning methods, or a combination thereof. Naturally, these approaches come with different strengths and

[1] also known as "a mystery" to non-scientists.

weaknesses that will be investigated in great detail in a later chapter.

Evaluating the performance of such fundamental frequency estimation algorithms requires a reliable ground truth to compare against. It may be calculated by a reference algorithm either from clean versions of the speech signal in question or from an additional laryngograph recording if available, or by measuring consistency of a single algorithm at various SNRs. The construction of a suitable ground truth and evaluation methods useful for comparing and evaluating algorithms will make up the remainder of this dissertation.

Thus, this dissertation accomplishes four distinct tasks to investigate the algorithmic estimation of fundamental frequency of speech:

- Part II introduces a set of tools for analyzing speech signals in the short-time Fourier transform, including contributions to phase spectral analysis methods and a new visualization technique for phase data.
- Part III describes a new algorithm for fundamental frequency estimation that uses phase spectral information as well as traditional techniques.
- Part IV constructs a new consensus ground truth method and database to overcome the challenges with existing ground truths for speech signals.
- Part V evaluates a large range of existing fundamental frequency estimation algorithms with traditional and new evaluation methods, databases, and ground truths.

Part II starts with Chapter 5, which introduces the foundational techniques necessary for analyzing speech signals in the rest of this dissertation. In particular, it discusses the properties of the short-time Fourier transform, which is a key ingredient in the analysis of speech signals, which disentangles signals along time and frequency. Doing so, however, introduces new parameters such as window functions and block lengths that need to be chosen with care to not fool our algorithms and eyes.

The chapter ends with visualization techniques for short-time Fourier spectra. This is critical for the analysis of speech signals, as our sophisticated auditory perceptions are unavailable to algorithms, and we must thus rely on information we can see with our eyes instead. This chapter also includes our first contribution, in the form of a new perceptually-smooth, circular color map purpose-built for visualizing phase spectra.

Chapter 6 expands this analysis to derivatives of the short-time Fourier transform, which conveniently re-integrate the flow of time into the short-time Fourier transform. Additionally, phase derivatives provide an easy interpretation for the otherwise obtuse spectral phases, and thus make them available to analysis. While parts of these techniques have been published before, their synthesis with specifically-adapted window functions is a minor contribution of this dissertation. The part concludes in Chapter 7 with a summary of lessons learned.

Chapter 8 in Part III applies these techniques to design and evaluate an algorithm for estimating the fundamental frequency of speech. The novelty of our approach is in using multiple signal representations to re-frame fundamental frequency estimation as a fine-grained voicing detection problem. This enables the algorithm to not just select a most likely fundamental frequency, but to predict accurately whether an estimate is salient or not. Thus the algorithm is comparatively reluctant to label a frame voiced, but much more precise when it does. These characteristics are evaluated and confirmed on a large corpus of speech and noise recordings and in reference to other common fundamental frequency estimation algorithms.

However, such claims of accuracy can only ever be made in comparison with ground truths, which are typically part of published speech corpora. The choice of speech corpus and ground truth therefore imparts unavoidable biases on every evaluation. To investigate these biases, Chapter 9 of Part IV examines various speech corpora in terms of their diversity, spectral characteristics, and fundamental

frequency ground truths. A number of significant differences are found, which have no doubt influenced many an algorithm in the past, and will continue to do so in the future.

To lessen their impact, Chapter 10 introduces a new ground truth, the *consensus truth*, which replaces the corporas' own estimates with the consensus of multiple fundamental frequency estimation algorithms. In contrast with common laryngograph-based ground truths, the consensus truth is derived from the same category of data as an algorithm's estimate, and is therefore more compatible in its failure modes and voicing decision. Apart from these details, results are very similar to published ground truths, and can therefore be used interchangeably with existing ground truths, or as a source of truth where no ground truth is available.

With this new and independent consensus ground truth, and a number of well-examined databases available, Part V evaluates the estimation characteristics of various fundamental frequency estimation algorithms in an unprecedented variety of circumstances. Chapter 11 introduces the parameters of this evaluation, including a historical retrospective on the algorithmic developments of the last thirty years, a literature review of algorithms and past comparison studies, and a thorough definition of performance metrics and computational considerations. The literature review shows little consensus among previously published studies, which underlines the need for rigorously defined performance metrics and evaluation parameters such as ours.

Thus, Chapter 12 examines these algorithms, corpora, and performance metrics in detail, in order to find the strengths and weaknesses of different approaches. In the end, this unprecedented comparison study reveals a number of hitherto unpublished details and characteristics, both of the algorithms' behaviors, and that of reference databases and common performance metrics. As a conclusion, we must abandon the notion of a *best* algorithm for all applications. Instead, there can merely be algorithms and corpora well-suited for particular tasks, and a general need for more diverse corpora and more comprehensive performance measures.

Finally, Part VI concludes this dissertation with a summary of the lessons learned, and a skeptical outlook of the field's future. After all, speech is inimitably human, and the variations between fundamental frequency estimation algorithms were found to be just as large as between different humans. This meta-analysis of fundamental frequency estimation turned out to provide far more interesting insights into the foundational values of speech analysis in general than the specifics of particular algorithms.

An Appendix then provides additional resources, such as notes on the implementation of the evaluation framework, source code, and summary profiles of the fundamental frequency estimation algorithms tested.

Part II

Analysis Techniques for Short-Time Speech Spectra

Where the necessary tools of speech analysis are laid out for analyzing the fundamental frequency of speech signals.

Chapter 5 formally introduces our main signal representation, the short-time Fourier transform, which disentangles signals in time and frequency. These spectra spectra reveal the harmonic or periodic structure of voiced speech, but they can only show one or the other, depending on its windowing parameters. Additionally, techniques for visualizing short-time Fourier spectra are discussed with respects to both technical limitations and those of human visual perception. The chapter ends with a new color map designed for displaying cyclic phase data without visual artifacts.

Chapter 6 expands the short-time Fourier spectra with an introduction of its derivatives in time and frequency. These are particularly useful for spectral phases, whose cyclic nature makes them otherwise difficult to interpret.

The part ends with a conclusion in Chapter 7.

Chapter 5

The Short-Time Fourier Transform

Human listeners experience sounds across many dimensions, such as timing, rhythm, loudness, pitch, coloration, timbre, and many more [98]. As rough categories, these can be grouped in *timing features*, where a property changes from phoneme to phoneme; and *spectral features* that describe the particular qualities of a single phoneme's sound.

This two-dimensional view of interpreting a sound in time and frequency is rooted in the physiology of the inner ear, which splits signals into frequency components that vary over time. But why adhere to these physiological limitations in signal processing? After all, there is no rule that signal processing algorithms need to follow human biology. Yet, speech is a special case, where both the production and perception of speech is an inherently human affair that derives all its meaning and interpretation solely from its use by humans. A physiological interpretation of these signals therefore seems particularly promising for technical interpretations as well.

This chapter will introduce various methods of splitting a speech signal along time and frequency, which all follow a common pattern: Splitting the speech signal into short, overlapping blocks, then transforming each block into the frequency domain with the help of the Fourier transform. These methods are commonly referred to as "time-frequency representations" or "spectrograms".

To calculate a spectrogram, a window function $w(n)$ is applied to the signal $s(n)$ that extracts a short section of the signal, commonly referred to as a "block". The Fourier transform of the block is

$$S(t, f) = \int_{-\infty}^{\infty} s(n) \cdot w(n - t) \mathrm{e}^{-i2\pi f n} \, \mathrm{d}n, \tag{5.1}$$

where the block is centered around a time t, and the Fourier transform is calculated for a frequency f. The window function is non-zero around its center $w(n = 0)$, and zero at $w(|n| > N)$, thus limiting the block to a length N.

In the following chapters, a variety of such spectrograms will be examined for the purpose of analyzing speech signals. These methods will vary the window function, the block length, and add post processing steps, to highlight different aspects of speech signals. We will refer to the form shown in Equation 5.1 as the short-time Fourier transform (STFT).

Figure 5.1 shows the STFT magnitude of a short speech segment. The lower half of the spectrogram shows horizontal line patterns, which represent the speaker's voiced speech as a fundamental frequency around 130 Hz, and parallel harmonics at integer multiples of the fundamental. Noisy consonants are shown as vertical bands across the entire frequency range with no harmonic structure. Formants are visible as broad spectral maxima superimposed on the harmonics around 500-1500 Hz.

The STFT in Figure 5.1 uses a raised-cosine Hann window $w_{\mathrm{hann}}(n)$, which is a common choice in speech analysis:

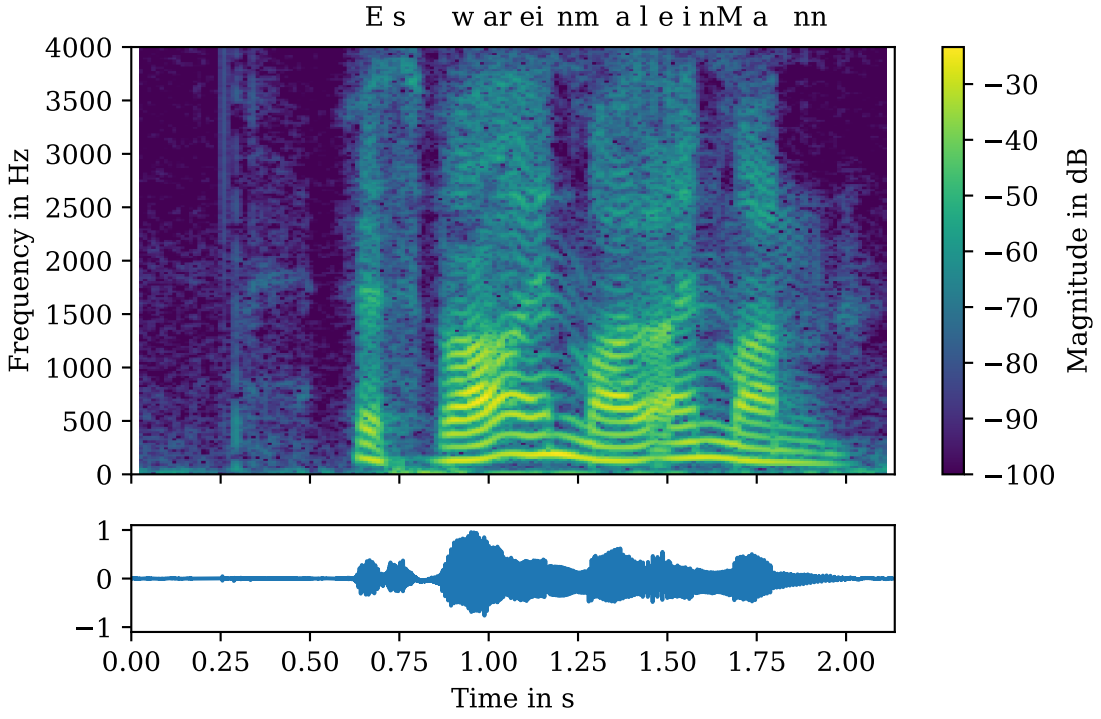


Figure 5.1: An STFT magnitude of the author speaking the short sentence "Es war einmal ein Mann" (top), and the signal waveform (bottom).

$$w_{\text{hann}}(n) = \begin{cases} \frac{1}{2}\left(1 + \cos(\frac{2\pi n}{N})\right) & \text{if } |n| < \frac{N}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

An algorithm called Fast Fourier Transform (FFT) is typically used to calculate an STFT with much fewer steps than Equation 5.1 might imply. Calculated this way, a time signal is decomposed into a complete set of coefficients for a regularly-spaced set of frequencies that fully describe the original signal. In many ways, large parts of the field of signal processing and modern life only became possible thanks to the invention of the FFT and its multitude of analytic and synthetic applications [129].

In the STFT, signals are representated as a two-dimensional matrix of "bins" along time and frequency, which are complex numbers that are often decomposed into a magnitude $|S(t,f)|$ and a phase $\angle S(t,f)$ component:

$$S(t,f) = |S(t,f)|\, \mathrm{e}^{i\angle S(t,f)} \tag{5.3}$$

If calculated with the FFT, the Fourier integral extends from $0 \dots N$ instead of Equation 5.1's definition of $-N/2 \dots N/2$, thus introducing a phase shift of $N/2$. To counteract this phase shift, the FFT-calculated $S_{\text{FFT}}(t,f)$ needs to be phase-shifted back:

$$S(t,f) = S_{\text{FFT}}(t,f) \cdot e^{-i2\pi f} \tag{5.4}$$

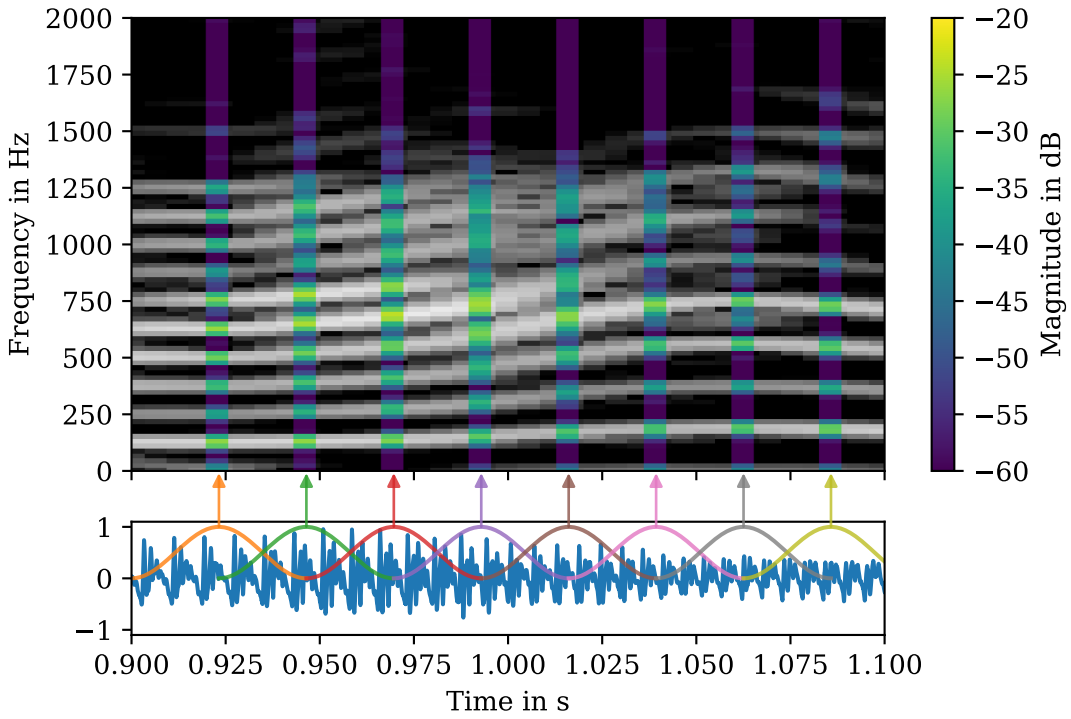


Figure 5.2: Illustration of the STFT: Cut the signal into short, overlapping blocks, apply a window, and Fourier-transform each block. Spectra in color are calculated from the signal blocks indicated by the colored window functions on the waveform. Window functions of greyscale spectra are not shown. All spectra are calculated with ca. 80 % overlap of 50 ms blocks and Hann windows.

Each STFT spectrum is derived from a signal block centered around a specific time index t. In practical applications, however, STFTs are rarely calculated for ever time index of the signal, as this would result in highly-redundant spectrograms due to the large overlap of neighboring blocks. Instead, STFTs are commonly only calculated for fewer instances, such as every block length N, or a fraction thereof. In the latter case, the time resolution of the STFT is commonly characterized by the overlap percentage between succeeding blocks.

Figure 5.2 shows an example of an STFT magnitude of a short speech segment, with a number of window functions superimposed on the waveform, and their STFT magnitude spectrum highlighted to illustrate how signal blocks are transformed into STFT spectra. Note how blocks are displayed centered around their center time t, where their raised-cosine window functions are maximal. This is in contrast to alternative definitions that define t as the start or end of each block, where our raised-cosine window functions would be zero.

The STFT magnitude is commonly displayed logarithmically in decibels (dB), which are $20 \cdot \log_{10}(|S(t,f)|)$, as this represents simultaneous signals as additive in the STFT, and generally compresses the STFT magnitudes to a more legible range. As this dissertation is only concerned with digital recordings, any use of decibels here is in reference to a sinusoidal signal with a full-scale amplitude of -1 or 1. This unit of measurement is sometimes abbreviated dB FS, which in this dissertation is synonymous with dB.

The STFT phase is harder to visualize and interpret than the STFT magnitude. Low-resolution Image graphs such as Figure 5.2 are in fact mostly inscrutable as phase angles vary too quickly in time. Phase angles rotate once per signal component period, which only becomes visible at very high

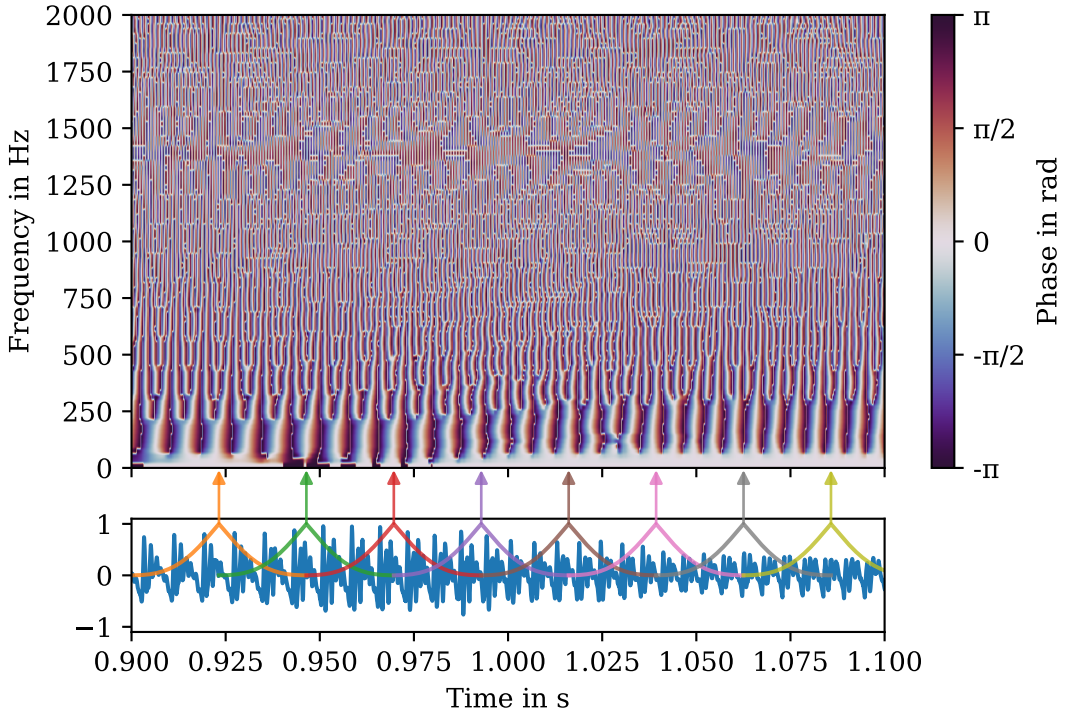


Figure 5.3: STFT phases of the same signal segment as in Figure 5.2, but with ca. 99 % overlap and Hann-Poisson windows. One phase rotation per signal component period, such as ca. 130 Hz for the fundamental, and ca. 260 Hz for the first harmonic.

overlaps, as shown in Figure 5.3. Chapter 6 will look into various alternative methods of interpreting and visualizing this data that does not require such high time resolution.

The rest of this chapter is organized as follows: The minimum block length requirements for speech processing are discussed in Section 5.1. If all the data in the speech signal is to be analyzed, consecutive blocks should overlap, which is discussed with a general overview over the properties of various window functions in Section 5.2. Section 5.3 will explore the challenges and techniques of visualizing STFTs in more detail.

5.1 Block Lengths

Speech signals are made from a rapid series of different configurations of the vocal organs. For the purposes of speech analysis, both the content of each configuration, and the sequence of configurations is of interest. Thus, a compromise has to be struck between long block lengths with considerable spectral resolution, and short block lengths that show detailed differences over time.

Specifically, voiced speech contains harmonics as closely spaced as 80 Hz for a deep voice. In order to resolve these individually, the harmonics need to be separated by at least two bins in the spectrum. Furthermore, voiced speech sounds remain relatively unchanging for at least 25 ms [110, 6], so block lengths should be no longer than that[1].

[1]Even though most speech processing publications use block lengths between 10 ms and 40 ms, actual evidence for that choice is surprisingly hard to find. Chapter 3.8 of [110] quotes a number of durations between 25 and 180 ms for phoneme durations, but also states that stress, speaking rate, and sequence influence phoneme duration, while coarticulation and varying pitch contours will introduce intra-phoneme spectral shifts.

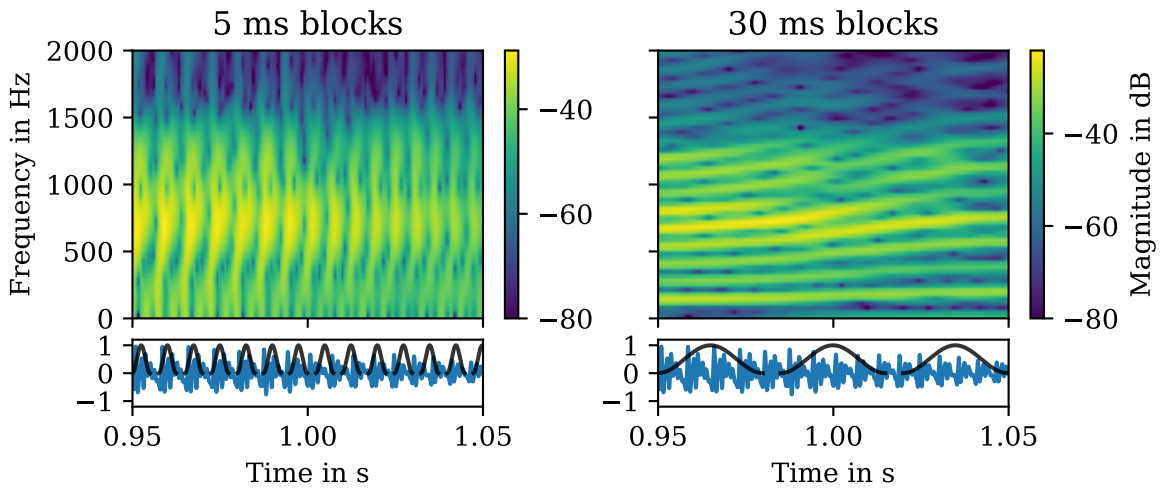


Figure 5.4: Two STFT magnitudes of the same voiced speech segment at different block lengths. At short block lengths, STFT appears periodic; at long block lengths, it looks harmonic. Window function shapes are shown in the bottom plot (different Y scale).

To accommodate both of these requirements at common sample rates, block lengths should be approximately within:

$$\frac{f_s}{2 \cdot 40 \text{ Hz}} \leq N \leq 0.020 \text{ s} \cdot f_s \tag{5.5}$$

For $f_s = 48\,000$ Hz this means block lengths $600 \leq N \leq 1\,200$, and for $f_s = 16\,000$ Hz it means $200 \leq N \leq 400$. For example, common audio codecs in GSM use a sampling rate of 8 000 Hz, and block sizes of 20 ms [156, 136].

However, the maximum window length is only applicable to the rectangular window, which includes each block in its entirety. Many window functions instead taper off towards zero at their beginning and end, in order to trade some time resolution for improved spectral acuteness. Such tapering windows render only the block center with good accuracy but attenuates block fringes. It is therefore reasonable to use significantly longer block lengths for these kinds of windows, as investigated in Chapter 5.2.4.

5.1.1 How Block Length Affects the STFT

At its source, voiced speech is produced by periodic opening and closing of the glottis, which releases puffs of air into the vocal tract. Thus, at short enough time frames, voiced speech is a periodic signal made from glottis pulses. However, all STFTs up to now showed non-periodic, harmonic patterns instead, with a fundamental frequency and multiple harmonics at integer multiples of the fundamental.

This duality of a periodic signal, giving rise to a harmonic spectrum, is a fundamental property of the STFT. The long window lengths used in all STFT graphs so far encompass multiple pulses and show their interactions as harmonic patterns. If window lengths are shortened to include only single pulses, the harmonic pattern is replaced with a periodic one. It is instructive to look at this duality, since viewing voiced speech as either periodic or harmonic highlights different aspects of its structure.

Figure 5.4 shows the STFT magnitude of a short segment of voiced speech at different block lengths. In the left graph, each block only contains a single glottis pulse, resulting in a periodic STFT magnitude. In the right graph, each block contains multiple glottis pulses, and shows a harmonic STFT magnitude.

The left "wide band" STFT with a short block length clearly shows how the speech signal is made from individual glottis pulses. The power and time of these pulses matches the shape and rhythm of the waveform, while the spectral distribution of the pulses shows the resonances of the vocal tract.

Interestingly, the pulses are slightly dispersed over time, with lower frequencies peaking earlier than higher frequencies. In the shown case, the blocks are so short that the STFT magnitude peaks at the fundamental frequency both at the waveform maximum, as well as the minimum with a zero in between. This multi-peak structure seems to repeat at a higher frequency as well.

This kind of fine-grained timing information is absent in the right "narrow band" STFT with a block length spanning multiple glottis pulses. Instead, the signal appears harmonic, as a series of parallel horizontal lines. The resonances of the vocal tract are still visible as varying partial magnitudes. In this graph, the most obvious feature is the slight frequency modulation, which sweeps the fundamental frequency from appr. 130 Hz to 190 Hz.

This has important ramifications for STFT spectra at different block lengths: Since harmonicity is a side-effect of multiple clicks in the same window, it can only develop if the block actually encompasses multiple clicks. If block lengths are shorter than that, there will no longer be any interaction between multiple clicks, and the STFT spectrum will lose its harmonic structure.

5.2 Window Functions

The first step in the short-time Fourier transform is to cut the signal into short, overlapping blocks. Then, a window function is applied to each block, which typically tapers to zero at the beginning and end. Finally, each windowed block is Fourier-transformed.

The purpose of the block window is to work around an inherent quirk of the STFT: cutting signals into short blocks introduces sharp transitions where the cut is made. Tapering window functions hide these block transitions by gradually attenuating the signal towards the ends of the block. With no amplitude at the beginning and end, there is no sharp transition anymore. However, this replaces the transition problem with a new problem: Now the shape of the signal has been tampered with, which, in turn, affects the spectrum.

Thus, the design of the window function has to strike a delicate balance between hiding block transitions, and changing the signal shape. Many window functions have been proposed for various purposes, and entire books have been written about them [124]. In this work, the main focus will be on the Hann window for magnitude spectra and the Hann-Poisson window for phase spectra. In addition, it is useful to first look at the effects of the non-tapering rectangular window.

5.2.1 The Rectangular Window

The rectangular window is a fancy way of saying that no tapering window function is used at all. All amplitudes are left as they are, and the transition problem is in full effect. However, describing this as an all-ones *window function* allows some abstract reasoning about the effects of block processing, independent of the window function or signal to be analyzed.

Formally, the rectangular window[2] of length N is

$$w_{\text{rect}}(n) = \begin{cases} 1 & \text{if } -\frac{N}{2} \leq n < \frac{N}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

As this window function is multiplied onto the signal, its spectrum is convolved with the signal spectrum. Figure 5.5 shows the rectangular window, and various ways of visualizing its spectrum.

[2]Technically, this is the symmetrical variant of it, which might not have N values 0 but is symmetrical around zero.

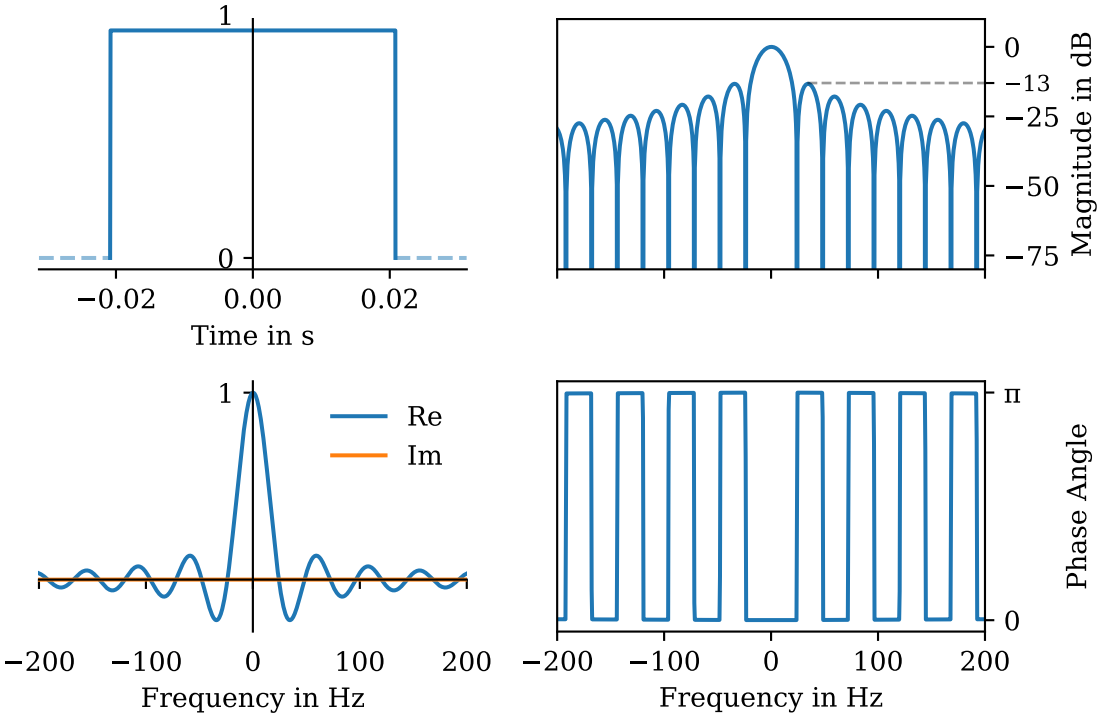


Figure 5.5: A rectangular window of length 0.04 s (top left), its real and imaginary spectrum (bottom left), its logarithmic magnitude spectrum (top right) and its phase spectrum (bottom right). Dashed lines for implied zeros outside of the window.

The spectrum has a broad central peak, called a *main lobe*, and a series of *side lobes* beginning at -13 dB, continuing to roll off at approximately 6 dB per octave. The spectrum amplitude oscillates around zero, with the zero crossings represented as zeros in the magnitude spectrum, and negative numbers as a phase angle of π [139].

In practice, this pattern is somewhat obscured by FFT's use of causal time indices, which centers the window around $t_{\text{mid}} = \frac{N}{f_s 2}$ instead of 0, and accordingly shifts all phases by $2\pi \frac{2f}{f_s} t_{\text{mid}}$. This has been removed in the visualizations by shifting them back to 0.

When the rectangular window's spectrum is convolved with the signal spectrum, the window spectrum blurs or smears every spectral bin into the lobe pattern, making the overall spectrum softer. Simultaneously, the phase of every bin is alternatingly flipped and unflipped in each side lobe.

Since every other window function also starts and ends with zeros, this smearing is a general fact of life when block processing and windowing. However, the shape of the smearing pattern and lobes is particular to the rectangular window.

5.2.2 The Hann Window

In speech analysis, the most widely used window functions are variants of the raised-cosine window, where the window is 1 in the center, and tapers off according to the shape of a raised cosine flank towards zero. The simplest of these forms is the Hann window:

This distinction is of little importance for our high-N applications with no reconstruction.

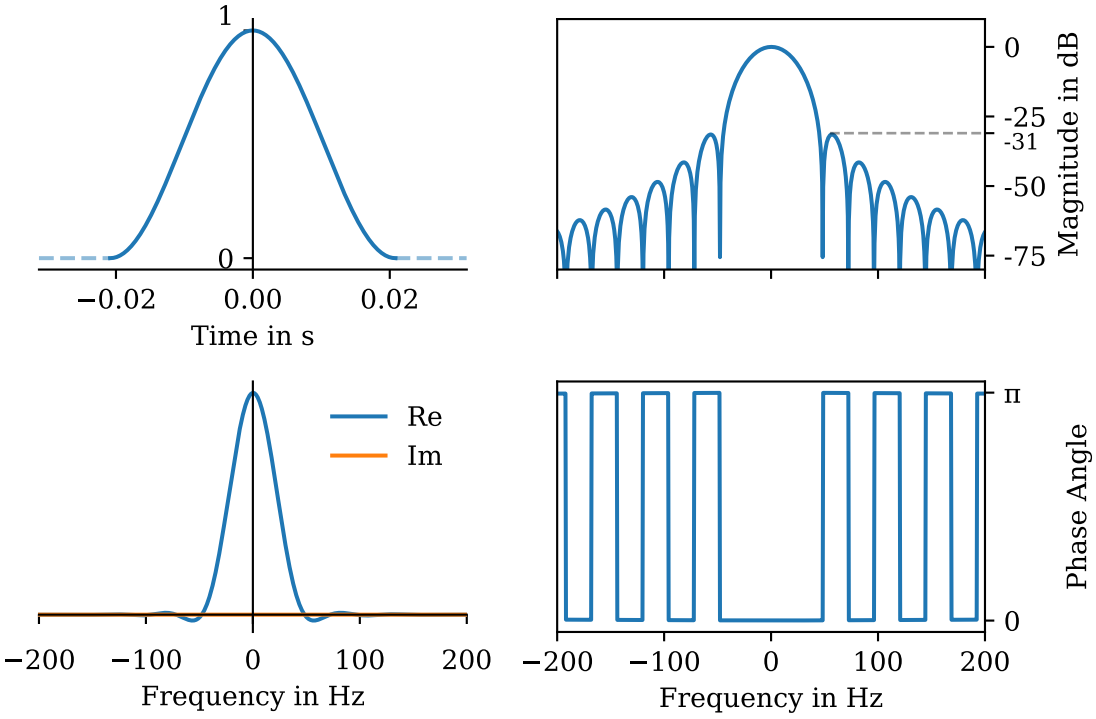


Figure 5.6: A Hann window of length 0.04 s (top left), its real and imaginary spectrum (bottom left), its logarithmic magnitude spectrum (top right) and its phase spectrum (bottom right). Dashed lines for implied zeros outside of the window.

$$w_{\text{hann}}(n) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{2\pi n}{N}\right)\right) & \text{if } -\frac{N}{2} \leq n < \frac{N}{2} \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

As this shape is a simple cosine signal, blocks with 50 % overlap add up to exactly the original signal[3]. This is of great utility for signal modification as it allows easy synthesis but is of little importance to the present analytic work.

Other popular variants of the raised-cosine window are the Hamming window, which tapers to near-zero instead of zero, or the Tukey window, which includes an area of ones in between the cosine flanks [139]. Variants of the raised-cosine window are so ubiquitous, in fact, that common software packages for calculating STFTs implicitly assume their use if no window function is specified[4].

The advantage of tapering the window towards zero is lower side lobes, albeit at the cost of a wider main lobe. In the case of the Hann window, the first side lobe is -31 dB relative to the main lobe, and further side lobes roll off at -18 dB per octave [139]. Figure 5.6 graphs the Hann window and its spectrum.

For speech analysis, the Hann window's lower side lobes afford a better separation between dominant sinusoidal tracks and the surrounding noise floor. The main lobe is considerably broader in comparison to the rectangular window, but it is still narrow enough to separate harmonics of even deep male voices at common block lengths and sample rates.

[3]The unsymmetric variant, that is.

[4]Matlab defaults to a Hamming window, Scipy to a Tukey window, and Matplotlib to a Hann window

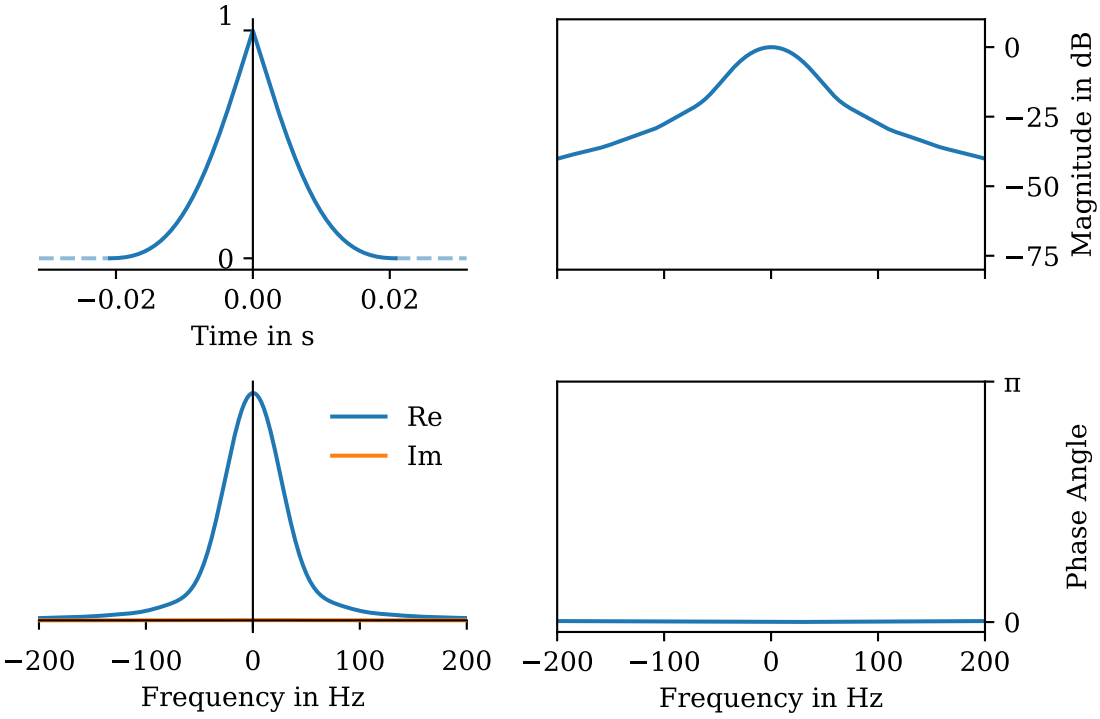


Figure 5.7: A Hann-Poisson window of length 0.04 s for $\alpha = 2$ (top left), its real and imaginary spectrum (bottom left), its logarithmic magnitude spectrum (top right) and its phase spectrum (bottom right). Dashed lines for implied zeros outside of the window.

5.2.3 The Hann-Poisson Window

In this work, much thought has been given to the phase spectrum of speech. Indeed, many aspects of the speech signal are not encoded in the magnitude spectrum, such as intra-block modulation, the precise time of clicks, or the precise frequency of a sinusoid. This information is instead encoded in the magnitude's often-overlooked sibling, the phase spectrum.

However, as the phase spectra in Figures 5.5 and 5.6 have shown, most window functions flip the spectral phases of half their side lobes. The STFT phase therefore suffers much heavier distortion than the comparatively benign smearing in the STFT magnitude. Similarly, the zeros between side lobes of most window functions create parallel ridges to sinusoidal tracks in the STFT, and thus prohibit the use of convex optimization methods[5] in the STFT magnitude [139].

Both of these issues are caused by the zeros and sign changes of the window spectrum. However, the Hann-Poisson window is a unique window function that does not have any zeros in its spectrum. The Hann-Poisson window is, perhaps quite overtly, the result of multiplying a Hann window with an exponential (or Poisson) window. Figure 5.7 shows the window function and its spectrum. The window function is given by:

$$w_{\text{hann-poisson}} = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{2\pi n}{N}\right)\right) \mathrm{e}^{-\alpha \frac{2|n|}{N}} & \text{if } -\frac{N}{2} \leq n < \frac{N}{2} \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

[5] also known as "hill-climbing" algorithms, which follow gradients to a local extremum.

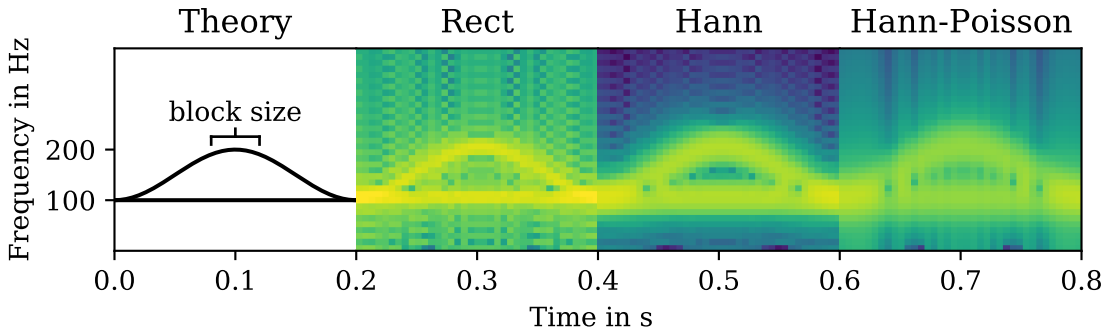


Figure 5.8: STFT magnitudes of two overlapping sinusoids, with their frequency tracks (left), and three different window functions (remainder). The three window functions use equal window lengths.

For any $\alpha \geq 2$, the Hann-Poisson window has no zeros, and a zero phase. Further increases of α increasingly smooths the side lobes and broadens the main lobe [139].

This smoothness comes at a cost, however, in that the "side-lobe" fall-off is not particularly steep and the "main lobe" smearing is rather large.

Nevertheless, analysis of the phase spectrum and derivatives of the magnitude spectrum might well benefit greatly from the absence of ripple and zeros provided by the Hann-Poisson window, which will be of great interest in later chapters.

5.2.4 Equivalent Rectangular Window Length

The preceding sections showed significant differences in main lobe width between window functions. The main lobe width is the most prominent cause of spectral smearing and should therefore be minimized. Figure 5.8 shows STFT magnitudes with the three window functions. The different main lobe widths are visible as the thickness of the spectral maxima, which is smallest for the rectangular window and largest for the Hann-Poisson window.

One way of decreasing the main lobe width is a longer window, which is usually limited by the length of stationarity within the signal, or 20–40 ms for speech signals. However, more tapered windows include very little data from the start and end of the window, so stationarity restrictions could accordingly be relaxed considerably for these kinds of windows.

Figure 5.9 shows the rectangular, Hann, and Hann-Poisson window at equivalent window lengths, where window lengths are enlarged so they include an equal amount of signal under the window. At equivalent window lengths, the Hann window is twice as long as the rectangular window, and the Hann-Poisson window 3.356 times as long.

Given these new window lengths, the right graph in Figure 5.9 shows their magnitude spectra. The graph shows no significant main lobe width difference between the window functions. If equivalent window lengths are used, main lobe width and smearing differences between windows can be ignored.

5.2.5 How Window Shape Affects the STFT Magnitude

As mentioned in the introduction to this chapter, applying any window function smears the spectrum. However, this smearing is a necessary side-effect of cutting the signal into blocks. Figure 5.10 shows the STFTs of two overlapping sinusoids, using various window functions at equivalent rectangular window lengths.

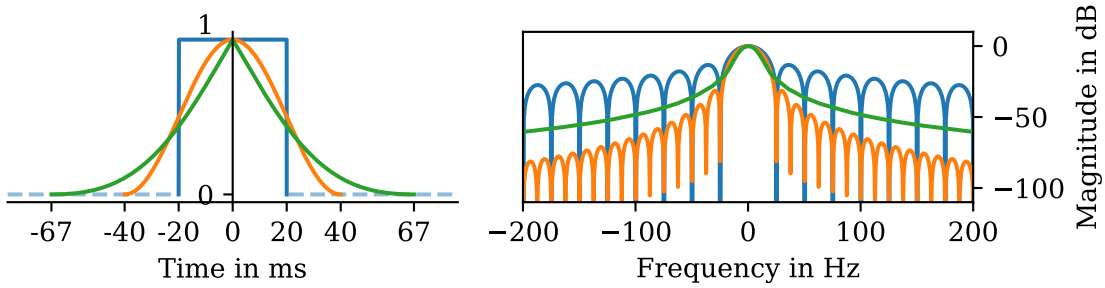


Figure 5.9: Rect (blue), Hann (orange), and Hann-Poisson (green) window at block lengths of equal time integral.

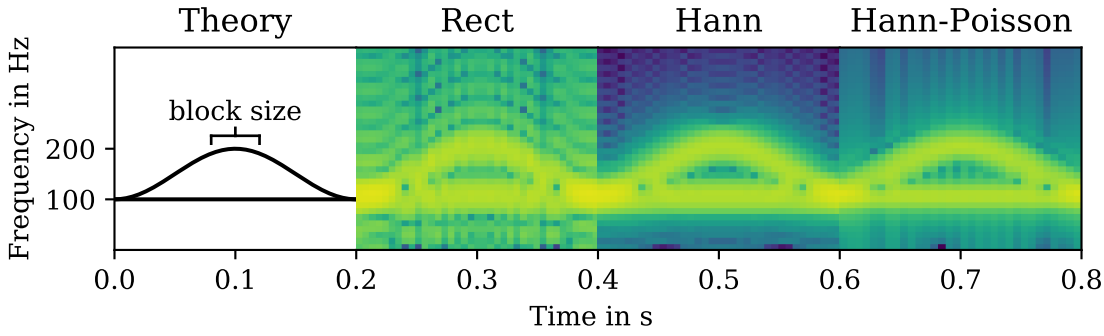


Figure 5.10: STFT magnitudes of two overlapping sinusoids, with their frequency tracks (left), and three different window functions (remainder). The three window functions use the same equivalent rectangular window lengths.

The graph shows how the lobe patterns in Figures 5.5–5.7 are represented in the STFT magnitude. Side lobes show up as parallel lines to the nearest of the two sinusoids and fall off significantly more quickly for the Hann window than for the rectangular window. Side lobe strength can be thought of as a sort of "noise floor" introduced by block processing and windowing, which attributes a better "signal-to-noise ratio" to the Hann window than the Hann-Poisson-window, and again than the rectangular window.

Since all three windows use the same equivalent rectangular window length, the main lobe widths are very similar. In this sense, main lobe width is purely an artifact of block processing in general, and more or less independent of the window function.

5.2.6 How Window Shape Affects the STFT Phase

STFT phases are notoriously difficult to interpret. Later chapters will show a number of ways of making STFT phases easier to interpret through phase derivatives. However, for simple sinusoids, the STFT phases can be visually understood as the argument of the sinus itself. Figure 5.11 shows STFT phases of two sinusoids, one modulated, one of fixed frequency.

The last panel of Figure 5.11 is easiest to interpret, which shows the STFT phases using the Hann-Poisson window. In the top half, the rate of phase rotation speeds up as the top sinusoid's frequency rises, then slows back down. The bottom half shows the bottom sinusoid's fixed frequency. The figure uses extreme overlaps of 98.4 % to show individual phase rotations.

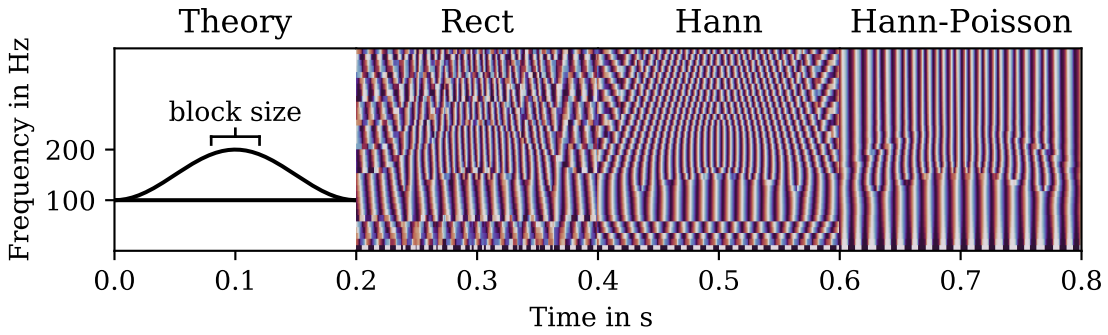


Figure 5.11: STFT phases of two overlapping sinusoids, with their frequency tracks (left), and three different window functions (remainder). The three window functions use the same equivalent rectangular window length. Phases shifted to remove block processing time shift.

The other two panels for the rectangular and Hann window show the same pattern as the Hann-Poisson window at the sinusoids' frequencies, but flipping phases for every other side lobe, in accordance with the phase flips described in sections 5.2.1–5.2.3. Thus, the Hann-Poisson window seems advantageous for phase analysis, as the absence of side lobes and phase flips should make the STFT phases easier to interpret.

Interestingly, the sinusoids' phases show in STFT bins far beyond their frequency. STFT bins clearly show the phase of the strongest, closest signal component, even if these components are of very low magnitude. This property can be useful for attributing individual STFT bins to certain signal components [36].

Between the two sinusoids in Figure 5.11 the STFT phases transitions sharply between the dominance of the higher, modulated sinusoid, and the lower, straight sinusoid. This and sharp either-or of STFT phases bins is a distinct contrast to the gradual superposition of STFT magnitude bins.

In complex real-world signals with additional noise, however, these simple patterns are not as easy to recognize. Additionally, most practical visualizations will use far less overlap, and thus will not resolve individual phase rotations, which further obfuscates the STFT's phase structure.

5.2.7 How Window Overlap Affects the Waveform

Tapering window functions reduce signal amplitudes at the beginning and end of each block. Many window functions, such as the aforementioned Hann and Hann-Poisson window, even reduce amplitudes to zero at these points. This, however, also means that the beginnings and ends of blocks are made invisible to the STFT.

Thus, in order to include all parts of the signal in the STFT, consecutive blocks need to overlap. Naturally, the narrower the window function, the more overlap is necessary. Figure 5.12 shows various overlap percentages for the Hann and Hann-Poisson window.

The figure shows how the sum of Hann windows is a straight line, which means that all parts of the signal are represented with equal amplitude in the STFT. The Hann-Poisson window, on the other hand, does not sum up to a straight line, and therefore emphasizes parts that fall near window centers.

In practice, this means that the Hann-Poisson window should always be used with higher overlaps. While this lessens the effect of the non-uniform window sum, there remains an amplitude modulation artifact, which might emphasize signal partials in harmony with the block rate.

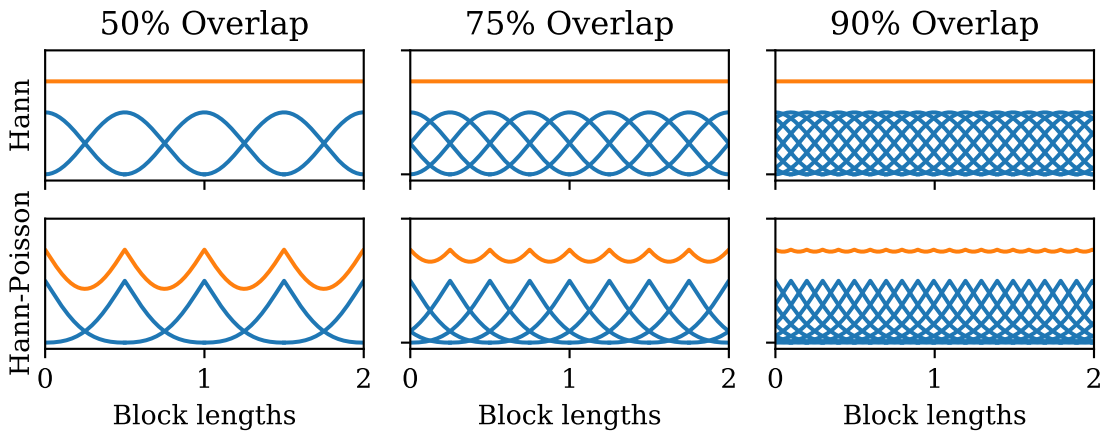


Figure 5.12: Sums of overlapping windows: Blue lines show overlapping windows. Orange line is sum of the windows (not to scale). Hann windows add to a steady sum, while the sum of Hann-Poisson windows is time-variant.

5.3 Visualizing STFTs

The experience of viewing a speech STFT is very different from hearing it. STFTs provide an objective interpretation of speech signals, disentangled along time and frequency in a way our perception is unable to. As such, STFTs offer a different kind of insight into the structure of signals than our ears allow.

In particular, a major drawback to human hearing is that we can only ever experience sounds holistically in real time. There is no way of "zooming in" on a particular detail, or of "freezing" a sound, and analyzing a particular instance on its own beyond repeated listening. By their very nature, audio qualia are characterized only by their flow through time, and never to be experienced in isolation.

The STFT overcomes this human limitation with a visual representation of sound, one we cannot directly experience as hearing but one we can analyze in great detail and independently from the inevitable flow of time.

Perhaps even more importantly, we can use the powerful pattern recognition of our visual system for finding structure in the STFT, even where our ears might struggle. So powerful are our visual pattern matching facilities, in fact, that it is often quite tempting to forego listening to sounds altogether and only focus on visual STFTs instead. This would be folly, however, as visual representations of STFTs are often incomplete in terms of frequency range and resolution, and are missing STFT phases.

Many audible patterns of speech have a structured representation in the STFT and can therefore be reasoned about visually. However, the human visual system is at least as complicated as the auditory system and has many quirks of its own. The rest of this chapter looks at various idiosyncrasies of the human visual system and what they mean for STFTs.

Some of these methods, such as dense sampling (5.3.1), are equally useful for human perception as to machine interpretation. Others, such as color maps 5.3.2, 5.3.3), are merely workarounds for deficiencies of human vision.

5.3.1 Resolution Requirements for Human Viewers

Technically, an STFT can be said to display a complete picture of a sound if every sample in the waveform is represented in at least one block with a reasonable amplitude. All three panels in Figure 5.13

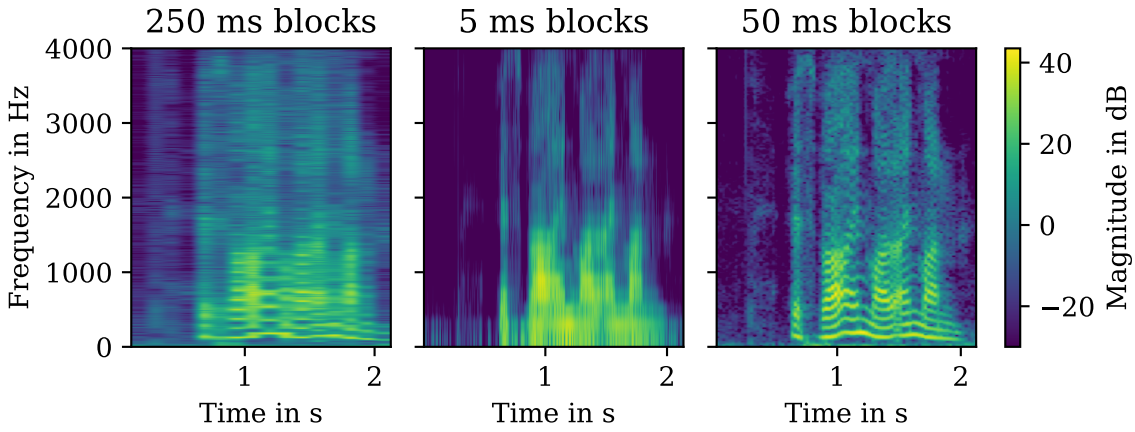


Figure 5.13: Magnitude STFT of a speech signal at different block lengths.

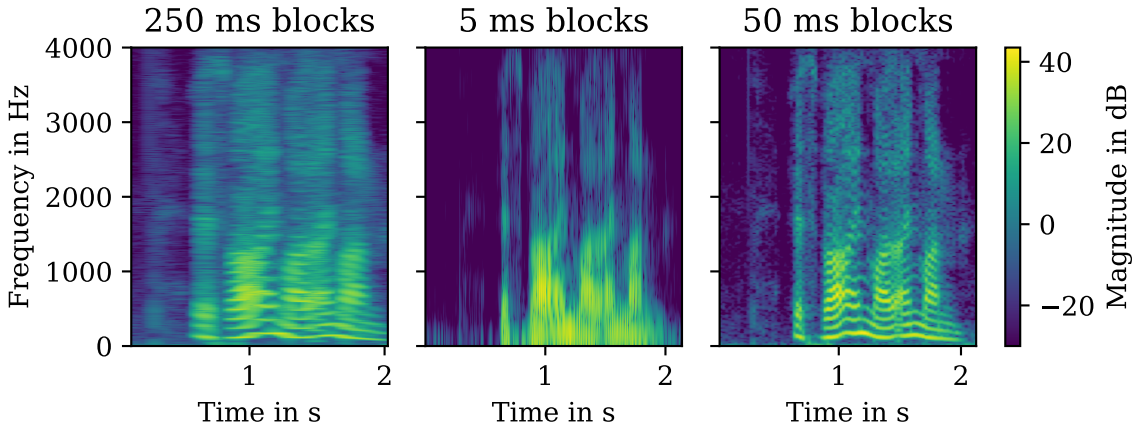


Figure 5.14: Magnitude STFT of a speech signal at different block lengths with oversampling. Same data as in Fig. 5.13, but with denser sampling in time (left), frequency (middle), and both (right).

show such complete STFT magnitudes, yet the first two are nigh-unreadable in comparison to the third.

The problem with Figure 5.13 is in how our visual system processes these images: We interpret these images based on contrasts between light and dark patches, and color differences between areas [81]. However, the strongest contrasts in the first and second panel are not between features of the signal, but in the sharp transitions between neighboring time-frequency bins. In the first panel, we see this as vertical banding along blocks, and in the second panel as horizontal banding along frequencies.

In order to focus our visual perception on the signal features that matter, we therefore need to ensure smoothness between neighboring time-frequency bins. Figure 5.14 shows the same signal as the previous figure at the same block lengths but uses denser sampling to hide the sharp transitions between bins. The left panel is sampled with 90 % block overlap instead of the customary 50 %. The center panel is oversampled in frequency by zero-padding the FFT by a factor of 10. Even the right panel was improved slightly by doubling the overlap and zero-padding.

5.3.2 Perceptually Uniform Color Maps

When drawing STFT graphs, magnitude and phase values are encoded as colors in a two-dimensional image graph. The color-maps used in this translation are of critical importance to our interpretation of the data. In fact, multiple studies have shown how a simple linear gray scale can significantly reduce error rates in medical applications in comparison to badly-designed color maps—despite years of familiarity with the color map in question and no familiarity with the linear gray scale [11].

The basic design criteria for a color map must be to maintain visual contrast proportional to the difference in value between bins [81, 12, 137, 87, 47]. The greater the difference between two values, the greater the visual contrast should be. As a secondary consideration, colors should degrade without artifacts when printed in black and white, or when viewed with color vision deficiencies or badly calibrated monitors or projectors.

In STFTs, specifically, most information is encoded in changes across small visual distances, where human vision is more sensitive to lightness differences than to hue. Color maps for STFTs should therefore primarily scale along lightness, and merely use hue differences as a supplement, but not as a main means for generating contrast. This automatically allows for printing in black-and-white as well.

Furthermore, color vision deficiencies are most common in differentiating between red and green hues, thus hue contrasts in this range should be avoided. Lest one assume that these are minor issues, color vision deficiencies affect a full 8 % of all males [138], or, as Wong puts it in *Points of View: Color Blindness* [167]: "if a submitted manuscript happens to go to three male reviewers of Northern European descent, the chance that at least one will be color blind is 22 percent."

These issues have come to a head in the last few years and have prompted multiple common software packages for signal processing to swap out their antiquated rainbow-style color maps for well-designed, perceptually uniform color maps along the blue-and-green hues[6] [95, 61].

It should also be noted that these considerations only fully apply to ratio scales, where a value can indeed be said to be "twice as high", and therefore map to a twice-lighter color. In a limited sense, such color maps are still useful for interval scales such as STFT phases, or even ordinal scales such as school grades. But they should be avoided for nominal scales, so as not to imply a rank order between different values.

For this reason, this work will use the perceptually uniform color maps *Viridis* for magnitudes, the circular, partly uniform *Twilight* for phases, parts of the uniform color map *Magma* for ordinal non-image graphs, and the perceptually-equal *Tableau* color rotation for nominal, non-image graphs. Figure 5.15 shows samples of all of these color maps, all of which are part of the software package Matplotlib [61], with *Twilight* being a recent addition to Matplotlib by the author of this dissertation[7].

The color map graphs for *Viridis* and *Twilight* follow Kovesi's test image [81], which shows both the full range of colors (left/center) and a high-frequency modulation. In a properly designed color map, the contrast between modulated parts should be uniform across the entire value range, which is the case for the color maps shown here.

5.3.3 A Circular Color Map for Phase STFTs

Phase STFTs contain angle values, which are circular in a range from 0 to 2π. Circular means that the "highest" phase angle of 2π is equal to the "lowest" value of 0. In fact, one could argue that no phase angle is "higher" or "lower" than any other angle, as they differ in *direction*, not magnitude. One should therefore not use a linear color map, such as the aforementioned *Viridis* or *Magma*, for

[6] Documentation of this change in Matlab: `https://matplotlib.org/3.1.1/users/dflt_style_changes.html` and Matplotlib: `https://bids.github.io/colormap/`

[7] See `https://github.com/matplotlib/matplotlib/pull/6254` for the corresponding pull request.

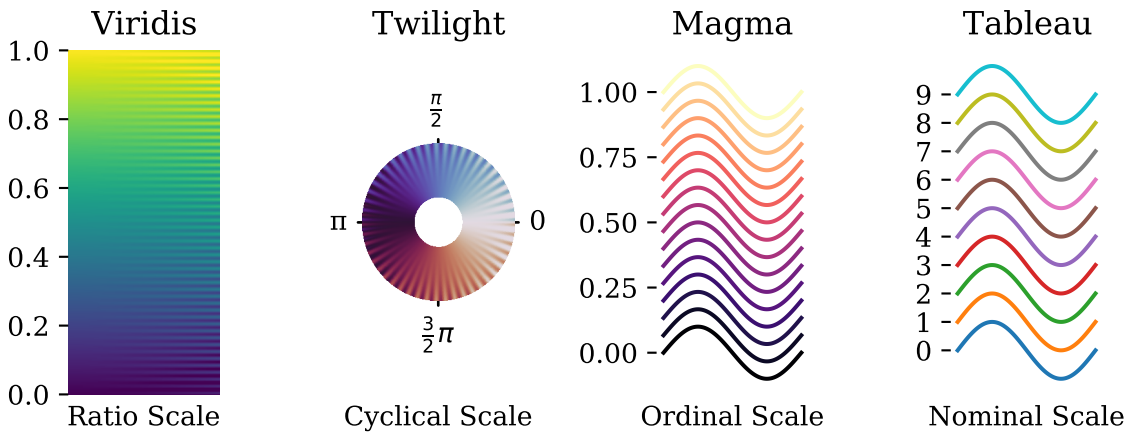


Figure 5.15: The color maps used in this dissertation. The left two panels use Kovesi's test image.

displaying phase angles, as these would introduce a sharp contrast at its minimal/maximal value where there should be a smooth transition. Instead, phase angles should be displayed with a color map that is perceptually uniform around the entire unit circle, which is commonly called a *circular* (or *cyclic*) color map.

Before this work, the only circular color map available in Matplotlib[8] or Matlab[9] was *HSV*, a much-maligned [12] rainbow color map with highly non-uniform contrast across its value range. In particular, *HSV* renders features in the red, green and blue range almost invisible, while yellow and cyan features are shown in great contrast. Moreover, it uses all primary colors and varies in lightness across its value range, which makes it unsuitable for printing or viewing with color vision deficiencies. Clearly, this is suboptimal for displaying STFT phases.

Since *HSV* is so obviously inadequate for STFT phases, most publications instead use a linear color map for displaying phases. This, however, introduces a sharp transition between "maximal" and "minimal" phase values, making the graphs harder to read, and putting undue emphasis on this transition. Figure 5.16 shows the cyclical but non-uniform color map *HSV*, the linear and uniform *viridis*, and the cyclical and uniform *Twilight*, which was designed as part of this work.

To remove the sharp transition between maximal and minimal values, a well-behaved circular color map should use a linear, brightness-scale color map for the value range between 0 and π, and the reverse for π to 0. This would accurately reflect the cyclical nature of phases, without any sharp transition. To remove the resulting ambiguity between the two halves of the unit circle, they should use colors of equal visual weight (lightness), but different hues [81].

Twilight was designed in the CIELAB color space, a color coordinate space that approximates human vision. In this representation, points of equal distance have equal visual contrast. Perceptual uniformity can therefore be ensured by designing the color map along points of constant distance. The color space represents color values in three coordinates: The lightness axis L^* scales from pure black at $L^* = 0$ to pure white at $L^* = 1$, while the two color axes a^* and b^* represent gray as $a^* = b^* = 0$, and green-red components as negative and positive values in a^*, and blue-yellow components in b^*.

Figure 5.17 shows *Twilight*'s color progression in the a^*/b^* hue plane and along the L^* lightness axis. In lightness, *Twilight* falls from off-white to off-black between 0 and π, then rises back to off-white at 2π. To differentiate the falling side from the rising side, they trace through red and blue, respectively. However, by ensuring that the rising and falling sides are exactly symmetrical in lightness,

[8] https://matplotlib.org/tutorials/colors/colormaps.html#cyclic
[9] https://de.mathworks.com/help/matlab/ref/hsv.html

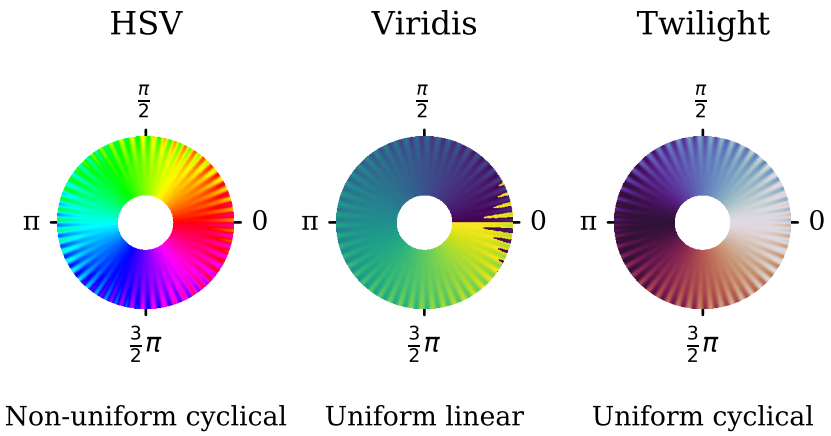


Figure 5.16: Color maps used for displaying cyclical data. This work introduces and uses *Twilight*.

both sides have equal visual weight. Red and blue were chosen as they remain discernible even with most color vision deficiencies.

The perceptual deltas in the bottom part of Figure 5.17 show the distance between *Twilight*'s color coordinates in CIELAB, and therefore an approximation of perceptual contrast. The symmetric lightness scaling between the first and second half of the color map introduces conspicuous low-contrast regions at the lightness reversals near 0 and π. The lightness reversals themselves are a necessary feature of a lightness-scale circular color map, as detailed earlier. To soften the L^* reversals somewhat, color coordinates were smoothed near the lightness reversals [81].

A slightly increased visual weight of the mostly-black and mostly-white regions remains near 0 and π in *Twilight*. However, this can in fact be advantageous for STFT phases, as areas of zero phase are often of particular importance and worth emphasizing a little. For this reason, slightly non-uniform lightness-scaling was deemed superior to an alternative design that uses perfectly uniform equal-lightness hue-scaling for visualizing STFT phases.

Figure 5.18 shows the color map with various simulated red-green vision deficiencies and in greyscale, for printing. By choosing colors outside of the red-green range, *Twilight* degrades without artifacts for most kinds of vision deficiencies. In greyscale, the ability to discriminate between the two halves of the unit circle is lost, but no further artifacts are introduced, and the two halves remain perfectly symmetrical.

In summary, *Twilight* is a perceptually uniform color map for cyclical phase data, much like *Viridis* and *Parula* are for ratio-scale data, designed according to the best practices published by [81].

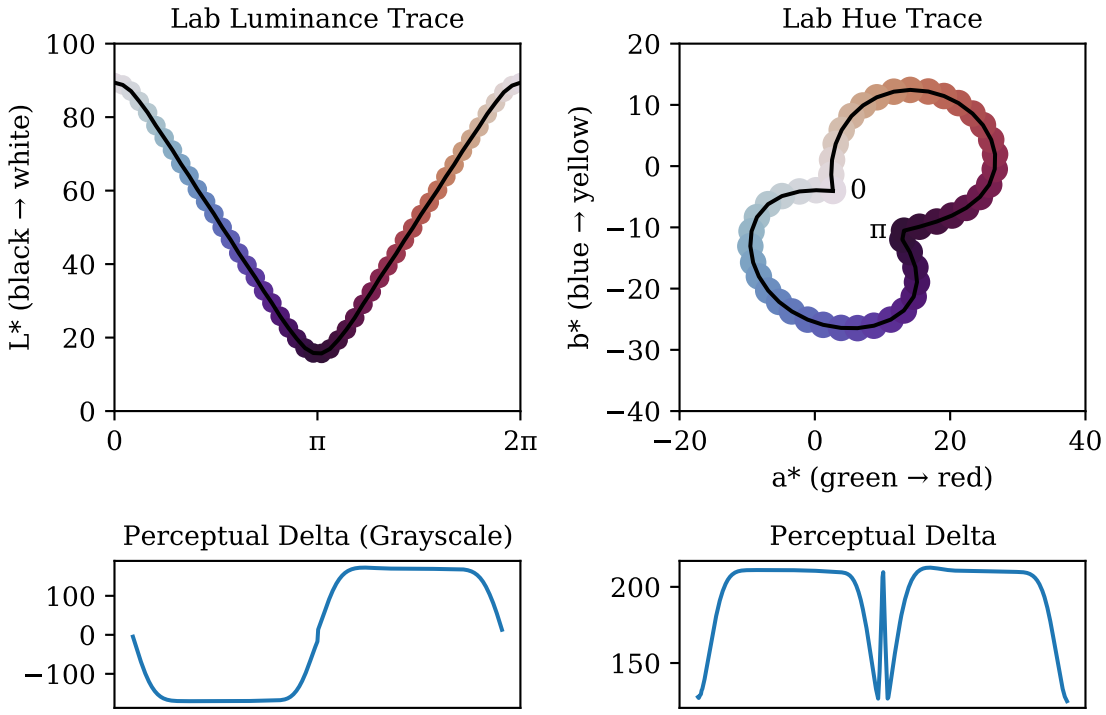


Figure 5.17: Traces of *Twilight* in the CIELAB color space, as well as perceptual contrast between neighboring points in full color and greyscale. Left side shows *Twilight*'s lightness progression and contrast, right side shows hue progression and contrast.

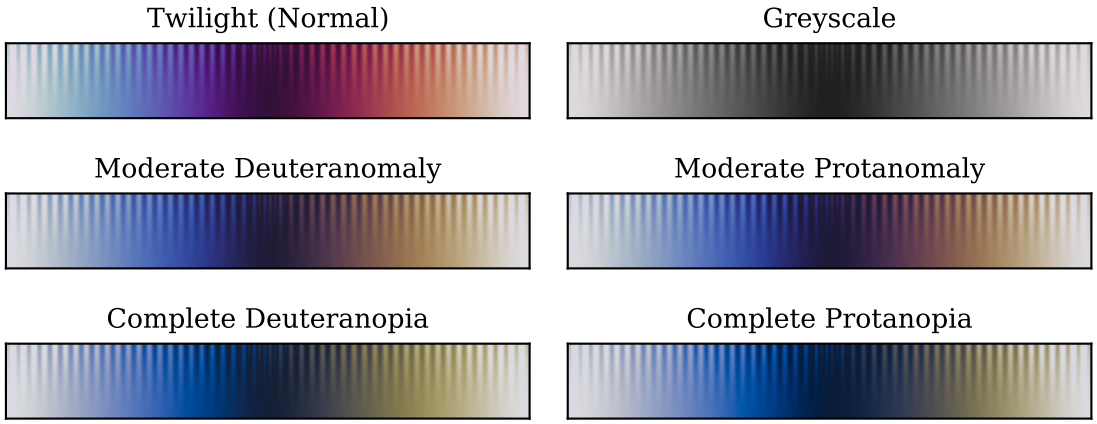


Figure 5.18: *Twilight* with various degradations: simulated red-green color vision deficiencies (bottom four graphs) and greyscale printing.

Chapter 6

Spectral Derivatives

Speech is a dynamic signal that is constantly changing over time and frequency. Signal parts span continuous areas of time and frequency, and with smooth transitions in both dimensions. It should therefore be no surprise that a great deal of information is encoded not only in steady-state sections of STFT spectra, but also the rates of change, or derivatives, of those spectra.

Well-known instances of such interesting derivatives are group delay and instantaneous frequency, which are loosely defined as the frequency and time derivatives of the phase spectrum, respectively. Group delay and instantaneous frequency can be interpreted as the time and frequency offset of the signal part that dominates a particular STFT bin. This offset information can be used to reassign the spectral energy to dominant transients as a way of sharpening spectral features [39, 116, ch. 5].

Perhaps more importantly, phase derivatives are a convenient way of side-stepping phase ambiguities due to wrapping and windowing delays. They are thus a particularly legible variant of STFT phases, which will be used in Chapter 8 to find harmonic structures in STFT phases.

Derivatives are not only useful for STFT phases, but for magnitudes as well. However, this information has historically found little application, although many of the phase derivations' uses could be implemented with magnitude derivatives as well. Reassignment, in particular, could just as well be implemented by hill-climbing magnitude derivatives, if a suitably smooth window function were used.

On the topic of window functions, much of the interpretation of STFT derivatives hinges on the choice of window functions, which govern both the slopes of the magnitude, and the clarity of phase derivations. Section 6.3 delves more into detail on this topic and the various interpretations of STFT derivatives. In general, the Hann-Poisson window with its smooth slope without intermittent zeros is a good choice for calculating STFT derivatives.

The following two sections 6.1 and 6.2 explores in detail two ways of calculating spectral derivatives, and their respective strengths and weaknesses.

6.1 Difference Method

A simple and straight-forward way of calculating spectral derivatives is the logarithmic magnitude and phase difference between neighboring STFT bins. While conceptually simple, the choice of time step Δt and frequency step Δf needs to select bins that are part of the same slope in order for the difference method to be meaningful. This often requires oversampling in time or frequency to obtain useful results.

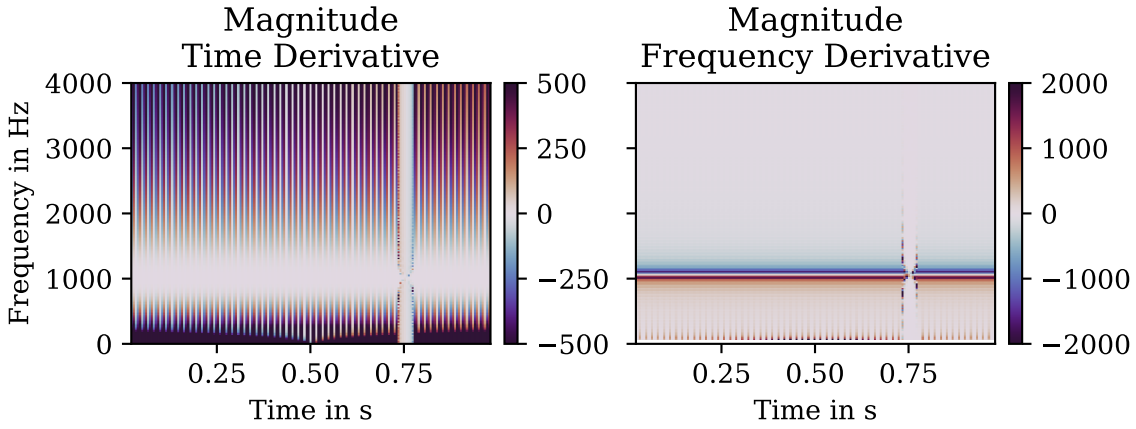


Figure 6.1: Derivatives of the STFT magnitude of a sinus at 1000 Hz and a delta impulse at 0.75 s analyzed with the Hann-Poisson window. Scaled to dB per $f_s/2$ and dB per $2N/f_s$, respectively.

Magnitude Derivatives

Both the STFT magnitude and its derivatives are most useful in the logarithmic domain, where signal spectra mix additively. Thus, to calculate the STFT magnitude derivatives, the difference is calculated between logarithmic magnitudes $S_{\mathrm{dB}}(t,f) = 20\log_{10}|S(t,f)|$:

$$\frac{\Delta S_{\mathrm{dB}}(t,f)}{\Delta t} = \frac{1}{\Delta t}\Big(S_{\mathrm{dB}}(t+\frac{\Delta t}{2},f) - S_{\mathrm{dB}}(t-\frac{\Delta t}{2},f)\Big) \tag{6.1}$$

$$\frac{\Delta S_{\mathrm{dB}}(t,f)}{\Delta f} = \frac{1}{\Delta f}\Big(S_{\mathrm{dB}}(t,f+\frac{\Delta f}{2}) - S_{\mathrm{dB}}(t,f-\frac{\Delta f}{2})\Big) \tag{6.2}$$

Neighboring STFT magnitude bins are usually smooth enough to be used for calculating magnitude frequency derivatives, as STFT magnitude spectra are already smoothed by windowing. Regardless, both time and frequency derivatives benefit from a smooth window function without zeros, such as the Hann-Poisson window.

Figure 6.1 shows magnitude derivatives of the sum of a sine and click signal. Graphs of these simple signals are provided here as a guide to how the derivatives represent common signal components, so as to help in interpreting more complex graphs later on.

In the frequency derivatives, sinusoids are represented as rising slopes towards the sinusoid's frequency and falling slopes thereafter. White components such as clicks do not have any frequency slope and are therefore zero. In the time derivative, clicks are represented with similar positive-negative ramps and sinusoids are zero.

Phase Derivatives

Phase derivatives are particularly interesting as they have a simple interpretation that is not obvious for the phase spectrum itself. Considering a sinusoidal signal component

$$s(t) = \mathrm{e}^{ift},$$

whose phase is accordingly

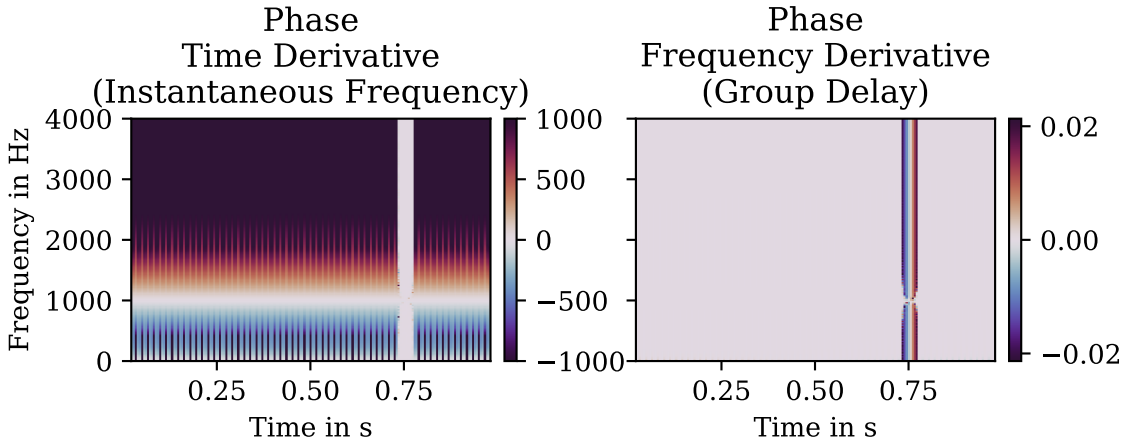


Figure 6.2: Derivatives of the STFT phase of a sinus at 1000 Hz and a delta impulse at 0.75 s analyzed with the Hann-Poisson window. The time derivative is clipped at ±1000 Hz for readability, but truly extends linearly to $f_s/2$. Instantaneous frequency in Hz, group delay in s.

$$\angle S_{s(t)}(t, f) = ft,$$

and the derivatives are f and t, respectively called *Instantaneous Frequency* and *Group Delay*.

To calculate phase derivatives from STFTs, phases of neighboring STFT bins must differ by less than 2π to avoid phase wrapping ambiguity. This can only be assured if the time delay Δt is one sample and the frequency delay $\Delta\omega$ is one bin.

$$\frac{\Delta\angle S(t,f)}{\Delta t} = \frac{1}{\Delta t}\Big(\angle S(t+\frac{\Delta t}{2}, f) - \angle S(t-\frac{\Delta t}{2}, f)\Big) \tag{6.3}$$

$$\frac{\Delta\angle S(t,f)}{\Delta f} = \frac{1}{\Delta f}\Big(\angle S(t, f+\frac{\Delta f}{2}) - \angle S(t, f-\frac{\Delta f}{2})\Big) \tag{6.4}$$

Figure 6.2 shows the instantaneous frequency and group delay for a sinus and click signal, scaled to Hz and seconds, respectively.

The instantaneous frequency represents sinusoids as linear ramps in the frequency direction, with a zero crossing at the sinusoid's frequency. Clicks are represented as zeros. The group delay represents clicks as linear ramps in the time direction with a zero crossing at the click's time, and sinusoids are zeros. This is exactly the opposite in the magnitude derivatives.

Both the instantaneous frequency and group delay are still phase differences that wrap at 2π. At one sample time delta, 2π corresponds to $f_s/2$ and at one bin frequency delta it is $N/2$, and therefore does not occur for steady signals such as the click and sinus in Figure 6.2.

However, larger Δt or Δf might be desirable for additional robustness to small signal disturbances such as noise, which would introduce phase wrapping. For speech signals with base frequencies ≥ 80 Hz, care should be taken that no wrapping occurs between neighboring speech harmonics by restricting Δt to $\Delta t \leq 80 \cdot N_{\text{FFT}}/f_s$. To restrict wrapping to frequencies ≥ 4000 Hz, Δf should likewise be limited to $\Delta f \leq 4000/f_s/2$.

6.2 Window Method

Instead of relying on differences between STFT bins for calculating spectral derivatives and dealing with the phase wrapping ambiguities, the derivatives can instead be calculated directly for each bin [92, 78]. This neatly sidesteps all wrapping issues but can require more computation and does not have the option of using longer step sizes.

Time Derivative

Given that the definition of the Fourier transform in Section 5 as

$$S(t,f) = \int_{-\infty}^{\infty} s(n) \cdot w(n-t) \mathrm{e}^{-i2\pi f n} \,\mathrm{d}n \tag{6.5}$$

references the time index t only in the window function $w(n-t)$, its time derivative need only be applied to the window function as well:

$$\frac{\mathrm{d}}{\mathrm{d}t} S(t,f) = \int_{-\infty}^{\infty} s(n) \cdot \frac{\mathrm{d}}{\mathrm{d}t} w(n-t) \cdot \mathrm{e}^{-i2\pi f n} \,\mathrm{d}n \,. \tag{6.6}$$

Thus the derivative of the spectrum can be derived merely by replacing the window function $w(n)$ with its time derivative $\frac{\mathrm{d}}{\mathrm{d}t} w(n)$.

The spectrum derivative decomposes into its magnitude and phase components as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} S(t,f) &= \frac{\mathrm{d}}{\mathrm{d}t} \Big(|S(t,f)| \cdot \mathrm{e}^{i \angle S(t,f)} \Big) && (6.7) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} |S(t,f)| \cdot \mathrm{e}^{i \angle S(t,f)} + |S(t,f)| \cdot i \frac{\mathrm{d}}{\mathrm{d}t} \angle S(t,f) \, \mathrm{e}^{i \angle S(t,f)}. && (6.8) \end{aligned}$$

Dividing by $S(t,f)$ removes the exponentials

$$\frac{\frac{\mathrm{d}}{\mathrm{d}t} S(t,f)}{S(t,f)} = \frac{\frac{\mathrm{d}}{\mathrm{d}t} |S(t,f)|}{|S(t,f)|} + i \frac{\mathrm{d}}{\mathrm{d}t} \angle S(t,f) \tag{6.9}$$

and thus gives the time derivative of the phase spectrum as

$$\frac{\mathrm{d}}{\mathrm{d}t} \angle S(t,f) = \Im \left(\frac{\frac{\mathrm{d}}{\mathrm{d}t} S(t,f)}{S(t,f)} \right) \tag{6.10}$$

and time derivative of the magnitude spectrum as

$$\frac{\mathrm{d}}{\mathrm{d}t} |S(t,f)| = |S(t,f)| \cdot \Re \left(\frac{\frac{\mathrm{d}}{\mathrm{d}t} S(t,\omega)}{S(t,f)} \right), \tag{6.11}$$

where $\Im(\cdot)$ and $\Re(\cdot)$ extract the imaginary and real part of a value, respectively.

However, the interesting part is the derivative of the logarithm of the magnitude spectrum. Since the derivative of the logarithm of a function is the derivative divided by the function, this can be calculated as

$$\frac{\mathrm{d}}{\mathrm{d}t} \log |S(t,f)| = \frac{\frac{\mathrm{d}}{\mathrm{d}t} |S(t,f)|}{|S(t,f)|} = \Re \left(\frac{\frac{\mathrm{d}}{\mathrm{d}t} S(t,f)}{S(t,f)} \right). \tag{6.12}$$

Frequency Derivative

The frequency derivative of the Fourier transform references f only in the exponential, and can thus be applied to the exponential only:

$$\frac{\mathrm{d}}{\mathrm{d}f}S(t,f) = \int_{-\infty}^{\infty} s(n)\cdot w(n-t)\cdot \frac{\mathrm{d}}{\mathrm{d}f}\mathrm{e}^{i2\pi fn}dn \tag{6.13}$$

$$= \int_{-\infty}^{\infty} s(n)\cdot i2\pi n w(n-t)\cdot \mathrm{e}^{i2\pi fn}dn \tag{6.14}$$

where, again, the derivative is accomplished merely by replacing the window function $w(n)$, this time with $i2\pi n w(n)$.

Following the same line of reasoning as the time derivatives, this gives

$$\frac{\mathrm{d}}{\mathrm{d}f}\angle S(t,f) = \frac{1}{2\pi}\Re\left(\frac{\frac{\mathrm{d}}{\mathrm{d}f}S(t,f)}{S(t,f)}\right) \tag{6.15}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}f}\log|S(t,f)| = \frac{1}{2\pi}\Im\left(\frac{\frac{\mathrm{d}}{\mathrm{d}f}S(t,f)}{S(t,f)}\right) \tag{6.16}$$

Figure 6.3 shows the magnitude and phase derivatives of the same sine-and-click signal as in the previous chapter using the window method. These graphs are almost identical to the derivatives using the difference method. However, the window method is an instantaneous measure of each STFT bin, whereas the difference method is calculated over a small time/frequency delta. Depending on the scale of the signal part to be analyzed, either method can be useful: Larger structures, especially in the presence of noise, are often more visible with a slight averaging imposed by larger time/frequency deltas and the difference method, while smaller structures are better resolved with the window method.

6.3 Applications

It is safe to say that the majority of all speech processing algorithms focus on rough shapes within the STFT magnitude. There are long-running arguments about the usefulness of additional data, such as STFT phases [113, 114, 91] or high-resolution spectra [8, 121, 34, 170, 157]. Yet without a doubt, there is additional information stored in phase data, as without it, no exact recreation of a speech signal is possible.

Part of the problem with interpreting phase data, and details within the STFT magnitude is a lack of visualization tools to make sense of them. In particular, phase spectra without frequency-shifting to the window center and with window spectra full of zero crossings are full of distracting phase reversals that can make it difficult to understand their content. Phase derivatives are a way of mitigating these issues; by focusing on the differences between neighboring STFT bins, many effects of phase wrapping and zero crossings can be glossed over.

Figure 6.4 shows such magnitude and phase derivatives for the same speech signal as the introductory Figure 5.1. All the features of the STFT magnitude are clearly visible in the derivatives as well: Onsets and offsets as vertical red or blue bands in the magnitude time derivative or phase frequency derivative, and sinusoids as horizontal red or blue bands in the magnitude frequency derivative or phase time derivative.

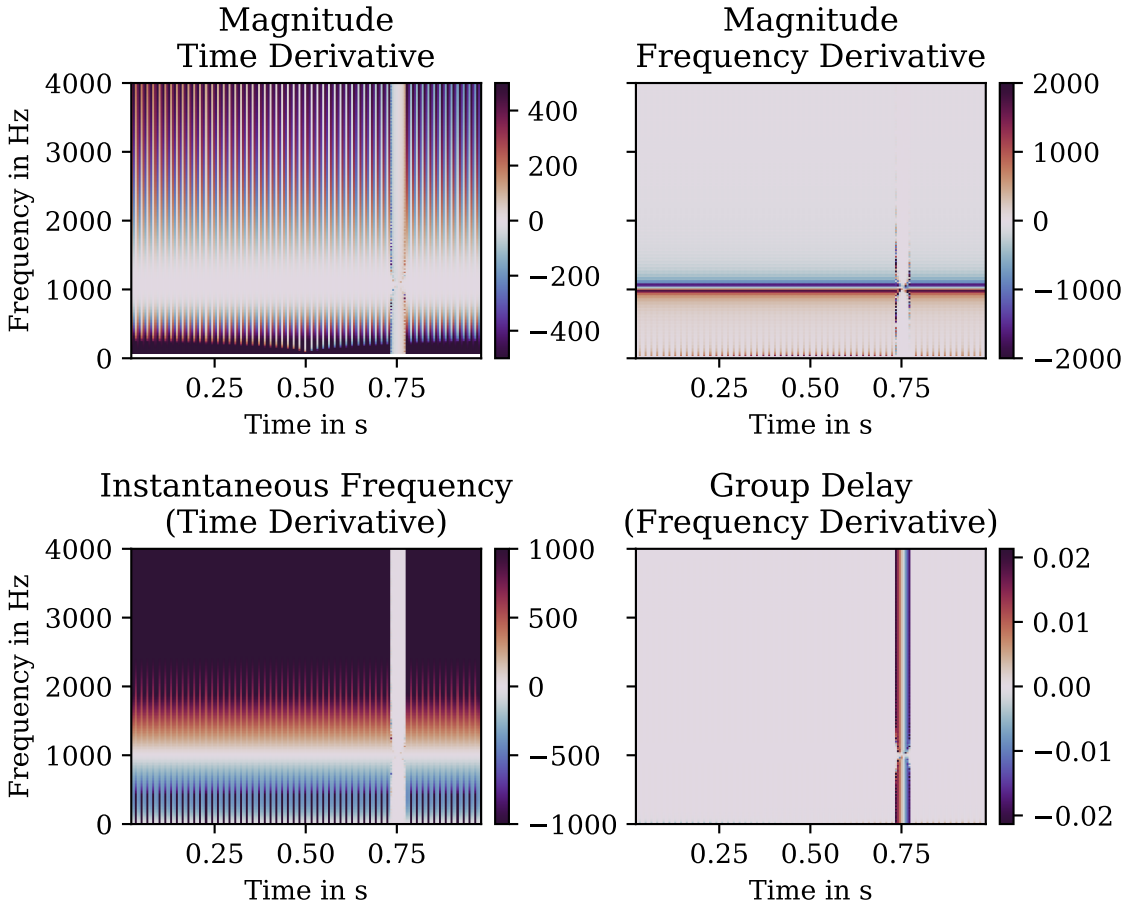


Figure 6.3: Time derivative (left) and frequency derivative (right) of the STFT magnitude (top) and STFT phase (bottom) of a delta impulse and a sinus STFT analyzed with the Hann-Poisson window. The time derivative is clipped at ±1000 Hz for readability, but truly extends linearly to $f_s/2$. Instantaneous frequency in Hz, group delay in seconds. Magnitude derivatives in dB per $f_s/2$ and dB per $2N/f_s$.

Additionally, phase derivatives have a clear interpretation as the time and frequency origin of an STFT bin's energy. Due to smearing and windowing, a signal component's energy is spread over multiple STFT bins. Energy is spread over time as multiple blocks sample sections of each signal component, attenuated by the window shape. Energy is spread over frequency as signal components are convolved with the window spectrum. While these smearing operations generally alter the component magnitude, its phase will still rotate with the original frequency and can be recovered from the phase time derivative. The precise moment of a click is similarly smeared across STFT time but can be recovered from the group delay. This information can also be used to reassign these components back to the original time/frequency location. In a similar way, phase derivatives are frequently used to increase the precision of magnitude spectral maxima [36, 34, 92, 121].

Phase derivatives can also be used as a signal feature in their own right. Figure 6.4 shows the instantaneous frequency and group delay of a speech signal, and clearly shows speech harmonics in the magnitude frequency derivative and instantaneous frequency as horizontal red-and-blue bands. This has been of particular use for fundamental frequency estimation [38, 20] and speech enhancement [82].

Figure 6.5 shows a shorter section of voiced speech in greater detail. The graphs show how the

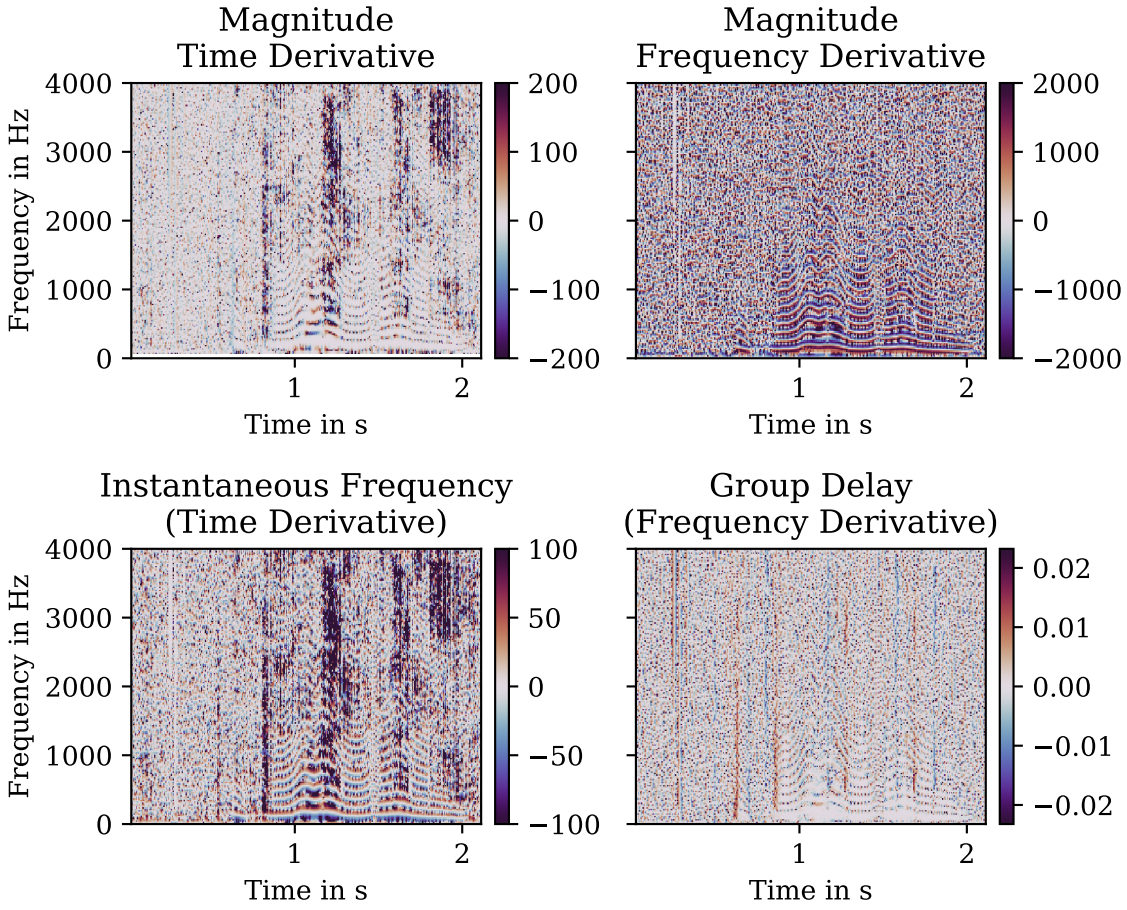


Figure 6.4: Time derivative (left) and frequency derivative (right) of the STFT magnitude (top) and STFT phase (bottom) of a speech signal STFT with the Hann-Poisson window. The time derivative is clipped at ±100 Hz for readability, but truly extends linearly to $f_s/2$. Instantaneous frequency in Hz, group delay in seconds. Magnitude derivatives in dB per $f_s/2$ and dB per $2N/f_s$.

STFT magnitude shows harmonics as a comb-like pattern of maxima, whereas the instantaneous frequency shows harmonics as a sawtooth-like pattern of negative-to-positive ramps around each harmonic. The window method shows more time detail, but is more susceptible to small noisy disturbances than the difference method.

Magnitude derivatives have not found much application in signal processing, yet. But since their shape is very similar to phase derivatives, they could be used in a similar way for reassignment[1], or indeed for reconstructing missing phase data.

[1]since exact time/frequency offsets are difficult to obtain from magnitude derivatives, iterative hill-climbing would have to be used instead.

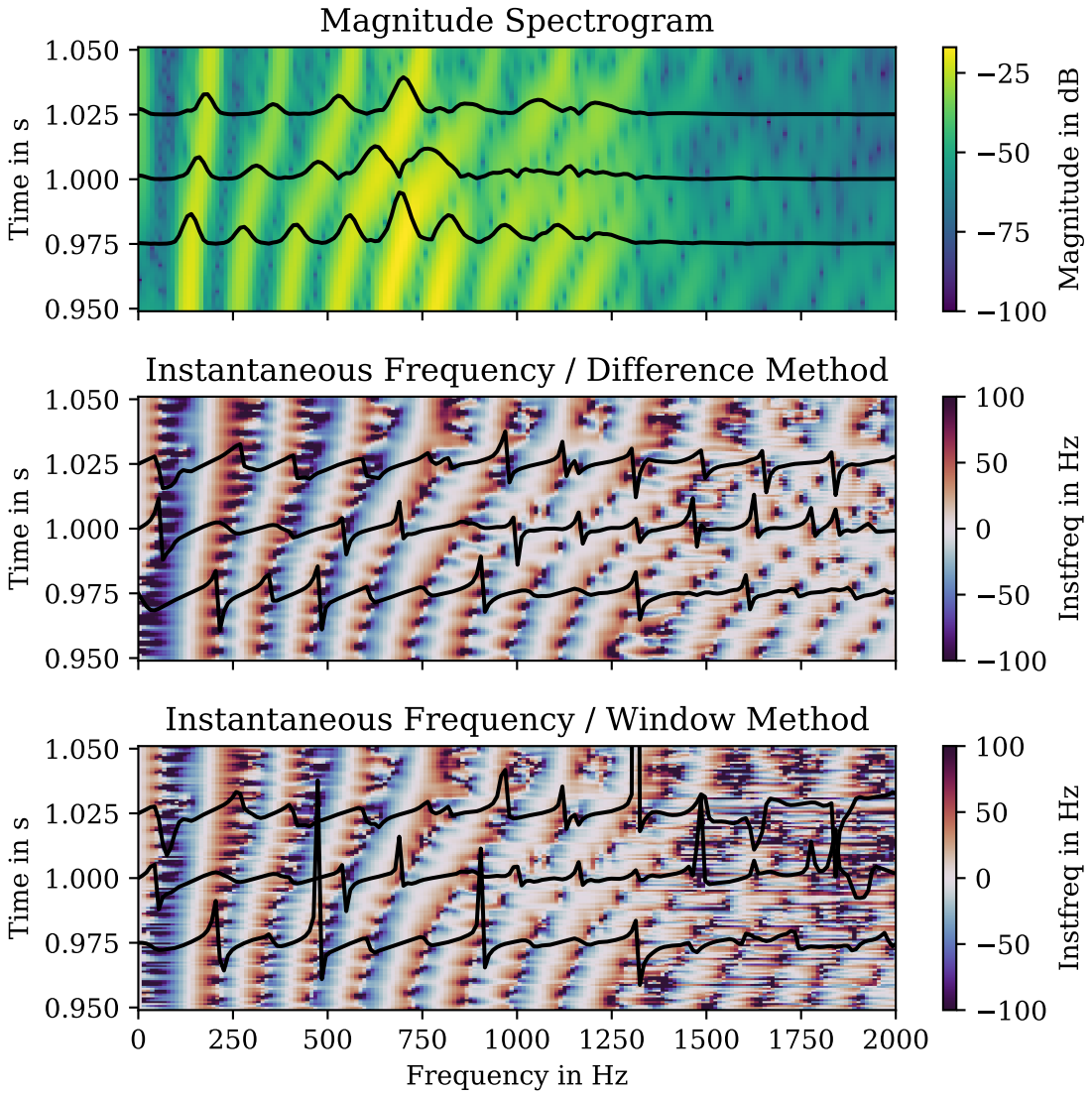


Figure 6.5: Harmonic patterns of a speech signal in the STFT magnitude (top), instantaneous frequency with the difference method at $\Delta t = {}^{100}/_{f_s}$ (center), and instantaneous frequency with the window method. Overlayed on top are black graphs of single spectra at 0.975 s, 1.0 s, and 1.025 s (not to scale, not in dB in case of magnitude STFT).

Chapter 7

Conclusions

This chapter introduced the basic mathematical and visual framework used throughout this dissertation for analyzing audio signals. Just as speech itself is a complex signal with many a different facet, so its analysis must also be done from multiple different viewpoints to obtain a complete picture.

The foundation of most of these perspectives is the short-time Fourier transform, an ingenious way of unfolding signals across time and frequency, and of revealing patterns otherwise hidden in their interplay. While not entirely unlike our own perception of speech as patterns both rhythmic and tonal, the STFT's true strength is how its many parameters can be adapted for the various challenges of speech analysis.

The STFT window function, in particular, plays a pivotal role in determining the kinds of features that can be easily resolved. For STFT magnitudes, a closer look was taken at the Hann window, which represents a reasonable trade-off between time and frequency resolution. For the phase domain, a different kind of window was found desirable that does not include phase reversals, in the form of the Hann-Poisson window.

To make these tools available to researchers, they must be mapped onto a representation that can be reasoned about visually. However, the human visual system is no less complex than its auditory side, and visualizations must therefore be constructed with just as much care as audio signals to avoid misinterpretations and biases. While these ideas have lead to the development of new, perceptually motivated color maps for STFT magnitudes in recent years, the same had not yet been done for STFT phases. This was resolved by introducing a new, perceptually uniform color map for STFT phases.

But windowing and visualization techniques only make visible the core of the problem with interpreting STFT phases: that its phase angles are circular in nature and cannot be visualized or reasoned about the same way as scalar magnitudes. To make sense of STFT phases, the notion of spectral derivatives was introduced in two variants. In the derivatives, at last, phases reveal some of their patterns to observers and algorithms. Surprisingly, these derivation techniques proved useful not only for phase angles, but for STFT magnitudes as well.

Thus we now have assembled a toolbox of techniques and interpretations for the following chapters. Where all previous discussions focused on theory and simple examples, we can now turn to the real world and start solving concrete problems, chiefly among them the analysis of speech signals and the search for its pitch.

Part III

Estimating the Fundamental Frequency of Noisy Speech

Wherein we introduce our first major contribution: An algorithm for estimating the fundamental frequency of voiced speech. This algorithm combines features from the magnitude spectrum and the phase spectrum into a voicing activity measure that gives an a posteriori probability for whether a block of audio data is voiced at a given candidate fundamental frequency. In contrast to similar algorithms, this is a true probability that can reject or accept any candidate frequency. Our algorithm can thus reject ambiguous estimates where simpler algorithms would mis-estimate instead.

A thorough evaluation section analyzes our algorithm's performance with an uncommonly large dataset of speech and noise recordings, and in comparison with notable reference estimators. We learned from this study that comparisons are surprisingly difficult. A large part of the remainder of this dissertation is motivated from this issue and will expand on the intricacies and problems of comparing fundamental frequency estimation algorithms, and how to judge the accuracy of their estimates.

Chapter 8

A Fundamental Frequency Estimation Algorithm

Abstract

The fundamental frequency of the human voice is an important feature for various speech processing applications such as speech enhancement, speech separation, and speech compression algorithms. This chapter presents an algorithm that probabilistically combines features from the magnitude spectrum and the phase spectrum. It then derives a direct pitch confidence measure, which avoids both octave errors and ambiguous estimates. The algorithm estimates relatively few frames as voiced, but remains reliable even with high levels of noise. These characteristics are examined with synthetic tone complexes and a large, freely available corpus of speech and noise recordings.

8.1 Introduction

Speech and language are fundamental to human communication. It encodes information as rapid sequences of sounds that are produced by acoustically exciting the vocal tracts. If this excitation is periodic, it produces a characteristic sound that we call the human voice [131]. This sound can vary in its intonation and pitch within a certain range of spectral envelopes and fundamental frequencies. Despite humans' ease in identifying and characterizing this sound, technical analysis remains difficult. Speech analysis is thus still an active area of research, with various applications such as speech enhancement, speech compression and modification, and the musical analysis of singing voices [21]. This chapter presents a new algorithm for estimating one of the defining properties of the human voice: its fundamental frequency.

We perceive short segments of the human voice as having a single pitch, but their spectra are made up of a series of harmonically related frequencies, each at an integer multiple of a fundamental frequency. This regular pattern is used for algorithmically detecting voiced speech, and for estimating its fundamental frequency. It is detectable in the frequency domain as evenly spaced tonal components, or in the time domain as a regularly repeating waveform.

In the time domain, such repetition can be detected by comparing short signal blocks at different starting times and estimating the fundamental frequency as the inverse time difference between two matching blocks [21]. This method has been implemented in a number of pitch estimation algorithms (or pitch determination algorithms, PDAs), which typically use variants of correlation [140] or differences [130] for comparing signals. Improvements include various pre- and post processing steps and machine learning techniques for improving pitch tracks [9, 150, 42, 24], and joint estimation in multiple frequency bands [65, 86, 151], where even non-harmonic signals at low signal-to-noise ratios (SNR)

remain locally periodic.

In the frequency domain, the fundamental frequency of the human voice can be estimated by comparing comb-like spectral templates at various fundamental frequencies with the short-time signal spectrum, and selecting the best match [106]. However, in real voice recordings, windowing and minute pitch shifts distort spectral peaks around the true comb pattern. More recent algorithms account for such inaccuracies by replacing the Dirac comb teeth with wider peaks [94]. Further improvements include various pre-processing steps and machine learning techniques for improving pitch tracks [73, 22, 50]. A separate branch of research derives their models directly from a theoretical signal model [126, 21, 104, 51] and confirms that comb-like patterns are indeed optimal for detecting pitched signals in noisy magnitude spectra.

However, both spectral combs and temporal self-similarity detect not only the fundamental frequency itself, but also higher and lower harmonics. This can lead to octave errors in the estimated pitch. Some algorithms suppress these errors by inserting negative comb teeth between the positive harmonic comb teeth, so that correlation maximizes the harmonic energy and simultaneously minimizes the subharmonic energy [149, 45]. Nevertheless, octave ambiguity is an inherent property of periodic signals in the time domain and the magnitude spectrum.

Some signal representations do not exhibit any octave ambiguity. One such representation is the phase spectrum and its derivatives, where harmonics are easier to separate from non-harmonic noise. As such, phase spectrum derivatives have been shown to be a meaningful representation of the human voice for pitch estimation [38, 114, 20]. However, the phase spectrum does not contain any loudness information, which makes fundamental frequency estimation susceptible to noise, and voice activity determination (VAD) difficult.

When comparing pitch estimation algorithms with human hearing, neurological research shows that human perception of sound integrates data from multiple domains that include information from both the magnitude spectrum and the phase spectrum, as well as numerous other sources [99, 160, 98, 54, 111]. In algorithmic terms, this indicates that combining PDAs from different domains could lead to higher estimation accuracy, particularly in the presence of noise.

The results of a PDA will depend critically on which parts of the speech are assumed to be pitched. It is therefore necessary to include or reference a VAD algorithm. When many speech parts are accepted as voiced, errors in pitch determination can be expected to be higher than when only the most salient voiced speech parts are selected.

Typically, these VAD algorithms use a different combination of features than the associated PDAs [93, 58, 150, 151, 42, 45]. This can be a problem for voiced signals with no unambiguous fundamental frequency, as in the presence of multiple pitches or fricatives. The accuracy of fundamental frequency estimation could therefore be improved if the VAD excluded not only unvoiced speech, but ambiguous estimates in general.

Today's state-of-the-art PDAs for human voices as outlined in the comparison studies [3, 151, 45, 50, 148, 51] favor estimators in either the time domain or the magnitude frequency domain with separate VAD algorithms such as *YIN* [24], *PEFAC* [45], *RAPT* [150], or *MBSC* [151]. Our proposed algorithm improves on this by combining multiple signal representations for both fundamental frequency estimation and VAD. As we will demonstrate, our VAD only selects unambiguously voiced speech, and thus produces sparse but highly reliable fundamental frequency estimates.

The following chapter will present a pitch estimation feature in the magnitude spectrum, followed by a feature in a derivative of the phase spectrum, and finally a method for combining these measures into a pitch confidence measure. All of these features will be illustrated with a simple, clean speech recording, whose spectrogram and fundamental frequency track is shown in Figure 8.1. The chapter will end with some notes on implementation. The next chapter will evaluate various aspects of this algorithm with real speech and noise recordings, and synthetic tone complexes in white noise, and discuss the results with respect to the aforementioned reference PDAs. Finally, this chapter will

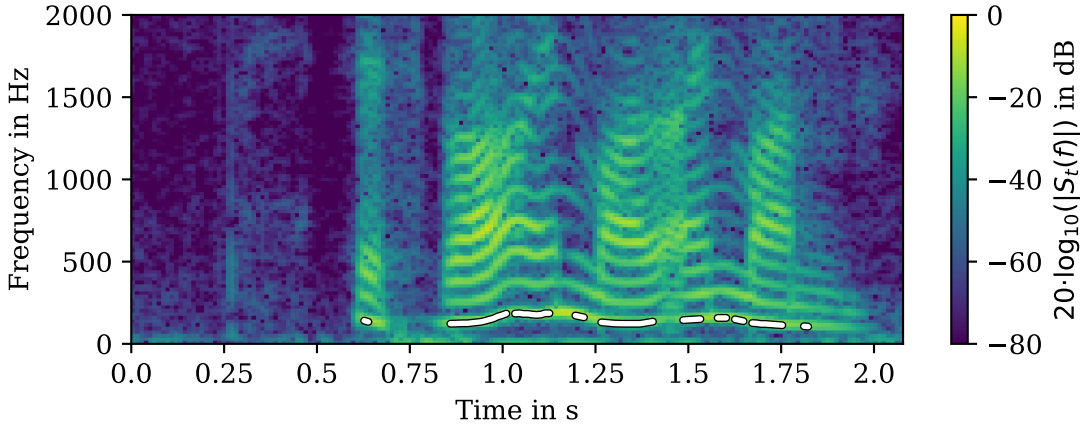


Figure 8.1: Magnitude spectrogram of a clean speech signal that will be used repeatedly for examples. The fundamental frequency is visible as the lowest sinusoidal track, and harmonics are visible as parallel tracks. A fundamental frequency estimate is highlighted as an outlined white line.

conclude with a summary of our main findings.

8.2 Proposed Algorithm

As outlined in the introduction, both the magnitude and phase spectrum can be used to estimate the fundamental frequency of the human voice. Indeed, a combination of these two signal representations seems advantageous in several ways: It uses more of the available information in the signal, and it can use phase spectral information to account for octave ambiguities in the magnitude spectrum. Finally, it can use the magnitude spectral information to help the phase spectrum discern salient parts from non-salient ones.

The presented algorithm integrates data from the short-time magnitude and phase spectrum, and is henceforth called *MAPS*, the *Magnitude and Phase Spectrogram* based fundamental frequency estimator. It implements an a posteriori combination of two features: A soft comb in the short-time magnitude spectrum, and a sawtooth-comb in the time-derivative of the short-time phase spectrum. The resulting pitch confidence measure inherently not only estimates the most probable pitch of each frame, but also provides a measure of confidence in that estimation, which excludes ambiguous estimates. This comparatively conservative approach to VAD makes the algorithm highly precise[1], albeit at the cost of its recall[2].

8.2.1 Voice in the Magnitude Spectrum

The fundamental frequency and every harmonic in a voiced tone complex create peaks in the magnitude spectrum, forming a comb pattern with comb teeth at regular intervals. In the short-time Fourier transform, each signal block is weighted with a window function, and consequently each short-time spectrum is convolved with the window's spectrum. The one-sided template magnitude spectrum T^M of an ideal harmonic tone complex with partials at integer multiples of the fundamental frequency f_0 can thus be described as

[1]Precision is "how many VAD-positive estimates are truly pitched?"
[2]Recall is "how many truly pitched values are VAD-positive?"

$$T^M(f, f_0) = \sum_{p=1}^{\infty} W(f - p \cdot f_0), \tag{8.1}$$

where p is the partial index, and $W(f)$ the spectrum of the time domain window with its maximum at $W(f = 0)$.

For speech processing, a typical window function is the Hann Window, whose Fourier transform is [68]

$$W_{\text{Hann}}(f) = \frac{1}{2}W_{\text{Rect}}(f) - \frac{1}{4}W_{\text{Rect}}(f + \frac{f_s}{N}) - \frac{1}{4}W_{\text{Rect}}(f - \frac{f_s}{N}), \tag{8.2}$$

with f_s being the sampling rate and

$$W_{\text{Rect}}(f) = \mathrm{e}^{-iN\pi\frac{f}{f_s}} \cdot \frac{\sin\left((N+1)\pi\frac{f}{f_s}\right)}{\sin(\pi\frac{f}{f_s})}. \tag{8.3}$$

This template spectrum T^M correlates perfectly with a pure harmonic tone complex at its fundamental frequency. However, it also correlates strongly with each harmonic of the tone complex. To mitigate these octave errors, the template should not only correlate positively with harmonics, but additionally correlate negatively with subharmonics. This was accomplished by subtracting the template mean, which inserts shallow negative valleys between the sharp positive comb peaks. Since the template spectrum is now zero-mean, this has the additional benefit of making the correlation sum invariant to the number of comb teeth (i.e. fundamental frequency). Furthermore, a low-pass filter at 1000 Hz is introduced, since physical speech contains mostly low frequencies. The complete template spectrum for realistic voiced speech is then

$$\overline{T^M}(f, f_0) = H^{LP}(f) \cdot \left(T^M(f, f_0) - \frac{2}{f_s} \sum_f T^M(f, f_0) \right), \tag{8.4}$$

where $H^{LP}(f)$ is the spectrum of a low pass filter with 24 dB per octave, similar to the low pass shape of a long-term average speech spectrum [17],

$$H^{LP}(f) = \begin{cases} 1 & \text{for } f \leq 1000 \text{ Hz} \\ 10^{-\log_2(\frac{f}{1000})24\frac{1}{20}} & \text{for } f > 1000 \text{ Hz}. \end{cases} \tag{8.5}$$

Figure 8.2 shows graphs of the template $\overline{T^M}(f, f_0)$ for two fundamental frequencies. Since each template's mean is zero, its power is constant across fundamental frequencies.

To compare the template with speech data, the absolute short-time magnitude spectrum $M(t, f)$ of the block at time t is correlated with a set of template spectra at various candidate f_0. This results in the magnitude domain feature

$$F^M(t, f_0) = \operatorname{corr}\Big(M(t, f), \overline{T^M}(f, f_0)\Big) \cdot \left(\frac{f_s}{2}\right)^{-\frac{2 \cdot f}{f_s}}, \tag{8.6}$$

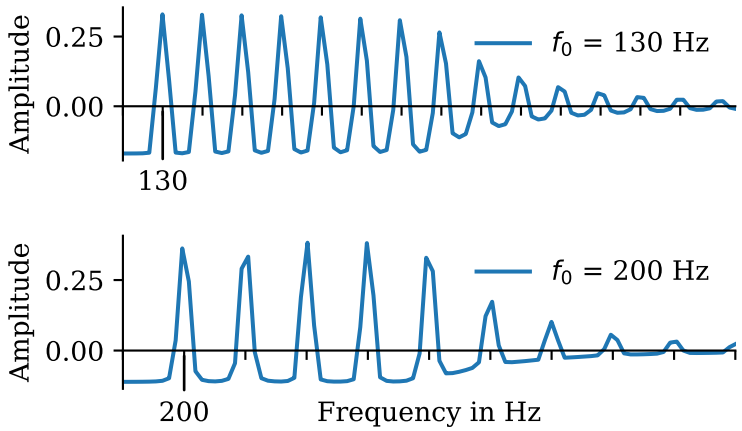


Figure 8.2: Templates for the magnitude spectrum for voiced speech at two typical fundamental frequencies of male and female speakers. Each template is a zero-mean comb with positive peaks at harmonics and negative troughs at subharmonics. High frequencies are attenuated, much like human speech.

where corr$(\cdot)$ denotes correlation and the last term weighs each frequency logarithmically, to emphasise the lower frequencies, where most voiced speech energy is concentrated by a local disturbance such as colored noise or a highpass filter.

Figure 8.3 shows $F^M(t, f_0)$ for the same clean speech signal as in Figure 8.1 for fundamental frequencies between 80 Hz and 450 Hz and 1024 point spectra for 30-ms signal blocks with 90 % overlap at a sampling rate of 48 kHz. The magnitude feature $F^M(t, f_0)$ is maximal at the fundamental frequency of the speech signal. Additional minor maxima are visible at each harmonic frequency. Thus, in this example, the fundamental frequency can be estimated without octave errors, but errors should seem likely if the fundamental were corrupted.

8.2.2 Voice in the Phase Spectrum

The phase spectrum itself does not show any obvious patterns indicative of human voice. However, there are variants of the phase spectrum that do show such patterns, such as its time derivative. The time derivative of the phase spectrum is commonly known as the instantaneous frequency (IF) [143]

This dissertation takes the somewhat unusual interpretation of phase spectra, where spectra are phase-shifted to the window center, as discussed in Equation 5.4 on page 35. This results in the IF representing a region dominated by a sinusoid as a linear frequency ramp that crosses zero at the frequency of the sinusoid, where a non-shifted IF would be equal to the sinusoid frequency modulo some phase wrapping constant. Our "unwrapped" IF is otherwise known as the instantaneous frequency deviation (IFD) [143, 83].

The phase spectrum in general is very sensitive to the STFT window function. Phase jumps in the window spectrum lead to destructive interference, which can hamper interpretation of the phase, as seen in Figure 5.11 on page 45. It is therefore advantageous to use a time window that does not contain any zeros in its spectrum, such as the Hann-Poisson window [139].

Figure 8.4 shows the IF and magnitude STFT of a speech signal. The comb-like magnitude pattern is clearly visible in the magnitude STFT. In the IF, the same signal produces a sawtooth-like pattern of repeating negative-to-positive ramps with one tooth per sinusoid. This pattern can be used to match voiced harmonic tone complexes in the IF, much like the comb pattern of the previous chapter

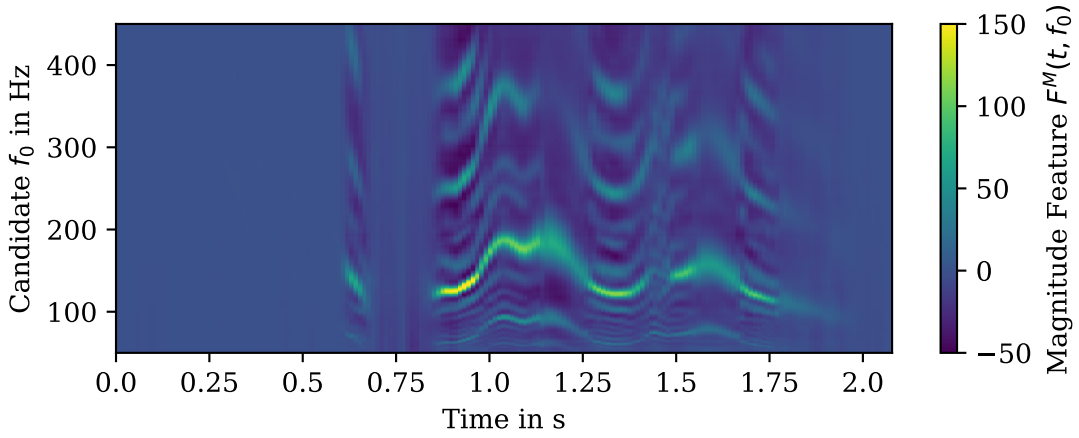


Figure 8.3: Magnitude domain feature for the same short, clean speech signal as in Figure 8.1, for candidate fundamental frequencies between 80 Hz and 450 Hz. The feature is maximal at the fundamental frequency between 100 Hz and 200 Hz, but octave ambiguities are clearly visible as parallel lines.

was used to match speech magnitude spectra.

Figure 8.5 shows the IF template $T^{\mathrm{IF}}(f, f_0)$ for harmonic tone complexes,

$$T^{\mathrm{IF}}(f, f_0) = \begin{cases} 0 & \text{for } f < \frac{f_0}{2} \\ f - \left\lfloor \frac{f}{f_0} \right\rceil \cdot f_0 & \text{otherwise,} \end{cases} \tag{8.7}$$

where the first term excludes the "zeroth" half-wave below the fundamental frequency, and the second term defines a linear positive-to-negative ramp around each partial frequency. The $\lfloor \cdot \rceil$ operator denotes rounding to the nearest integer.

To compare the template spectrum with speech data, a set of phase domain templates at different candidate f_0 is subtracted from the IF spectrum $\mathrm{IF}(t, f) = \mathrm{IF}(t, \frac{\omega}{2\pi})$ of the block at t. This results in the phase domain feature

$$F^{\mathrm{IF}}(t, f_0) = \left| \mathrm{IF}(t, f) - T^{\mathrm{IF}}(f, f_0) \right| \cdot \left(\frac{f_s}{2} \right)^{-\frac{2 \cdot f}{f_s}} \tag{8.8}$$

where the last term weighs each frequency logarithmically as in Equation 8.6. In contrast to the magnitude template, both the IF and the IF template are naturally zero-mean, and does not need to compensate for variations in difference between candidate fundamental frequencies.

Figure 8.6 shows $F^{\mathrm{IF}}(t, f_0)$ for a clean speech signal for fundamental frequencies between 80 Hz and 450 Hz and 1024 point spectra for 30-ms signal blocks with 90 % overlap at a sampling rate of 48 kHz. It can be seen that $F^{\mathrm{IF}}(t, f_0)$ is minimal at the fundamental frequency of the speech signal. Additional minima are visible in speech pauses, but no octave ambiguities are present.

8.2.3 Combination of Features

The previous two sections introduced two features for estimating the fundamental frequency of human voices. It was shown that both features have certain ambiguities: the magnitude feature is susceptible to octave errors at integer multiples of the fundamental frequency, while the phase domain feature

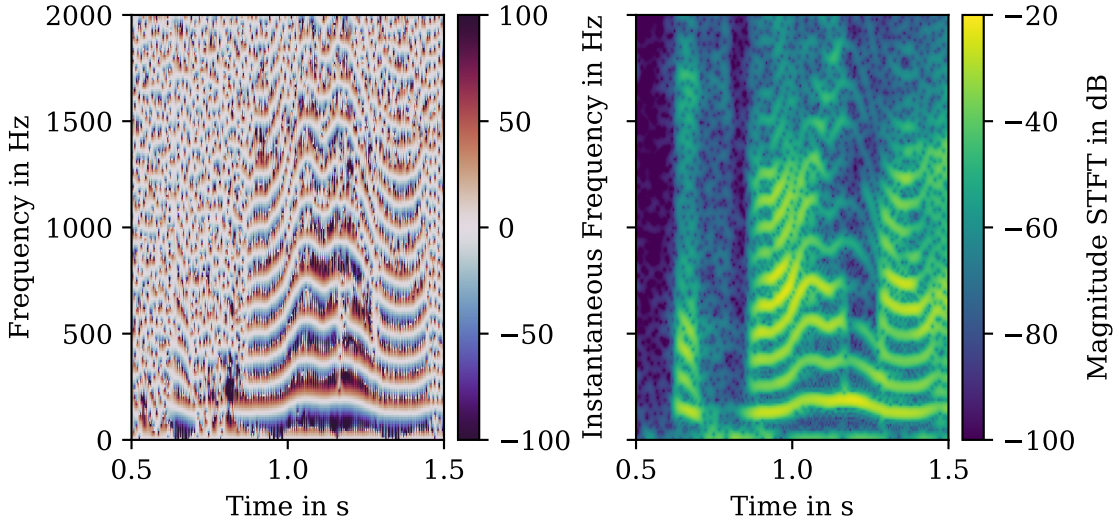


Figure 8.4: IF (left) and magnitude STFT (right) of a voiced speech signal. The IF represents sinusoids as linear negative-to-positive ramps with the zero crossing at the sinusoid. The magnitude STFT represents sinusoids as maxima. The graphed IF uses the Hann-Poisson window, the magnitude STFT the Hann window.

shows spurious minima during silent passages. Since these error modes can never happen at the same time, a combination of both features should avoid both errors.

According to Bayes' theorem, a combined *a posteriori* probability for voicing $P(\text{voice}|F^{\text{IF}}(t, f_0), F^M(t, f_0))$ can be derived from a prior probability of voicing $P(\text{voice})$, a marginal likelihood of magnitude and phase-feature values $P(F^{\text{IF}}(t, f_0), F^M(t, f_0))$, and a likelihood of feature values within voiced data $P(F^{\text{IF}}(t, f_0), F^M(t, f_0)|\text{voice})$:

$$F^C(t, f_0) = P(\text{voice}|F^{\text{IF}}(t, f_0), F^M(t, f_0)) \tag{8.9}$$

$$= \frac{P(\text{voice}) \cdot P(F^{\text{IF}}(t, f_0), F^M(t, f_0)|\text{voice})}{P(F^{\text{IF}}(t, f_0), F^M(t, f_0))}. \tag{8.10}$$

Each of these probabilities[3] was calculated as a histogram from a large database of speech recordings with known fundamental frequencies mixed with various noises. The histogram used 20 equally-spaced bins within the 1-99 percentile of both dimensions. $F^M(t, f_0)$ and $F^{\text{IF}}(t, f_0)$ were calculated for 200 f_0 for each test sample, and marked *voiced* if they were the closest f_0 to the true fundamental frequency according to the dataset's ground truth. The prior probability $P(f_0)$ is then the ratio between the number of voiced f_0 and all f_0. The marginal likelihood $P(F^{\text{IF}}(t, f_0), F^M(t, f_0)))$ is the 2D histograms of all combinations of $F^{\text{IF}}(t, f_0)$ and $F^M(t, f_0)$, and the likelihood $P(F^{\text{IF}}(t, f_0), F^M(t, f_0))|\text{voice})$ is the 2D histogram of all voiced combinations of $F^{\text{IF}}(t, f_0)$ and $F^M(t, f_0)$.

However, the resulting posterior probability of voicing $F^C(t, f_0)$ only covers combinations of $F^M(t, f_0)$ and $F^{\text{IF}}(t, f_0)$ that were found in the training data. To reduce the influence of random variations, the histogram was smoothed with a gaussian kernel with $\sigma = 3$. High-value histogram bins that were not found in the training data were filled by extrapolation towards high $F^M(t, f_0)$ and low

[3] A more correct term would be "frequencies", from the frequentist interpretation of Bayes' theorem, but we use the term "probabilities" to avoid ambiguities with signal frequencies.

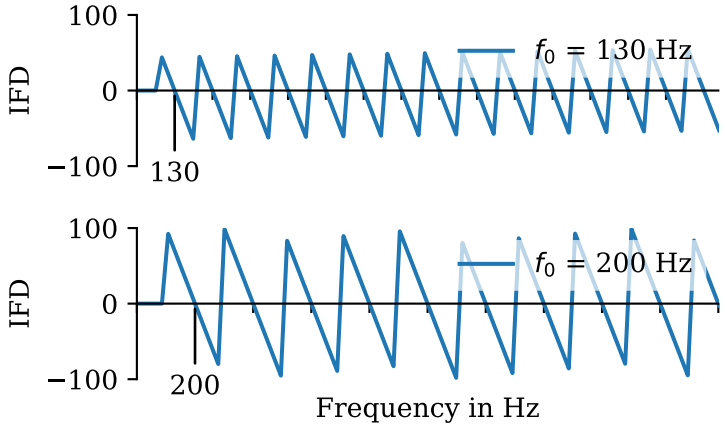


Figure 8.5: IF templates for voiced speech for two typical fundamental frequencies of male and female speakers. Each template is a sawtooth comb with zero crossings at harmonic frequencies.

$F^{\mathrm{IF}}(t, f_0)$. The pitch confidence can be seen in Figure 8.7, together with the decision threshold for the voicing decision. It will henceforth be called *pitch confidence*, since it measures the probability of estimating a correct fundamental frequency, instead of the more common maximum likelihood fundamental frequency. This allows the pitch confidence to be used as both PDA and VAD, instead of typical VADs based on features external to the PDA.

Figure 8.7 shows that bins are classified as voiced if $F^M(t, f_0)$ is high and $F^{\mathrm{IF}}(t, f_0)$ is low. If the voicing decision were based on either feature on its own, it would have to include an area of low $F^M(t, f_0)$ or high $F^{\mathrm{IF}}(F_0)$, and thus false positives. This confirms the earlier statement that each feature suffers from ambiguities that are absent when combined.

We used clean speech recordings from the pitch tracking database of the Graz University of Technology (PTDB-TUG) [120] and natural noise recordings from the Queensland University of Technology noise database (QUT-NOISE) [25] to calculate the pitch confidence. Both databases are freely available under open licenses. The speech recordings were split into a 60 % training set, which was used for the above calculation, and a 40 % test set, which was used for the evaluation below. Speech and noise were mixed with SNRs[4] ranging from -20 dB to 20 dB. In total, 50 repetitions with random speech recordings × 22 noise types at random starting times × 9 SNRs = 9,900 combinations of signal and noise were used to calculate the pitch confidence $F^C(t, f_0)$.

Figure 8.8 shows the pitch confidence for a clean speech signal. In comparison to Figure 8.3 and 8.6, there are no octave ambiguities as were in the magnitude domain and no spurious minima in quiet parts as were in the phase domain, and the fundamental frequency track stands out prominently.

Finally, the estimated fundamental frequency track $\hat{f}_{0_t}$ is calculated by Viterbi-searching $F^C(t, f_0)$ for a track that maximizes $F^C(t, f_0)$ and a transition probability between consecutive f_0:

$$\hat{f}_{0_t} = \underset{f_0, \hat{f}_{0_{t-1}}}{\operatorname{argmax}} \left(F^C(t, \hat{f}_{0_{t-1}}) \cdot \tau(\hat{f}_{0_{t-1}}, f_0) \cdot F^C(t, f_0) \right) \tag{8.11}$$

where $\tau(f_{0,t-1}, f_{0,t})$ is a transition probability from fundamental frequency $f_{0,t-1}$ to $f_{0,1}$, which penalizes large frequency jumps in order to create a smoother track with fewer disturbances. The transition probability was set to be proportional to the frequency quotient

[4]SNR calculated only from active speech segments, not pauses or leading/trailing silence.

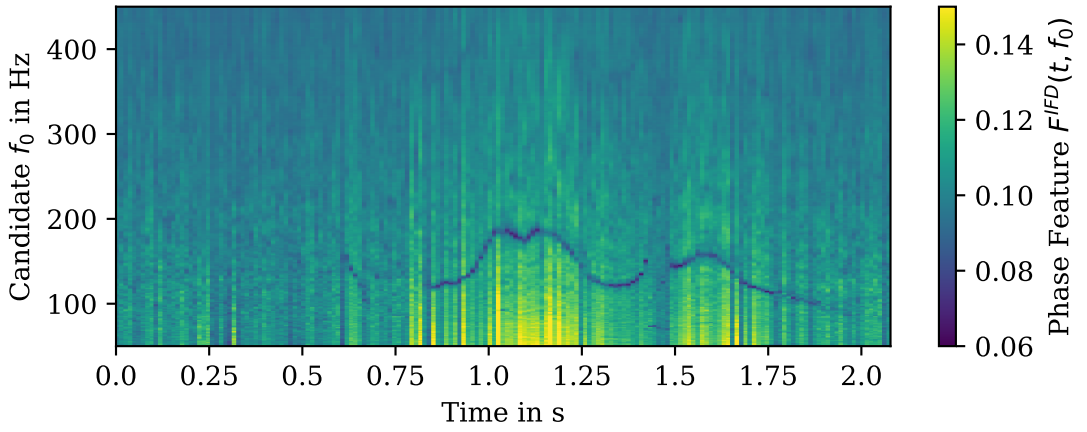


Figure 8.6: Phase domain feature for the same short, clean speech signal as in Figure 8.1 for candidate fundamental frequencies between 80 Hz and 450 Hz. The feature is minimal at the fundamental frequency between 100 Hz and 200 Hz but varies with signal level.

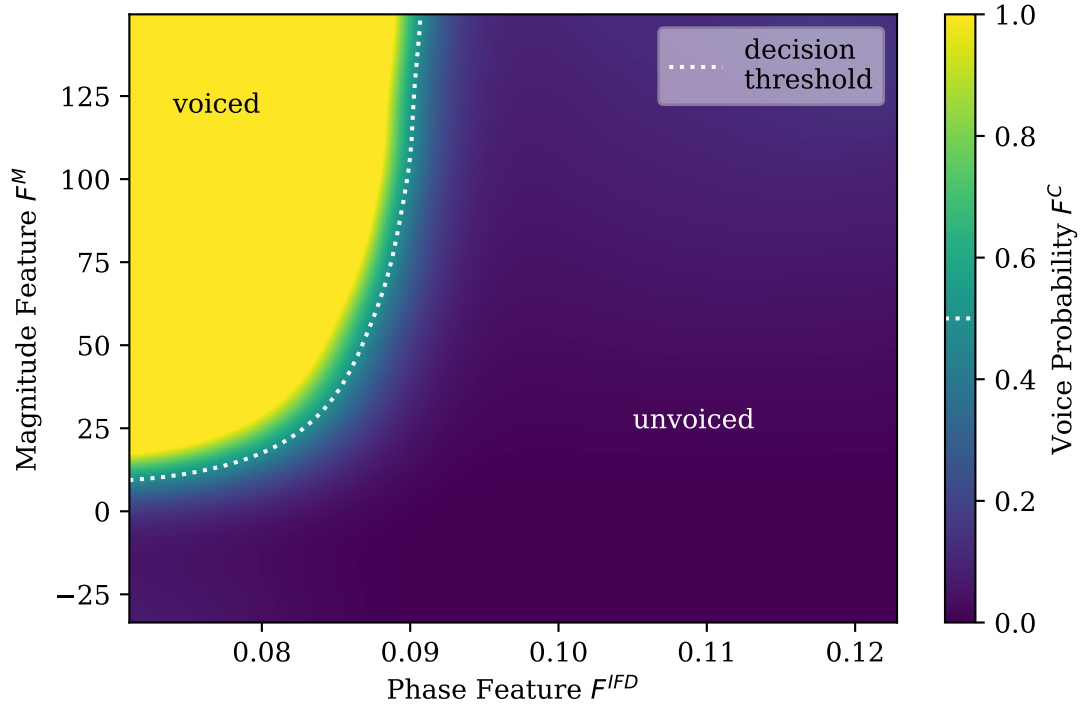


Figure 8.7: Pitch confidence $F^C(t, f_0)$ for various combinations of the magnitude domain feature $F^M(t, f_0)$ and the phase domain feature $F^{\text{IF}}(t, f_0)$. Only samples that have both high $F^M(t, f_0)$ and low $F^{\text{IF}}(t, f_0)$ exhibit a high pitch confidence. The graph area spans all feature values in the 1–99 percentile of the training data. Note that the magnitude domain feature is a correlation between spectral magnitudes and a real-valued template, which can be negative. The phase domain feature is an absolute difference, which is always positive.

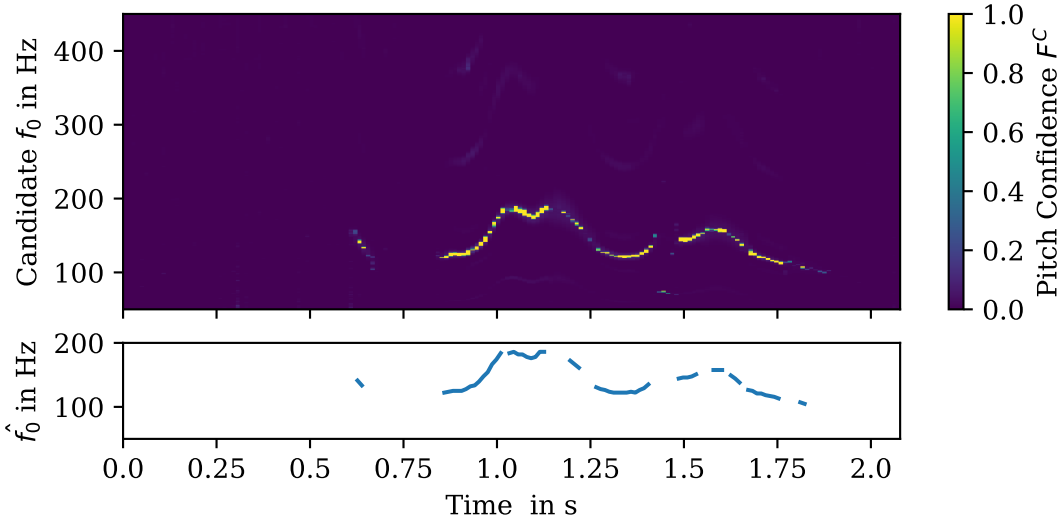


Figure 8.8: Pitch confidence for the same short, clean speech signal as in Figure 8.1 for candidate fundamental frequencies between 80 Hz and 450 Hz (top) and resulting estimated fundamental frequency track (bottom). Pitch confidence is maximal at the fundamental frequency between 100 Hz and 200 Hz. The artifacts from both Figure 8.3 and 8.6 are markedly reduced, and the fundamental frequency track is sharply defined. The fundamental frequency track excludes $F^C < 0.5$ and the spurious maxima near 1.4 s.

$$\tau(f_{0,t-1}, f_{0,1}) = \min(f_{0,1-1}, f_{0,1}) / \max(f_{0,t-1}, f_{0,t}). \tag{8.12}$$

The Viterbi search is typically implemented with a dynamic programming algorithm that searches for the globally optimal path through $F^C(t, f_0)$ from back-to-front and automatically finds the optimal start and end point.

The estimated fundamental frequency of frame at time t is then $\hat{f}_{0_t}$, and its pitch confidence is accordingly $F^C(t, \hat{f}_{0_t})$. The resulting pitch track can be seen in the bottom panel of Figure 8.8.

A meaningful voicing decision can be made at $F^C(t, \hat{f}_{0_t}) > 0.5$, where the probability of voice becomes greater than its inverse probability of non-voice. Note that this criterion does not estimate the probability of general voice activity within a frame, but the specific confidence that a pitch can be estimated accurately at the current frame.

8.2.4 Implementation and Parameters

In order to be able to compare the template spectra with real spectra in a computer program, a minimum frequency resolution is required. The comb-like pattern of the magnitude template needs at least two bins of separation between neighboring sinusoids. The sawtooth-like phase template can only be found with at least three bins per sinusoid. For typical human voices with fundamental frequencies as low as 80 Hz, and a sampling rate of 48 kHz, this requires an FFT length of at least 48000 Hz $/\frac{80 \text{ Hz}}{3} = 1800$ samples. Thus, the algorithm uses a default block length of 2048 samples, and a hop size of 1024 samples.

The phase domain feature is derived from the difference between two highly overlapping phase spectra. The distance Δt between these two blocks needs to be large enough to have good numerical

accuracy, but small enough so that there is no phase wrapping between neighboring sinusoids. The overlap for pitch frequencies up to 450 Hz should therefore be $\Delta t < \frac{1}{450\ \text{Hz}} \cdot 48000\ \text{Hz} = 107$ samples. In practice, this can be implemented as splitting each block of length 2048 into two sub-blocks of length $2048 - 107 = 1941$ samples.

In the final step of the algorithm, the frequency resolution of the fundamental frequency estimate $\hat{f}_{0_t}$ is dependent on how many candidate fundamental frequencies are evaluated for each of the templates. Since pitch is perceived with logarithmic accuracy with respect to its frequency [98], it makes sense to space fundamental frequency candidates logarithmically in the range of human voice pitches. A frequency resolution of 200 candidates between 50 Hz and 450 Hz was deemed sufficient, and extends a bit beyond the normal range of human voice pitches in order to include both children and pathological voices.

8.3 Evaluation

To evaluate the characteristics of this *Magnitude and Phase Spectrogram* (*MAPS*) pitch estimator, we used two datasets: First, the upper bound of the theoretical accuracy of *MAPS* was evaluated with 7800 synthetic, speech-like, harmonic tone complexes at various fundamental frequencies in various levels of white noise. Second, *MAPS* was evaluated with 5720 combinations of speech recordings from the PTDB-TUG [120] corpus, mixed with background noise recordings from the QUT-NOISE [25] corpus and white noise at various signal-to-noise ratios. The PTDB-TUG corpus contains microphone recordings and laryngograph-derived pitch tracks of 20 English speakers for 4720 TIMIT sentences. As mentioned above, these were separated into a 60 % training set and a 40 % test set. The experiments only used the test set. The QUT-NOISE corpus contains over 10 hours of various real-world background noise recordings such as traffic noise or cafeteria noise.

In order for our results to be comparable with other publications, we additionally included the three widely cited PDAs *RAPT* [150], *YIN* [24], and *PEFAC* [45] in this evaluation. *RAPT* is a variant of the time domain auto-correlation method at multiple sampling rates and dynamic programming to select pitch tracks. *YIN* is a time domain PDA that compares short signal blocks using the difference function and an emphasis for low lags. *PEFAC* works in the magnitude log-frequency domain with a soft comb and uses a Gaussian mixture model for its voicing decision. The main feature of *PEFAC* is somewhat similar to the magnitude feature in *MAPS*, albeit with a different soft comb and in a logarithmically warped spectrum. These PDAs were selected for being widely cited by comparison papers such as [3, 151, 45, 50, 148, 51]. See Chapter 11 for more information on these PDAs. The ground truth pitch is taken from the PTDB-TUG speech corpus, which used a modified *RAPT* PDA on laryngograph recordings.

Gross Pitch Error

To evaluate the algorithm's estimation accuracy similarly to how humans perceive pitch, the *Gross Pitch Error* (GPE) [127] was used. The GPE is the percentage of pitch estimates that deviate from the true pitch by more than ± 20 %. This is similar to prosodic speech perception for Western languages, where the overall shape of pitches is more important than the precise frequency values. Figure 8.9 shows gross pitch errors for synthetic speech in noise and speech recordings in recorded noise, for both the ground truth's voicing decisions and the algorithms' own voicing decisions.

The theoretical performance of each algorithm tends towards zero GPE for positive SNRs. At negative SNRs, errors rise quickly. The same pattern can be seen for realistic signals, although zero GPE remains elusive, and errors start rising earlier than in the synthetic case. Using the voicing decisions from *MAPS* instead of the ground truth improves accuracy significantly, since it limits the evaluation to high-confidence frames, even approaching the theoretical maximum from the synthetic

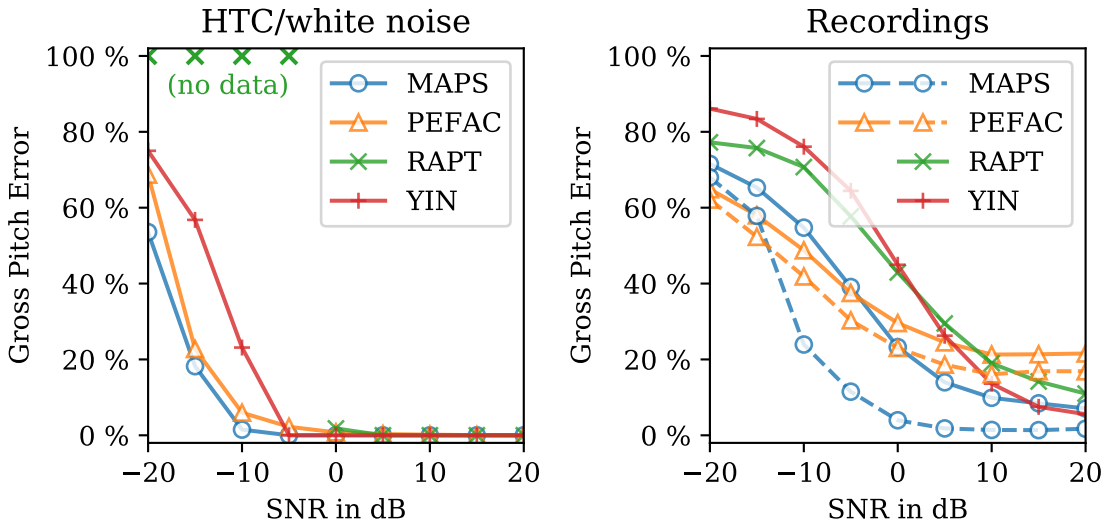


Figure 8.9: Pitch estimation accuracy for synthetic, speech-like, harmonic tone complexes in white noise (left) and speech recordings in recorded noise (right) over SNR. The Gross Pitch Error (GPE) is the percentage of estimates with a difference greater than ± 20 % between estimated and true pitch for voiced frames. Solid lines use the ground truth for voicing decisions, dashed lines use the PDA's own voicing decisions (if available).

evaluation. A similar improvement can be seen for *PEFAC*, albeit not as strongly. Section 8.3 will investigate this difference in more detail.

Fine Pitch Error

Figure 8.10 shows the algorithm's precision as the *Fine Pitch Error* (FPE) [127]. The FPE is the mean error of fundamental frequency estimates within the ± 20 % bounds given by the GPE. This is a numerical analysis of the absolute precision of the estimation, useful for singing voices and musical applications.

Again, synthetic results show near-zero FPEs for positive SNRs and quick deterioration at negative SNRs. Realistic results show a similar shape, but generally worse precision. Using the algorithms' own voicing decisions again improves precision significantly. FPEs however do not approach zero at high SNRs, but instead converge on about 1.5 % FPE. This is likely caused by small pitch errors in the ground truth itself, which can not be estimated by PDAs and therefore register as FPEs. Part IV will look into this topic in more detail, and will investigate alternative ground truths and speech corpora for evaluating PDAs.

For synthetic signals, mean errors of *MAPS* in Figure 8.10 did not quite reach zero, since the default fundamental frequency search space of *MAPS* is set to 200 points within 50 Hz–450 Hz, which yields a frequency resolution of $^{450}/_{50}{}^{1/200} = 1.01 \triangleq 0.5$ % FPE. However, the realistic evaluation shows that this precision is sufficient for real-world signals, in that it is higher than the maximum attainable precision in the data set. If higher precision is desirable, the density of the fundamental frequency search space can be adjusted accordingly. Again, the FPE shows that not every frame has an unambiguous fundamental frequency, and precision can be improved by selecting only high-confidence frames.

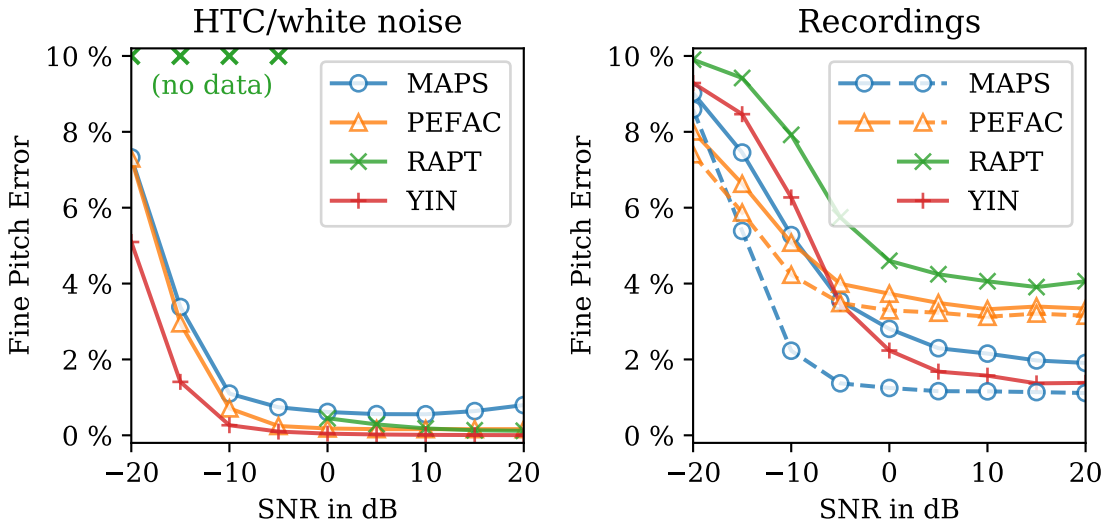


Figure 8.10: Fundamental frequency estimation precision for synthetic speech in white noise (left) and speech recordings in recorded noise (right) over SNR. The Fine Pitch Error (FPE) is the MAD of fundamental frequencies within ± 20 % of the true fundamental frequency, for voiced frames. Solid lines use the ground truth for voicing decisions, dashed lines use PDA's own voicing decisions (if available).

Voicing Determination

So far, *MAPS* has been shown to be highly accurate and precise because only high-confidence frames are selected for fundamental frequency estimation. Figure 8.11 shows a detection error trade-off graph, which characterizes the trade-off between false negatives and false positives in the algorithm's voicing decision. Since ambiguous estimates are discarded by the VAD algorithm, *MAPS*' false negative rates are high. However, its positive VAD decisions almost always result in accurate and precise estimates, as evidenced by the negligible false positive rate and very low GPEs. Furthermore, this remains true even at low SNRs, even though false negative rates clearly deteriorate.

The trade-off between false positives and false negatives can also be adjusted by the VAD threshold. A more conservative threshold selects fewer frames as voiced, and generally leads to higher precision and fewer gross pitch errors. Figure 8.12 illustrates this connection for *PEFAC* and *MAPS*. However, while high thresholds can generally push most VADs towards 100 % precision, they do not necessarily push GPEs towards zero.

MAPS is an unusual PDA in that it is a joint pitch estimator and voice activity detector. Its "VAD" explicitly selects frames with a clearly defined pitch, instead of merely voiced ones. Because of that, precision and GPE accuracy are tightly coupled for *MAPS*, whereas PDAs with mismatched VADs tend to retain some base level of residual GPEs due to ambiguously voiced frames.

It should also be noted that the ground truth's voicing decisions are based on laryngograph measurements, which sometimes remain periodic even when the microphone signal is not, for example during phoneme transitions or noisy fricatives [110, ch. 3.3]. This exacerbates false negatives to some extent, and implies that real-world false negative rates are likely not as egregious as in Figure 8.11.

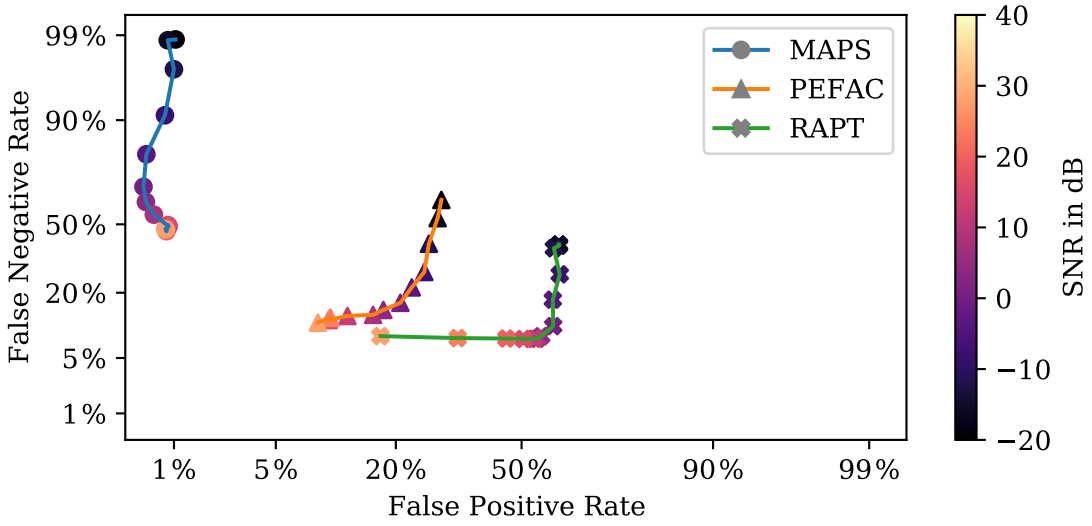


Figure 8.11: Voicing decision properties as a detection error trade-off graph. False negative rates are plotted against false positive rates on probit-warped axes for various SNRs and PDAs.

Signal Conditions

Figure 8.13 shows the influence of various noises and fundamental frequencies on the estimation accuracy of *MAPS*. The relatively small spread between GPE lines in the left graph shows that the type of noise does not have a major influence on estimation accuracy. Only very stationary noises such as car noise and white noise are significantly different from the other noises with more temporal variance and tonal components. If the algorithm's own voicing decision is used, the SNR of stationary noises affects accuracy less. This is not as apparent for non-stationary noises, where *MAPS* likely picks up tonal components in the background noise, which result in GPE errors. If the background noise is stationary and non-tonal, however, accurate and reliable estimation remains possible even with extreme SNRs, although the number of positive voicing decisions becomes small.

The right side shows that estimation accuracy is generally better for high-frequency voices at positive SNRs and low-frequency voices at negative SNRs. Since different noises mask different frequency ranges, low-frequency voices with denser harmonics are easier to detect at low SNRs. At positive SNRs, however, the wider harmonic spacing of higher-frequency voices facilitate their correct estimation.

8.4 Conclusions

The main contribution *MAPS* makes is the Bayesan combination of a feature in the magnitude spectrum and a feature in the phase spectrum. This reduces the ambiguities in either feature, and provides a probabilistic measure for pitch confidence. Since the confidence is used for both fundamental frequency estimation and the voicing decision, pitch estimates are highly reliable and accurate. Its reliability can be ascribed to the algorithms' preference to rejecting ambiguous frames over giving uncertain estimates, which results in almost perfect precision at the cost of some recall.

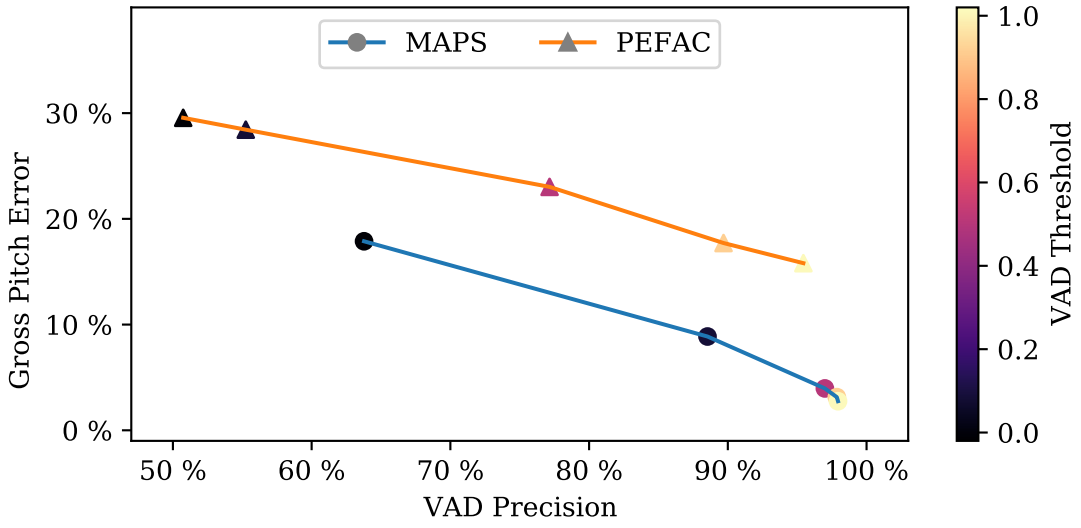


Figure 8.12: Effect of changing the VAD threshold on precision and GPE scores at zero dB SNR. Precision rises with threshold, while GPE scores fall. Near zero threshold, GPE scores reach ground truth levels from Figure 8.9. GPE scores improve with higher thresholds, but remain bounded by an algorithm-specific minimum.

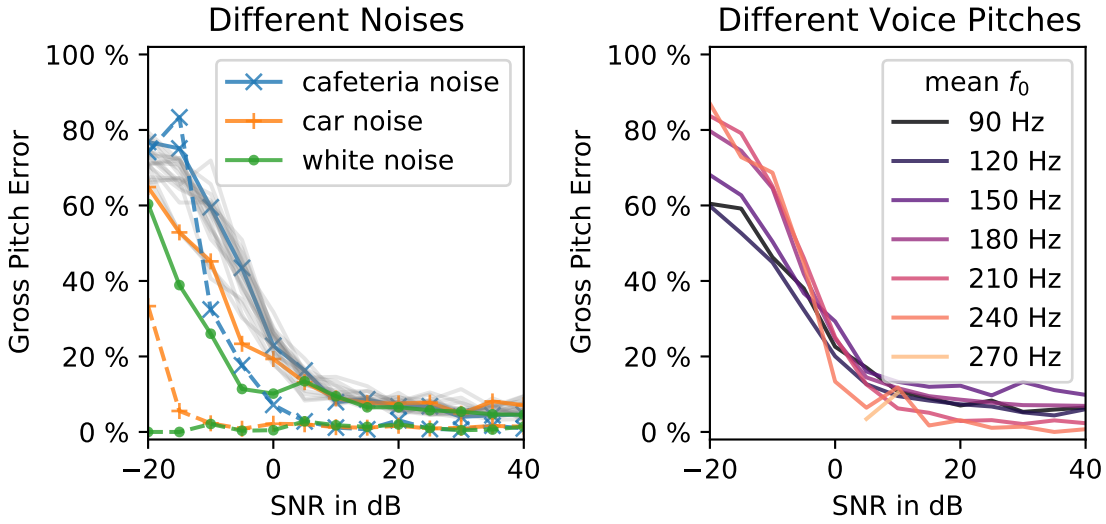


Figure 8.13: Estimation accuracy of *MAPS* in GPE for different noises (left) and base frequencies (right) over signal-to-noise ratio. The left side shows GPEs for all available noise types with ground truth voicing decisions as solid lines, with three prototypical noises highlighted in color and dashed lines for the PDA's own voicing decision. The right side shows GPEs for ground truth voicing decisions for the test set of speech/noise recordings against their mean fundamental frequencies.

Part IV

Defining Truth in Fundamental Frequency Estimation

Where we realize that any measure for accuracy of fundamental frequency estimation algorithms is only as valid as the ground truth it is judged against. Thus, a better measure for accuracy requires a well-chosen corpus of speech recordings, and a defect-free ground truth. While this information would have no doubt been of great use in the construction of the fundamental frequency estimation algorithm in Part III, it was not yet available at the time, and it would be disingenuous to retrofit it into the narrative.

In this part, Chapter 9 explores existing corpora and their suitability for evaluating fundamental frequency estimation algorithms. It finds significant problems and differences between these corpora, both in the quality and quantity of their recordings, and in their fundamental frequency ground truths, if any are included.

Chapter 10 addresses the need for a defect-free ground truth for evaluating estimators, by introducing a *Consensus Truth*. In contrast to most existing ground truths, the consensus truth neither relies on categorically mismatched laryngograph measurements, nor on error-prone estimates from a single reference algorithm. It is thus better suited for evaluating the accuracy of fundamental frequency estimation algorithms, especially in edge cases.

Chapter 9

Speech Databases for Pitch Determination

Abstract

The development of pitch determination algorithms (PDA) is reliant on speech and noise corpora for training and evaluation. These corpora are designed to be diverse enough to capture all the major phonemes and pronunciations of everyday speech, yet compact enough to remain computationally feasible. This chapter explores a number of such speech and noise corpora, quantifies their differences and biases, and investigates their popularity in published literature on PDAs. These analyses highlight a number of issues in these corpora, particularly with regards to speaker diversity and cross-corpus comparability. In general, the *PTDB-TUG* speech corpus and the *QUT-NOISE* noise corpus are found to be appropriate default corpora for the development of fundamental frequency estimation algorithms.

9.1 Introduction

Human speech is a fundamentally human signal. It is created by humans and most typically perceived by humans. As such, the pitch of human speech can only be accurately assessed by human listeners. However, human pitch perceptions are unavailable to human-machine interactions such as automatic speech recognition or voice activity detection systems. Machines must therefore rely on some other form of ground truth for estimating the pitch of human speech. These non-human pitch estimates are referred to as *fundamental frequency* instead of *pitch*, and a plethora of databases and algorithms have been published for estimating and evaluating it.

However, it must be stressed that the notion of *accuracy* in fundamental frequency estimation algorithms (or pitch determination algorithms) for human speech can only ever be claimed with respect to such a ground truth, which is necessarily derived from another algorithmic estimation, not from actual human pitch perceptions[1]. The choice of ground truth is therefore a decision of great consequence in the design of a fundamental frequency estimation algorithm: Varying algorithms for estimating the fundamental frequency may easily result in an implicit bias when used as a ground truth.

An alternative method for evaluating PDAs is the use of synthetic speech with a known fundamental frequency. However, this merely shifts the artificiality from the estimation algorithm to the generation algorithm, and does not improve the validity of PDA accuracy for human speech.

[1] excluding manually-annotated pitch data, which are impractical for large datasets.

Thus, databases of pre-recorded speech remain a necessity. Several speech databases have been published for the purposes of fundamental frequency estimation, which provide large collections of recordings of human speech and various metadata, such as a fundamental frequency ground truth or laryngograph recordings. The latter are measurements of the openings of the vocal folds during phonation, which can be used as an alternative source signal for fundamental frequency ground truth.

However, laryngograph recordings are not without their own problems. Most importantly, vocal folds can intermittently vibrate even though the mouth is closed, and thus have a clearly defined fundamental frequency where there is no speech. Similarly, vocal folds may vibrate clearly in transitions where the rest of the vocal tract is not yet in resonance and the speech signal has no clear pitch yet. The technical apparatus of laryngographs furthermore cannot distinguish between partially closed and closed vocal folds, and thus misrepresent certain types of phonations [5, 110]. In general, laryngograph recordings can (locally) exhibit tonality independent of the acoustic speech signal or the acoustic energy, and are therefore somewhat problematic as a source of ground truth [110].

On a more philosophical level, laryngographs imply that the truth about pitch is to be found at the source of production, the vibrations of the vocal cords. Alternatively however, it could be argued that pitch is instead a perception, and should be measured not at the source, but from the perceived acoustic waveform instead. Both viewpoints have merit, but PDAs necessarily estimate the latter, as they do not have access to a measurement of vocal fold vibrations. A production-based ground truth would therefore expect slightly more errors in PDA estimations than a perception-based one.

Regardless of these philosophical musings, any ground truth fundamental frequency has to be derived from a fundamental frequency estimation algorithm, either from the speech recordings or from laryngograph recordings. This circular reasoning can be valid, as ground truth estimates are derived from clean recordings, while PDA evaluations are instead conducted with mixtures containing noise recordings. However, this assumes close to perfect recording conditions of the speech database, which may not be true for older databases. Furthermore, both the choice of speakers and sentence lists may or may not be a good fit for the reference PDA chosen as ground truth, which can bias the speech database for some applications.

All of these factors form a type of database-dependent bias that makes comparisons between PDAs difficult unless the same databases are used for both PDAs. This issue is exacerbated for data-driven algorithms that are trained for a particular dataset, either explicitly in a machine learning algorithm, or implicitly by optimizing algorithm parameters manually during the development process.

This chapter therefore seeks to quantify differences between speech databases typically used for fundamental frequency estimation, and to ascertain their suitability for comparison studies. These differences should not be interpreted as shortcomings of the databases themselves, but merely a record of their compatibility with other databases.

9.1.1 Databases

Commonly used speech databases contain clean recordings of English speech, usually accompanied by metadata such as information on the speakers and utterances, as well as ancillary measurements of laryngographs or articulographs. Purpose-built databases for fundamental frequency estimation typically also contain a fundamental frequency ground truth.

The following databases were selected for being widely cited in publications on fundamental frequency estimation. All but *TIMIT* are freely available on the internet, with *TIMIT* being a commercial database purchasable from the Linguistic Data Consortium [40]. Table 9.1 summarizes the following corpora and their popularity in the literature.

Table 9.1: Speech corpora for fundamental frequency estimation by number of mentions in publications on fundamental frequency estimation between 1990-2020. The columns f_0 and Lar. denote whether the corpus has a fundamental frequency ground truth and laryngograph recordings.

Corpus	# Mentions	f_0	Lar.
KEELE	106	✓	✓
TIMIT	80		
FDA	63	✓	✓
PTDB-TUG	22	✓	✓
CMU-ARCTIC	19		✓
MOCHA-TIMIT	5		✓

CMU-ARCTIC [80]

http://www.festvox.org/cmu_arctic/

The *CMU-ARCTIC* speech synthesis databases were built by the Language Technologies Institute at Carnegie Mellon University for the purpose of speech synthesis research in 2003. It consists of 1132 phonetically balanced sentences from out-of-copyright texts from Project Gutenberg, read by 18 speakers, for a total of 13:53 h in 15603 audio recordings.

Aside from the audio recordings, it contains phonetic labels and laryngograph signals. The dataset is freely available under the terms of a BSD-style free software license.

It was selected for being cited in comparison studies such as [125, 23, 30, 32].

Paul Bagshaw's database (aka *FDA*/*CSTR* database) [5]

http://www.cstr.ed.ac.uk/research/projects/fda/

Paul Bagshaw's database for evaluating pitch determination algorithms was created as part of Bagshaw's Ph.D. thesis in 1993, and contains 100 samples read by two speakers in audio recordings six minutes in length. The sentences were chosen by Bagshaw explicitly for containing phonemes whose pitch is difficult to estimate. Additionally, it comes with a fundamental frequency ground truth derived from laryngograph recordings, which are included as well.

This database is otherwise called the "*FDA* database" or "*CSTR* database", and was one of the first publicly available databases for pitch estimation. For that reason, it has been widely used for evaluating PDAs, particularly in the nineties and early 2000s. Recent citations in comparison studies include [162, 49, 161, 1, 163].

The database was published as part of Bagshaw's Ph.D. thesis without any explicit licensing text and has been included for being widely cited.

KEELE [122]

https://lost-contact.mit.edu/afs/nada.kth.se/dept/tmh/corpora/KeelePitchDB/

The Pitch Extraction Reference Database created at Keele University in 1995 is another early database specifically built for pitch estimation. It contains six minutes of ten speakers, each reading Aesop's Fable "The Northwind and The Sun", a phonetically balanced text, complete with laryngograph recordings and a fundamental frequency ground truth.

Like *FDA*, this database has been cited widely due to its early publication date, but has remained popular to date, e.g. in [152, 162, 115, 161, 30, 46, 163, 91].

While the original public FTP links to the database are no longer active, the above URL provides a mirror that was still available in 2019. No specific licensing text is given beyond "The database

is intended to be open", "the database should be open, easily obtainable, and practical", and "The database [...] is the first step towards a public database to aid evaluation of pitch extraction algorithms. The database is open and external contributions and remarks are welcome." [122].

MOCHA-TIMIT [168]

http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

The *MOCHA-TIMIT* database was created at the Queen Margaret University College in 1999, for the purpose of training automatic speech recognition systems, using the same sentence list as *TIMIT*. It is however well-suited for fundamental frequency estimation applications since it includes both laryngograph and articulograph recordings in addition to its 4028 samples of 4:38 h of speech recordings. In *MOCHA-TIMIT*, different subsets of 460 sentences are read by ten speakers, including three speakers in the main corpus and seven speakers in the "unchecked" set.

Some of the recording conditions in the "unchecked" set are not as clean as in the main set, but still usable for fundamental frequency estimation.

The data is "free for non-commercial use", according to the database's website [168].

PTDB-TUG [119]

https://www.spsc.tugraz.at/databases-and-tools/ptdb-tug-pitch-tracking-database-from-graz-university-of-technology.html

The *PTDB-TUG* pitch tracking database from Graz University of Technology from 2011 is one of the biggest and most complete speech databases for fundamental frequency estimation. It contains 9:36 h of 4718 speech and laryngograph recordings from 20 speakers, reading different sets 236 of *TIMIT* sentences each, as well as a fundamental frequency ground truth.

It is particularly relevant today as one of the largest databases built specifically for fundamental frequency estimation, and is particularly popular for comparison studies such as [142, 2, 49, 84, 128, 147, 174, 28, 66, 145].

The database is freely available under the terms of the Open Database License [37].

TIMIT [40]

https://catalog.ldc.upenn.edu/LDC93S1

The *TIMIT* Acoustic-Phonetic Continuous Speech Corpus is a commercial database sold by the Linguistic Data Consortium for the development and evaluation of automatic speech recognition systems. The database contains 6300 samples of 630 speakers reading 10 phonetically rich sentences each, for 5:23 h of audio recordings, as well as time-aligned orthographic, phonetic and word transcriptions.

This is another early database from 1993, originally distributed on CD-ROMs, that has been cited widely, both in comparison studies [89, 88] as well as numerous other publications.

9.2 Literary Survey

These databases were selected either for being particularly popular in publications on fundamental frequency estimation, or for being particularly well-suited for this task. To assess the relative popularity, we conducted a literary search of the last thirty years (1990-2020).

Publications were searched for in the IEEE database, Springer Link, and the INTERSPEECH conference archives and included if their title contained the key words "frequency" or "pitch", and "speech" was mentioned anywhere in their metadata. These results were filtered manually to only include publications relevant to fundamental frequency estimation of speech, based on their title. In

total, the resulting dataset contains 851 publications across hundreds of journals. Table 9.2 lists the most prolific journals for fundamental frequency estimation research, according to this dataset.

Table 9.2: Most prolific journals for publications on fundamental frequency estimation of speech. All journals with at least five publications were included in this table.

Journal	# Papers	Years
International Conference on Acoustics, Speech, and Signal Processing	159	1990-2019
INTERSPEECH Conference	139	1996-2019
Transactions on Audio, Speech, and Language Processing	54	1994-2019
European Signal Processing Conference	49	1996-2018
Workshop on Applications of Signal Processing to Audio and Acoustics	10	1999-2017
Transactions on Signal Processing	9	1991-2015
International Journal of Speech Technology	8	1999-2019
Electronics Letters	7	1993-2007
International Conference on Signal Processing	7	2002-2012
Signal Processing Letters	7	2008-2018
International Symposium on Circuits and Systems	5	2005-2009
EURASIP Journal on Audio, Speech, and Music Processing	5	2014-2018
other	391	1990-2019

To investigate use and popularity of the speech corpora over time, Figure 9.1 graphs the volume of their references over time. Due to the nature of this literature survey, this merely tracks *mentions* of the corpus names, which includes false positives where corpus name are used out of context. This is likely to overestimate *KEELE* and *CSTR/FDA* in particular, as any mention of the Keele University and the Centre for Speech Technology Research (CSTR) of the University of Edinburgh is included as a mention of the corpus. Also, the results of *TIMIT* necessarily include mentions of *MOCHA-TIMIT*.

Furthermore, this analysis relies on a full-text search of the articles in the dataset, which can be error-prone or impossible for older articles, where the PDFs sometimes include no text, or only garbled text from low-quality scans. This is likely to under-estimate popularity in older articles prior to 2000.

Nevertheless, the graph shows the use of corpora steeply increasing with the gaining popularity of the Internet in approx. 2005. This occurred in parallel with greater processing power becoming available and the analysis of larger datasets becoming possible. In the 2010s, this led to increased use of the larger internet-only corpora such as *PTDB-TUG* and *CMU-ARCTIC*, and a relative reduction in the smaller *FDA* and *KEELE* datasets. An exception to these trends is the *TIMIT* dataset, which is both large in size, and enduringly popular, likely due to its very early CD-ROM distribution that apparently remains available at many campuses.

This literature survey dataset is of limited utility in this chapter, but will be revisited in Chapter 9.2 with a more detailed analysis of fundamental frequency estimation algorithms.

9.3 Data Diversity

A key feature of these speech corpora is that they are designed to include a representative set of speech recordings that translate well to real-world speech usage. On the one hand, this means including as wide a variety of speakers and phonemes as possible in order to capture the complexities of natural speech completely. On the other hand the distribution of features should not be too wide so as to remain as close to natural speech as possible, and to not introduce any unnatural biases.

Earlier corpora in particular were also limited by the processing power available at the time, which

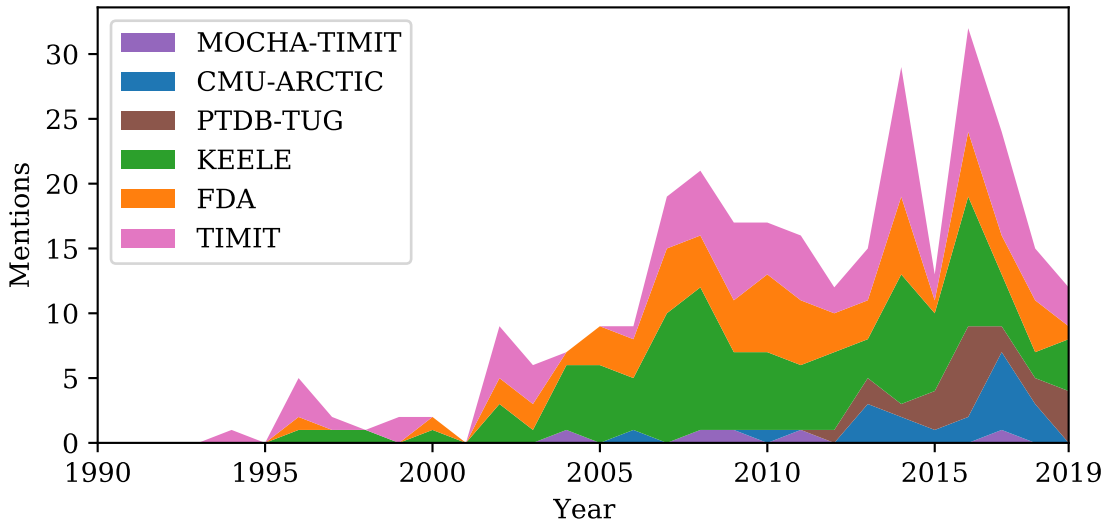


Figure 9.1: Mentions of various speech corpora over time as a stacked area chart. Search queries included common alternative names and spellings. Total height is sum of all mentions per year.

made handling larger data sets impractical. In fact, the very first fundamental frequency estimation algorithms from the 1960s took hours of computation per second of audio recording [105], which made any kind of larger analysis impossible. Larger datasets of multiple minutes of audio material became practical only in the 1990s, and led to the publication of the *FDA* and *KEELE* corpora, each with a little more than five minutes of speech recordings. Distribution of larger datasets remained limited to physically shipping CD-ROMs, which only commercial entities such as *TIMIT*'s Linguistic Data Consortium could feasibly undertake. Thus, truly large datasets only appeared when the Internet was widespread and fast enough to transport meaningful amounts of data in the 2000s, with multi-hour datasets such as *PTDB-TUG*, *CMU-ARCTIC*, and *MOCHA-TIMIT*.

These larger corpora include a greater variety of speakers and more diverse sentence lists. A particularly popular sentence list is the *TIMIT prompts*. *TIMIT*, *MOCHA-TIMIT*, and *PTDB-TUG* use variants of these sentence prompts, which are, according to the *TIMIT* documentation:

> The text material in the TIMIT prompts [...] consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. The dialect sentences (the SA sentences) were meant to expose the dialectal variants of the speakers [...]. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest [SX sentences]. [...] The phonetically-diverse sentences (the SI sentences) were selected from existing text sources - the Brown Corpus (Kuchera and Francis, 1967) and the Playwrights Dialog (Hultzen, et al., 1964) - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts.
>
> —from the *TIMIT* corpus documentation (*readme.txt*) [40]

The *TIMIT* prompts include well-known sentences such as "She had your dark suit in greasy wash water all year." (SA1), "Don't ask me to carry an oily rag like that." (SA2), "Seamstresses attach

zippers with a thimble, needle, and thread." (sx420), or "Everything went real smooth, the sheriff said." (SI453).

While these sentences cover a broad range of phonemes, they have been criticized for using an old-fashioned style and unrealistically complicated sentences for natural speech [80].

The *KEELE* corpus uses recordings of Aesop's Fable "The Northwind and The Sun", which is a phonetically balanced text recommended by the International Phonetic Association for showing phonemic contrast of various English accents [64]. However, the text has been found to lack a few common phonemes, but overall still a good choice for a phonetically balanced text [29].

CMU-ARCTIC is intended for training speech synthesizers and uses an original set of phonetically balanced sentences from out-of-copyright books from Project Gutenberg [85]. Sentences were selected for using a modern style, as well as being phonemically diverse.

Paul Bagshaw's *FDA* corpus was designed for evaluating fundamental frequency estimation algorithms. It is explicitly biased towards aperiodic phonemes such as fricatives, nasals, liquids and glides, which are described as particularly difficult to estimate.

In summary, most of these corpora use sentence lists optimized for phonemic diversity. While the concrete aims of the sentence lists differ, and have been criticized in some contexts, these issues are unlikely to be a problem for fundamental frequency estimation, so long as they cover most phonations and pitches. Still, there remains a trade-off between speaker variety and phoneme variety, that each corpus solves differently, and an obvious benefit to larger corpora.

9.4 Voice Activity

Speech corpora fill a precarious dual role, both as a means for evaluating PDA performance, and an optimization target while developing PDAs. Any bias in the corpora is likely to be incorporated into the algorithms as well. While these issues are certainly more prevalent in today's machine learning-oriented workflows than in yesteryear's theory-motivated signal processing designs, there always remain implicit assumptions about the kinds of signals an algorithm is expected to be exposed to.

One particularly under-reported aspect of such a corpus bias is its ratio between speech and silence. All examined corpora include a few seconds of silence before and after each speech recording, probably in part as an artifact of the recording setup, and in part to give PDAs some time for signal adaptation before the beginning of the speech data. Yet, these silent passages might ultimately bias the PDAs' voice activity determination (VAD) system for or against speech activity, or implicitly condone random pitch estimates at low signal levels if not otherwise controlled for.

Figure 9.2 illustrates the total length of recordings in each corpus, and the amount of silence in each corpus. As previously noted, total lengths differ dramatically between corpora. But perhaps surprisingly, a large percentage of each corpus' recordings consist of silence. This graph considers as silence any 25-ms block that has less energy than the average of the 5th and 95th percentile block energy in dB, which was empirically found to be a robust estimator for discriminating speech from silence. The strongly divergent speech/silence ratios will undoubtedly bias PDAs and VADs for different signal sparsities.

Equally important as the total lengths and ratios between speech and silence is where that silence occurs. Figure 9.3 shows histograms of both speech and silence durations, split into silence before the speech, silence during the speech, and silence after the speech. Any 25-ms block was considered speech or silent by the same criterion as above, but excluding single-block outliers in the silence before and after speech, which could occur due to recording glitches such as clicks.

The top graph highlights how the *KEELE* corpus uses fewer, but much longer, recordings than the other corpora in this set. Apart from that, the speech segments in all corpora are of similar length, possibly slightly longer in *PTDB-TUG*. A more significant difference appears in the bottom graph

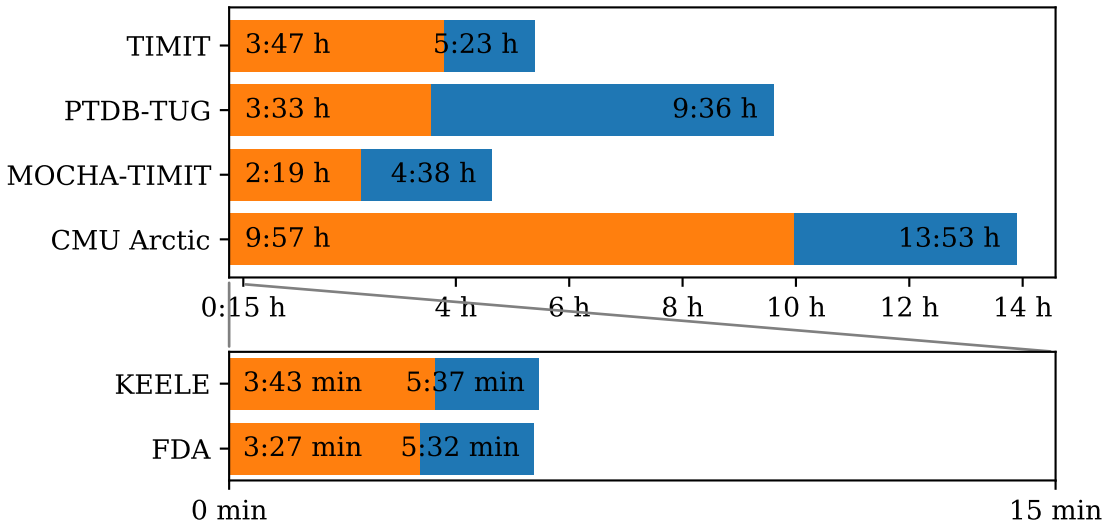


Figure 9.2: Total length (blue) and speech length (orange) of various speech corpora. Note the time scale difference between the top and bottom bars.

in the silences, where *CMU-ARCTIC*, *FDA*, and *TIMIT* usually have less than a second of silence before the speech recording, whereas *MOCHA-TIMIT*, and particularly *PTDB-TUG* generously pad each recording with a few seconds of silence.

These last differences could be of great importance for PDAs that include a denoising or noise normalization phase, which frequently assume a few frames of undisturbed background noise at the beginning of each recording. The performance of such PDAs should be expected to be stronger on *PTDB-TUG* and *MOCHA-TIMIT* than on less padded corpora.

The amount of silence during each speech recording is comparable between corpora, with the obvious exception of *KEELE*, and silence after the recording follows a similar distribution as the pre-silence above.

Since the long recordings in the *KEELE* corpus is such an outlier in these preliminary discussions, we developed a modified version of the *KEELE* corpus, which simply cuts each *KEELE* recording in about a dozen shorter pieces, and thus brings it more in line with the other corpora. This corpus is included in the graphs as *KEELE*-mod.

9.5 Long Term Average Speech Spectrum

Due to the different speakers and sentence lists, the spectral makeup of the corpora is bound to differ. Figure 9.4 shows long term average speech spectra of the corpora. Strikingly, there are significant differences in overall loudness between the corpora, with *TIMIT* being most quiet and *CMU-ARCTIC* being loudest. This might be problematic for cross-corpus comparisons of PDAs with absolute decision thresholds.

The high-frequency ends of the spectra illustrate the different sampling frequencies of the corpora, where *CMU-ARCTIC*, *MOCHA-TIMIT*, and *TIMIT* are using 16 kHz, *KEELE* and *FDA* at 20 kHz, and *PTDB-TUG* with 48 kHz. Depending on the intended application, these differences might be very significant, not, perhaps, due to the additional spectral content, which reflects rather low-energy above 8 kHz, but mostly as some PDAs are optimized for a particular sampling rate and frequently do not work optimally for different ones.

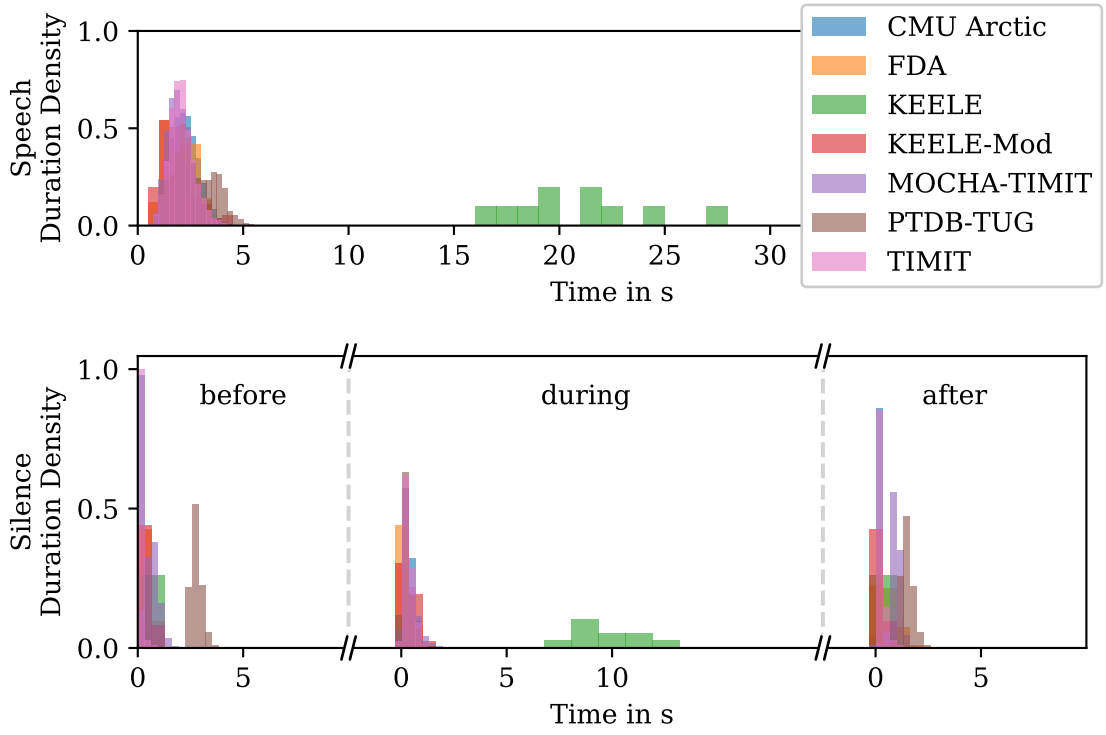


Figure 9.3: Histograms of length of speech (top) and silence before, during, and after the speech in each recording. Time scaling on top and bottom are identical.

The main point of the long-term spectra is their spectral composition. Most corpora show three local spectral maxima around 150 Hz, 300 Hz and 600 Hz [17]. In the literature, these three maxima usually rise in magnitude with frequency, which has been replicated similarly in *KEELE*, *PTDB-TUG*, *FDA*, and *TIMIT*. However, *CMU-ARCTIC* and *MOCHA-TIMIT* depart somewhat from published long-term spectra and emphasize the lower maximum more strongly, while *TIMIT* has a particularly strong high maximum.

These biases towards lower frequencies in *CMU-ARCTIC* and *MOCHA-TIMIT*, or higher frequencies in *TIMIT*, can be significant for cross-corpus comparisons: In the presence of background noises, noises will predominantly mask the lower-energy spectral regions, which correspond to different frequency ranges in these corpora. Moreover, PDAs use features derived from different spectral regions and should therefore be expected to perform differently for the aberrant corpora.

Although no mention of this is present in the corpora's documentation, it seems reasonable to assume that the unusual rising slope of *TIMIT* between 100 Hz and 500 Hz is an indication of a high pass filter used during recording.

Reassuringly, at least, all corpora fall off similarly towards low and high frequencies, with a steady slope of appr. 12 dB/octave.

9.6 Level Distribution

Another consideration for the comparability of corpora is their level range differences, and whether all speakers were recorded at the same level. To investigate this, Figure 9.5 shows histograms of the corpora's signal levels. As expected from the previous section, corpora with more silence have stronger peaks on the quiet side of the histograms, whereas corpora with more speech are skewed more towards

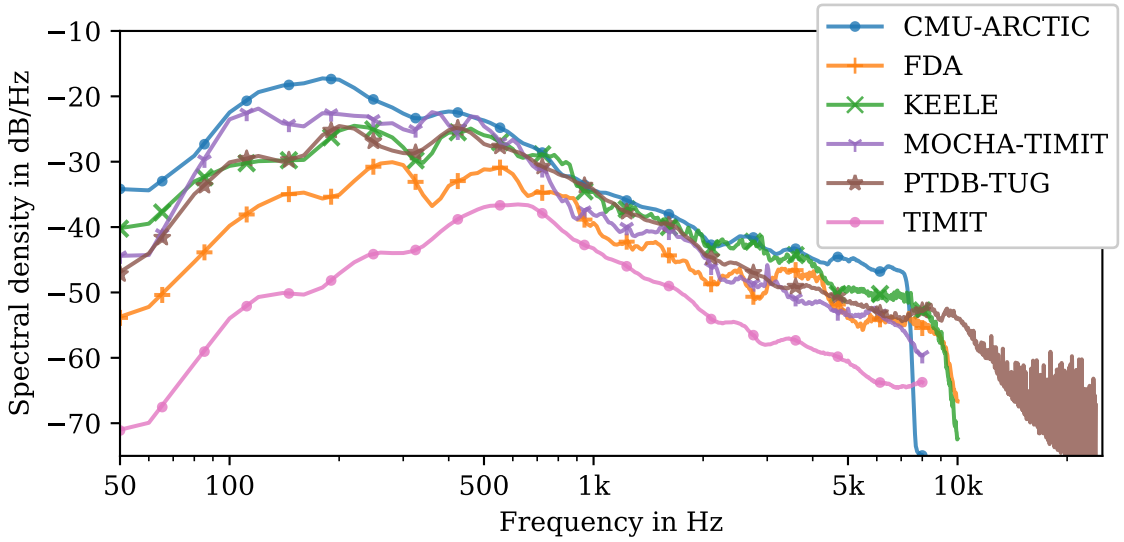


Figure 9.4: Long term average speech spectrum of the corpora. Spectra are 50-ms Welch spectra using a Hann window, averaged over all speech blocks of all recordings of each corpus.

the loud side.

However, the relative levels of the loud end and the quiet end vary strongly between corpora. As we already learned from the previous section, *TIMIT* and *FDA* have quieter maxima than the other corpora, and *CMU-ARCTIC* is clearly the loudest. Moreover, however, the level difference between silence and speech of *MOCHA-TIMIT*, *FDA*, and *TIMIT* is about 10 dB better than *CMU-ARCTIC*'s, *KEELE*'s, or *PTDB-TUG*'s. Though unlikely to be of importance for fundamental frequency estimation, these differences can potentially fool simple voice activity determination algorithms, often part of older PDAs.

A particularly interesting case of this is *MOCHA-TIMIT*, which evidently used two different recording setups with different levels of background noise.

The main histogram peak in most corpora looks roughly Gaussian, except for *CMU-ARCTIC*, whose main peak is asymmetric with a sharper fall-off towards the maximum level of 0 dB than towards lower levels. This might indicate the presence of a compressor in the recording setup, even though this is not mentioned in [80].

9.7 Voiced vs. Unvoiced speech

For the purposes of fundamental frequency estimation, we are mostly interested in voiced speech. The *un*voiced speech is of interest only for voicing decisions, if included in a PDA. This voicing decision, however, is a matter of some urgency, as voicing false positives may present un-estimable frames as estimation errors, or hide difficult sections behind false negatives.

The bias towards voiced and unvoiced speech in a corpus is therefore very relevant for comparing corpora. Figure 9.6 summarizes the amounts and ratios of voiced and unvoiced speech in each corpus. All of these results used the *consensus truth* (introduced in Chapter 10) as their fundamental frequency estimate and voicing decision, to have a comparable data base.

As the data shows, the ratios between voiced and unvoiced speech vary greatly between corpora, from 3/2 in *FDA* and *PTDB-TUG* to 7/2 in *CMU-ARCTIC* and *MOCHA-TIMIT*.

For training computationally complex PDAs, it might make sense to select a corpus with a high

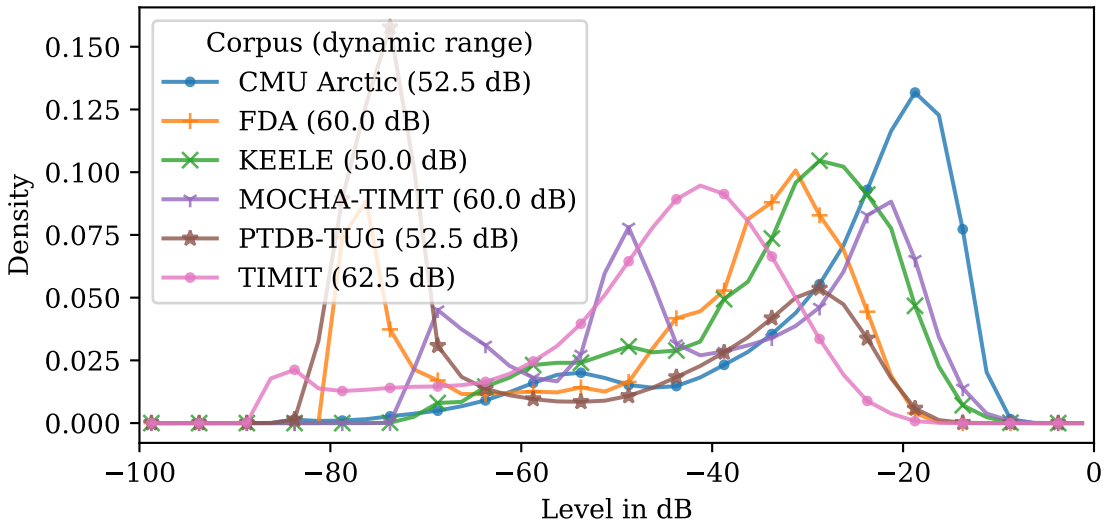


Figure 9.5: Level histograms of 50-ms blocks of each corpus, with level range in the legend. The level range is the level difference between silent parts and speech, as measured by the level difference between the first and last density greater than one percent.

voiced-to-unvoiced ratio in order to optimize the computation time on useful operations. For training voicing determination algorithms, however, a more balanced corpus might be more sensible.

9.8 Fundamental Frequencies

What is crucial to algorithmic pitch determination is the fundamental frequencies themselves. Figure 9.7 shows histograms of fundamental frequencies in the corpora, again determined using the *consensus truth* (see Chapter 10) for comparability.

Clearly, all corpora show two distinct maxima for male and female voices, respectively. The relative height of the maxima, however, varies between corpora, indicating different distributions of male and female voices. *PTDB-TUG*, *KEELE*, *FDA*, and *MOCHA-TIMIT* are more or less balanced, whereas *CMU-ARCTIC* and particularly *TIMIT* show significantly more male voices than female voices.

This should be an important consideration, not only for gender equality, but also because PDAs frequently estimate low or high voices with different accuracy. For comparisons between different PDAs, such a biased corpus would skew results unrealistically in favor of PDAs that deal better with matching voices. If training a PDA, it might actually impose the corpus bias onto the PDA. Such a gender imbalance is thus a serious matter, particularly in the widely cited *TIMIT* corpus, where the bias is strongest.

Lastly, the graph shows the curious case of the *FDA* corpus, where the female voice seems unreasonably high-pitched.

Instead of seeing this high-pitched voice as a defect, however, it might in fact be beneficial to view the absence of such voices in the other corpora as the true problem. All of their sentences were spoken by very normal voices, thus under-representing exceptionally deep male voices, or especially high female voices, or even higher children's voices in the process. Algorithms trained and evaluated with these corpora are unlikely to work well for such outlier voices, creating an unnecessary diversity issue. It might therefore be a great research opportunity to assemble a more diverse corpus specifically for fundamental frequency estimation that explicitly includes such outliers.

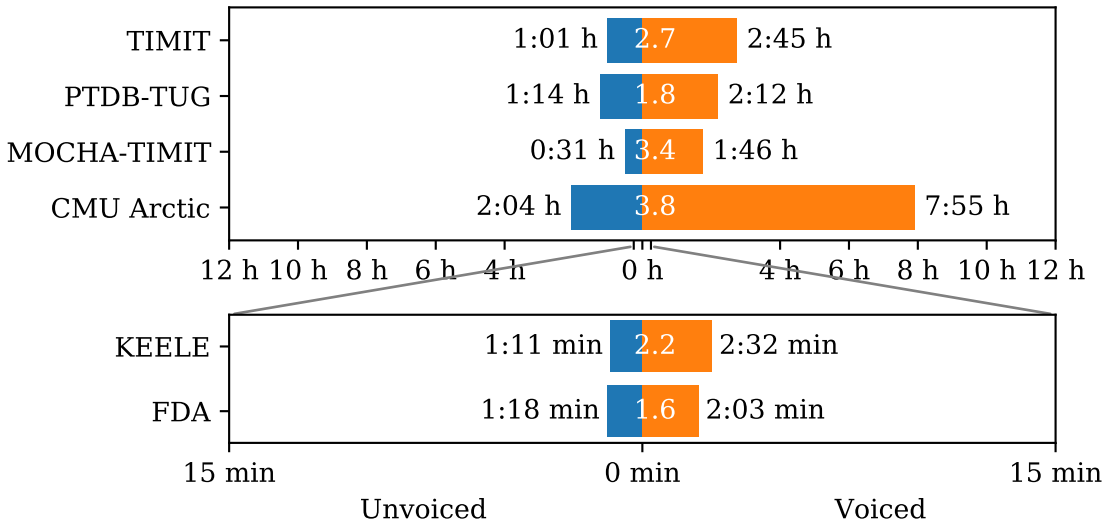


Figure 9.6: Unvoiced speech (blue) and voiced speech (orange) of various speech corpora, with the ratio between the two in white text in the center. Note the time scale difference between the top and bottom bars.

Another aspect of fundamental frequencies in speech is their change over time. Figure 9.8 shows a histogram of changes in pitch between consecutive time frames. In general, intra-utterance pitches appear to decrease more than they increase in all corpora. This effect is slightly stronger for *CMU-ARCTIC* and slightly weaker for *PTDB-TUG* and *TIMIT*. One oddity, however, is *TIMIT*, whose maximum is significantly stronger than the other corpora's, indicating more monotone voices in this corpus than in the other corpora.

9.9 Background Noises

In many comparison studies of PDAs, their performance is evaluated with variable background noise at a range of SNRs. Two of the most popular noise databases used for this purpose are *NOISEX*, and *QUT-NOISE*.

NOISEX [155]

`http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`

This noise corpus was designed for studying the effect of additive noise on speech recognition systems. It contains 15 recordings of about one minute each, taken from the RSG.10 noise dataset [144]. Among these recordings are noise recordings from a machine gun, the noise inside various military vehicles and airplanes, as well as factory floor noise, car noise, office noise, babble noise, and synthetic pink and white noise, each recorded at 20 kHz.

QUT-NOISE [26]

`https://research.qut.edu.au/saivt/databases/qut-noise-databases-and-protocols/`

The rather larger *QUT-NOISE* background noise corpus consists of 13:39 hours of various natural background noises recorded at 48 kHz in stereo for evaluating voice activity determination algorithms. This includes 20 half-hour-long recordings of cafés, kitchen and living room ambient noises, multiple

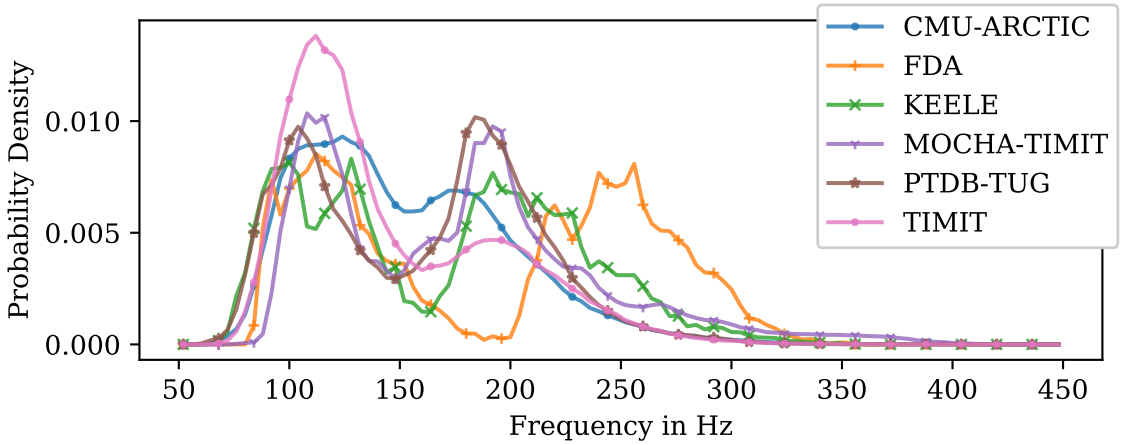


Figure 9.7: Histogram of fundamental frequencies of the speech recordings in each corpus.

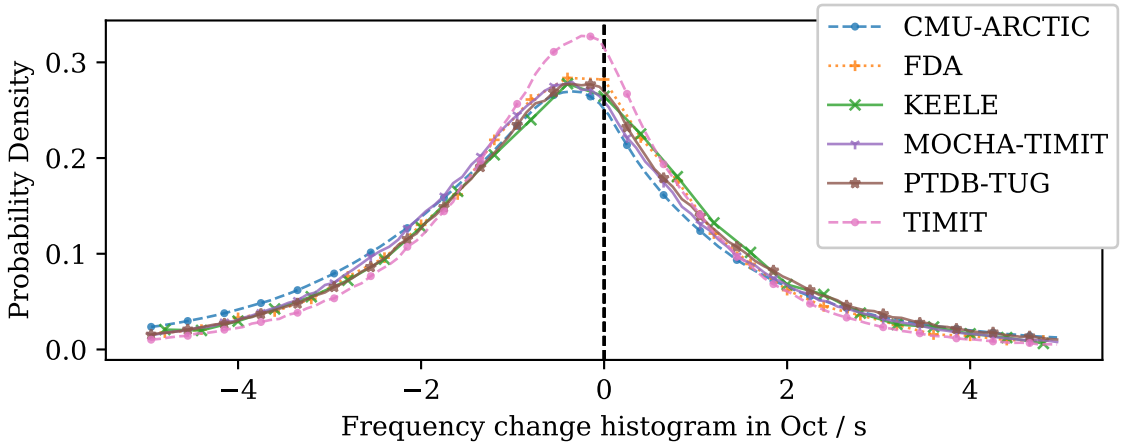


Figure 9.8: Histogram of fundamental frequencies changes of the speech recordings in each corpus, normalized to octaves per second.

street noise recordings, car noises with open and closed windows, as well as reverberant parking lot and swimming pool noises. Some recordings include preliminary calibration sweeps at the very beginning of the recording. Care should be taken to not include these in evaluations.

Figure 9.9 shows long-term spectra of these noise databases, as well as spectra of each noise recordings therein. In general, *NOISEX* are louder on average and somewhat more variant between recordings. Some of the *NOISEX* recordings feature strong tonal elements (jet engine whine) above the range of speech fundamental frequencies. Listening to the examples, each recording is very uniform and steady.

In comparison, *QUT-NOISE* recordings are more varied within each recording, with transient events such as honking cars or a droning truck engine driving by, or the clanging of a dropped utensil in the kitchen recording.

These characteristics are confirmed in Figure 9.10, where the range of levels in *QUT-NOISE* is significantly broader than in *NOISEX*. The graph also shows the few outlier files in *NOISEX*, which are less loud than the others in the lower maximum near -25 dB. The minor maximum at -65 dB is an artifact of the machine gun noise, which is relatively silent between each shot.

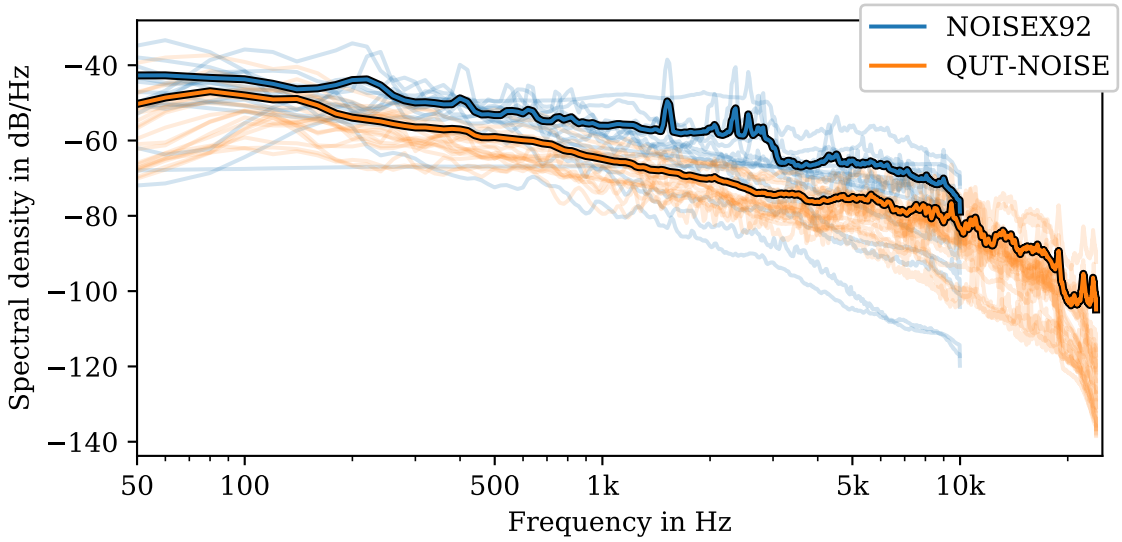


Figure 9.9: Long-term average spectra of two background noise databases, with spectrum of each noise recording as faint lines.

Thus the more stationary *NOISEX* recordings are particularly useful for small-scale, reproducible evaluations (and military applications), whereas the more varied *QUT-NOISE* provides a more complete assessment of the vagaries of real-world background noises useful for larger-scale studies.

9.10 Conclusions

The choice of speech and noise corpora for training and evaluating fundamental frequency estimation algorithms is an important one. As this chapter has shown, there are significant differences between the corpora, which makes cross-corpus comparisons difficult and might bias PDAs trained with any one particular corpus.

If only one speech corpus had to be chosen, it would probably be *PTDB-TUG*, as it contains a sizable amount of data, with a diverse speaker set and sentence list, has a neutral gender distribution, and has no obvious other defect. By a similar argument, *QUT-NOISE* is a reasonable choice for a noise corpus.

However, it must be stressed again that all of these corpora are biased strongly towards WEIRD voices, as in *Western, Educated, Industrialized, Rich, and Democratic*. There are few strong dialects, colloquial idioms, or extraordinary voices in these corpora, and the noise corpora are clearly optimized for the WEIRD world that also provides the funding and framework for this dissertation. The corpora also do not include shouting voices, emotional voices, whispering, or singing voices, or indeed languages other than English. While likely not a major problem for fundamental frequency estimation and voice activity determination technologies, it is nevertheless clearly a far cry from actually capturing all the full varieties of human speech, as we humans understand it.

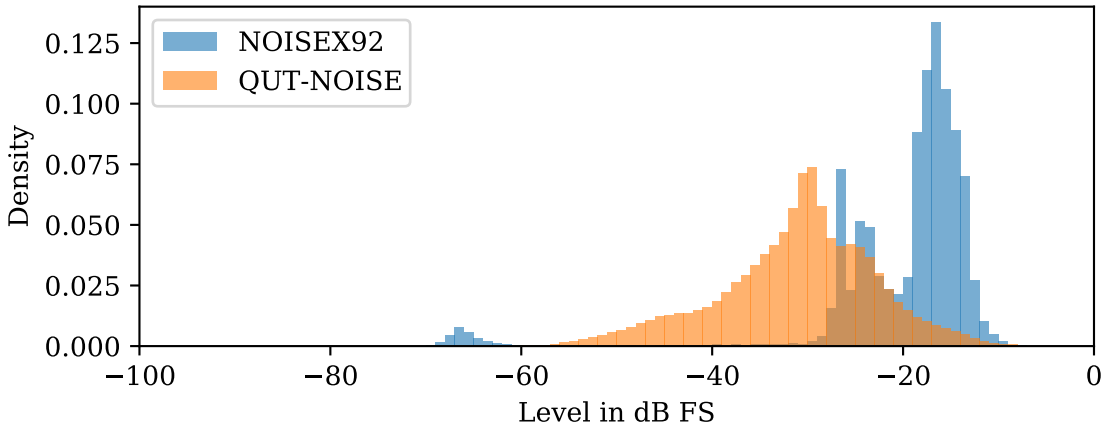


Figure 9.10: Loudness histogram of 50-ms blocks of both corpora.

Chapter 10

Consensus Truth

Abstract

A large number of PDAs have been developed as features for various speech processing tasks such as speech recognition, speaker identification, and speech compression. To evaluate the accuracy of these algorithms, their estimates can be compared against a known ground truth fundamental frequency. However, the ground truths in typical speech corpora were calculated by single reference PDAs, inheriting their biases, and differing between corpora. Furthermore, ground truths are typically derived from laryngograph recordings, whose fundamental frequency activity is different from that of acoustic recordings. We therefore propose a new method for deriving a fundamental frequency ground truth from the consensus of a number of state-of-the-art fundamental frequency estimation algorithms, which can be calculated from acoustic recordings alone, is more robust than a single algorithm's estimate, and not subject to the ambiguities of laryngograph estimates.

10.1 Introduction

The pitch of the human voice is an essential characteristic of speech and communication. It is, however, a human perception that cannot be measured objectively; Instead, we rely on estimating the fundamental frequency of voiced speech using computer algorithms. Thus without access to an objective truth, this raises the question of how to define a truth for evaluating such algorithms.

Common speech corpora that include a fundamental frequency ground truth [122, 5, 4, 119] derive them from laryngograph measurements that record the vibrations of the vocal folds at the origin of voiced speech using adapted PDAs. However, vocal fold vibrations do not necessarily lead to sounds with a unique fundamental frequency: complex vocal tract movements, such as phoneme transitions, onsets, and offsets, can obscure the fundamental frequency such that the acoustic emission is unvoiced or multi-pitched, even though the vocal folds vibrate differently [110].

Figure 10.1 shows both an acoustic recording and a laryngograph recording of a short sentence, to illustrate these differences. Onsets and offsets, particularly the onset of "had" and the offsets of "she" and "your" are more clearly harmonic in the laryngograph recording than in the acoustic recording, whereas some harmonicity remains visible in the acoustic recording at the end of "dark" and "suit" that is missing in the laryngograph. The laryngograph recording additionally shows some pitch doublings in "she", and the ends of "dark" and "suit" that is less pronounced in the acoustic recording. For reasons such as these, laryngograph-based fundamental frequency ground truths are not ideal for evaluating PDAs.

Another approach to evaluating PDAs, especially in noise, is to rely on a reference PDA's estimate of clean speech signals as the ground truth [32, 67, 101, 83, 3]. This enables the use of speech corpora

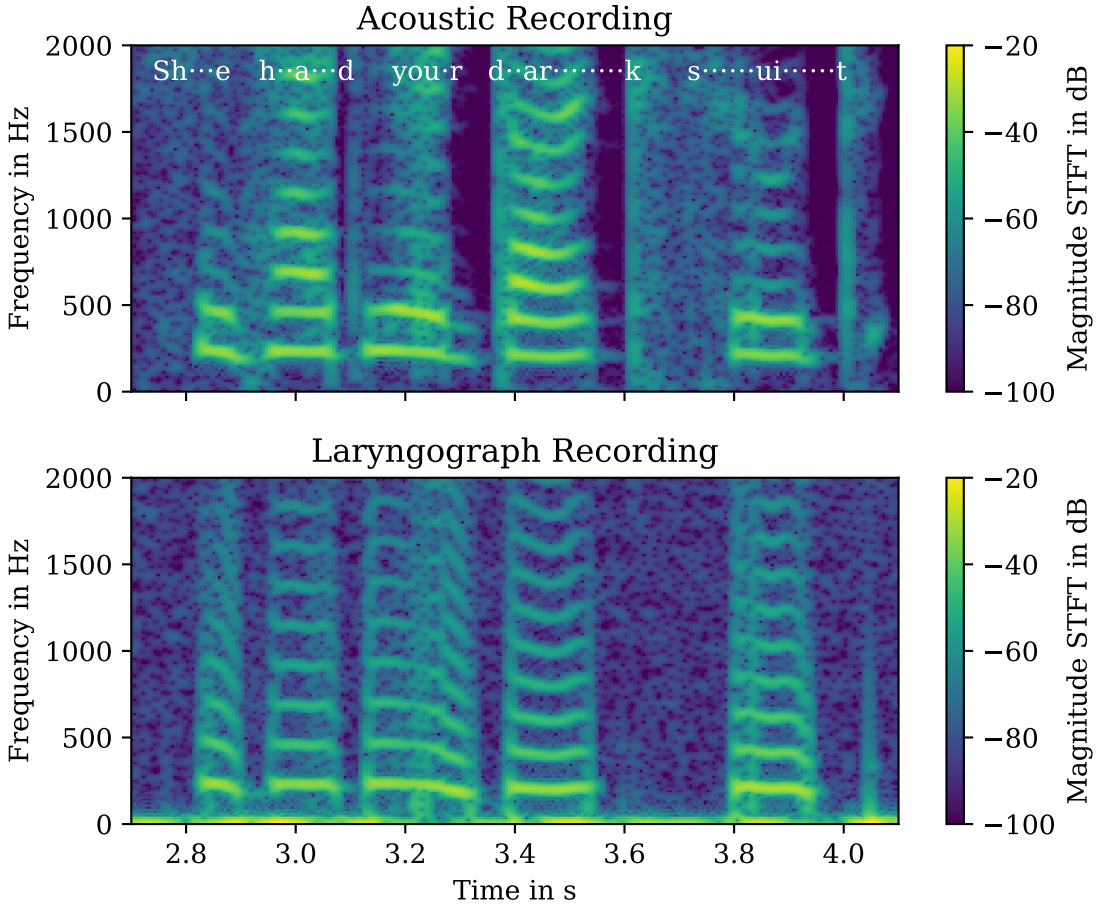


Figure 10.1: Magnitude STFTs of an acoustic speech recording (top) and a laryngograph recording (bottom). Laryngograph shows more detailed frequency tracks at onsets and offsets, especially near 2.9 s and 3.9 s, and significantly different harmonicity near 3.3 s. The signal is sample F06/sa1 from the *PTDB-TUG* corpus of a woman saying "She had your dark suit in greasy wash water all year".

without an existing ground truth or laryngograph recordings, including special-purpose corpora such as pathological voices [77, 134] or foreign languages [19, 57]. However, the quality of such a ground truth is inherently dependent on the PDA used as reference, and will inherit its biases.

This chapter introduces a new fundamental frequency ground truth that is solely derived from acoustic recordings and does not exhibit ambiguous laryngograph pitches and algorithm-specific idiosyncrasies by requiring multiple PDAs to agree on a common fundamental frequency. This measure, henceforth called *consensus truth*, can be calculated from arbitrary clean speech recordings, and should reflect the acoustic tonality of speech better than laryngograph-derived fundamental frequency ground truths. This study provides pre-calculated consensus truth tracks for a number of popular speech corpora, namely *FDA* [5, 4], *KEELE* [122], *MOCHA-TIMIT* [168], *PTDB-TUG* [119], and *TIMIT* [40]. These results are available on the companion website to this dissertation at `https://bastibe.github.io/Dissertation-Website/`.

The selection of PDAs for calculating the consensus truth is bound to be somewhat arbitrary, as fundamental frequency estimation is still an active area of research with new PDAs being developed regularly. For this study, PDAs needed to provide reference implementations, accurately estimate the fundamental frequency of arbitrary speech signals, and have a reasonable run time for short audio

samples.

Recent comparison studies [147, 148, 3, 31] selected *BANA* [102], *DNN* [50], *PEFAC* [45], *PRAAT* [9], *RAPT* [150], *SRH* [31], *STRAIGHT* [73], *SWIPE* [18], *MBSC* [151], *YAAPT* [69], and *YIN* [24] as the most reliable PDAs for various speech analysis tasks. Recent publications on fundamental frequency estimation additionally used *CREPE* [79], *DIO* [101], *KALDI* [42], *SACC* [86], *SAFE* [22], *SIFT* [93], *SHR* [149], and *NLS* [103] as common reference PDAs [67, 163, 66, 1, 2, 32]. All of these PDAs provide reference implementations and were shown to estimate fundamental frequencies reliably for clean and noisy speech signals.

Each of these PDAs are based on reasonable signal models that were developed with a particular application in mind, such as various measures for harmonicity or periodicity in a number of different signal representations. The consensus truth represents a sort of grand average of these underlying concepts, and thus a more holistic view of fundamental frequency estimation across multiple areas of application and theories of operation.

10.2 Methods

To calculate the consensus truth, fundamental frequency estimates from all PDAs in Table 10.1 were calculated for every clean speech recording from the speech corpora *FDA* [5, 4], *KEELE* [122], *PTDB-TUG* [119], *TIMIT* [40],and *MOCHA-TIMIT* [168]. These corpora were selected for being widely used for fundamental frequency estimation tasks [163, 2, 66, 1, 102, 69, 18, 147, 148, 31], as detailed in Chapter 9. Three corpora, *KEELE*, *FDA*, and *PTDB-TUG*, already include a fundamental frequency ground truth. The two remaining corpora, *TIMIT* and *MOCHA-TIMIT*, do not, and are commonly used with another PDA as ground truth[45, 16, 88, 166].

For each speech recording in every corpus, fundamental frequency estimates of all PDAs were latency-corrected and resampled to a common time base of one estimate every millisecond, which was chosen significantly higher than the PDAs' time bases to account for timing differences. Latency correction is necessary, since various PDAs position their fundamental frequency estimates at either the start or the center of each block. PDA latencies were estimated by minimizing their estimation errors for various lags and short modulated tone complex of a known fundamental frequency. Estimates were assumed unvoiced where the ground truth voicing probability was smaller than 50 % or no fundamental frequency was available.

To calculate the consensus truth, we first define a consensus predicate $C^{(n)}(t)$ that checks whether the fundamental frequency estimate $\widehat{f}_0^{(n)}(t)$ of the nth PDA at time t is within a ± 20 % consensus range of the median estimate $P_{0.5}(t)$ over all PDAs, similar to the commonly used Gross Pitch Error [127] measure:

$$C^{(n)}(t) = \left| 1 - \frac{\widehat{f}_0^{(n)}(t)}{P_{0.5}(t)} \right| < 0.2 \tag{10.1}$$

A consensus voice activity decision $\overline{\mathrm{VAD}}(t)$ at time t is then defined as the fraction of PDAs that estimated a voiced fundamental frequency within the consensus range:

$$\overline{\mathrm{VAD}}(t) = \frac{1}{N} \sum_{n=1}^{N} \left[\!\!\left[C^{(n)}(t) \wedge \mathrm{VAD}^{(n)}(t) \right]\!\!\right]_{\mathrm{I}} \tag{10.2}$$

where N is the number of all PDAs, $\mathrm{VAD}^{(n)}(t)$ is the binary voicing decision of the nth PDA, and $[\![\cdot]\!]_{\mathrm{I}}$ is the Iverson bracket, which is 0 or 1 depending on the logical proposition inside.

Table 10.1: Fundamental frequency estimation algorithms used to calculate the consensus truth

Name	URL to download
AUTOC [140]	reimplemented from publication
AMDF [130]	reimplemented from publication
BANA [102]	http://www2.ece.rochester.edu/projects/wcng/code.html
CEP [105]	reimplemented from publication
CREPE [79]	https://github.com/marl/crepe
DIO [101]	http://www.kki.yamanashi.ac.jp/~mmorise/world/english/
DNN [50]	http://web.cse.ohio-state.edu/pnl/software.html
KALDI [42]	https://github.com/LvHang/pitch
MAPS (Chapter 8)	https://bastibe.github.io/Dissertation-Website/maps/index.html
MBSC [151]	http://www.seas.ucla.edu/spapl/shareware.html
NLS [103]	https://github.com/jkjaer/fastF0Nls
PEFAC [45]	http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
PRAAT [9]	https://github.com/praat/praat
RAPT [150]	http://www.speech.kth.se/wavesurfer/links.html
SACC [86]	http://labrosa.ee.columbia.edu/projects/SAcC/
SAFE [22]	http://www.seas.ucla.edu/spapl/weichu/safe/
SHR [149]	https://mathworks.com/matlabcentral/fileexchange/1230
SIFT [93]	reimplemented from publication
SRH [31]	https://github.com/covarep/covarep
STRAIGHT [73]	https://github.com/HidekiKawahara/legacy_straight
SWIPE [18]	http://www.cise.ufl.edu/~acamacho/english/curriculum.html
YAAPT [69]	http://www.ws.binghamton.edu/zahorian/yaapt.htm
YIN [24]	http://audition.ens.fr/adc/

A frame is defined as *voiced* by majority vote if $\overline{\mathrm{VAD}}(t) \geq 0.5$, and *unvoiced* otherwise. PDAs that do not have a VAD are assumed to classify every frame *voiced*, and thus determine voicing based solely on $C^{(n)}(t)$.

The consensus fundamental frequency $\overline{\mathrm{f}_0}(\mathrm{t})$ is then selected as the mean of all estimates within the consensus range:

$$\overline{\mathrm{f}_0}(\mathrm{t}) = \overline{\mathrm{VAD}}(t) \sum_{n=1}^{N} \hat{f}_0^{(n)}(t) \left[\!\!\left[C^{(n)}(t) \wedge \mathrm{VAD}^{(n)}(t) \right]\!\!\right]_{\mathrm{I}} \tag{10.3}$$

where $1/\overline{\mathrm{VAD}(t)}$ is the number of all voiced PDAs.

10.3 Evaluation and Discussion

The fundamental frequency consensus truth was designed as a reproducible and reliable method for estimating the fundamental frequency of clean speech recordings for the purpose of evaluating the performance of fundamental frequency estimation algorithms. It should therefore behave similarly to the ground truths available in *PTDB-TUG*, *KEELE*, and *FDA*, yet not be subject to some of their shortcomings. Small differences in voicing decisions and fundamental frequency are to be expected due to numerical variations, and the differences between microphone and laryngograph recordings.

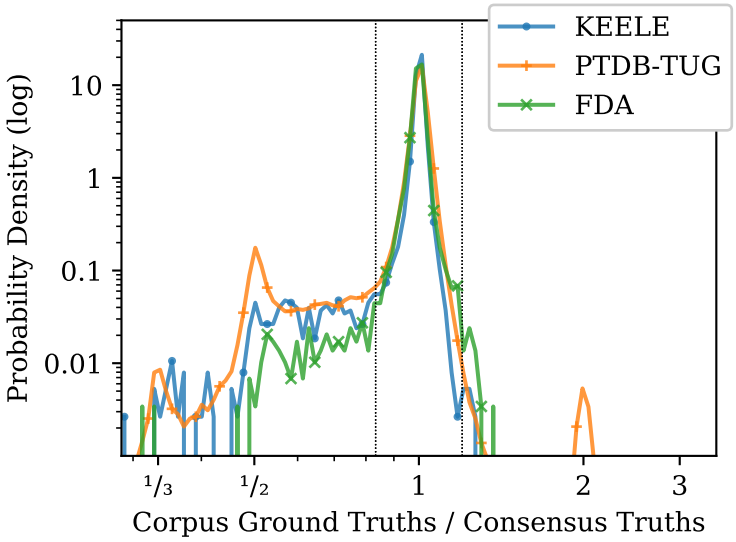


Figure 10.2: Probability density of fundamental frequency differences as the quotient of the corpora's ground truths and the consensus truth. Dotted vertical lines mark ±20%, or gross pitch errors.

A summary of these differences between the corpora's ground truths and the consensus truth are shown in Figure 10.2. The graph shows the probability distribution of estimation differences as the quotient between the corpora's ground truths and the consensus truth, for frames where both truths are voiced. Differences are mostly small, with a slight bias towards lower frequencies in the ground truths, and small but notable octave errors in *PTDB-TUG*'s ground truth.

Table 10.2 interprets these differences as gross and fine pitch errors as if the consensus truth were a PDA: gross pitch differences denote the percentage of consensus truth fundamental frequencies within ±20% of the ground truth. Fine pitch differences are the mean absolute difference within ±20%, for frames where both truths are voiced. The table shows gross pitch differences to be generally rare, and fine pitch differences small, with the biggest differences for *PTDB-TUG*, due to the octave errors shown in Figure 10.2. These metrics are well below the margin of error for evaluating fundamental frequency estimation algorithms in noise but may be significant for clean speech evaluations.

Table 10.2: Fundamental frequency estimation differences between the consensus truth and the corpora's ground truths. Gross differences are the percentage of frequencies outside ±20% of the consensus truth, and fine pitch differences are the mean estimation difference within ±20%.

Corpus	Gross differences	Fine differences
PTDB-TUG	3.5 %	2.16
KEELE	2.3 %	1.39
FDA	1.0 %	1.72

Voicing decision differences are summarized in Table 10.3. In general, variations between the corpora's ground truths' voicing decisions and the consensus truth's voicing decisions are small. All three corpora's ground truths label slightly more frames as voiced than the consensus truth. This is as expected, since the ground truths' laryngograph recordings can be periodic during onsets, offsets, and phoneme transitions, while the rest of the vocal tract is not yet in resonance, and the acoustic speech signal is not yet periodic. The opposite case of consensus-positive decisions for ground truth-negative

frames are less frequent. The overall number of voiced frames in the consensus truth is smaller than in the corpora's ground truths by 0.21 % for PTDB-TUG, 0.63 % for FDA, and 3.01 % for KEELE. As these frames are deemed unvoiced by the majority of PDAs, this reduction is probably justified, and will result in fewer spurious voicing detection errors when evaluating PDAs.

Table 10.3: Voicing decision comparison of the corpora's ground truths and the consensus truth.

Consensus	*PTDB-TUG*		*KEELE*		*FDA*	
	Voiced	Unvoiced	Voiced	Unvoiced	Voiced	Unvoiced
Voiced	20.15 %	3.19 %	44.83 %	1.79 %	35.32 %	2.04 %
Unvoiced	3.40 %	73.26 %	4.80 %	48.58 %	2.67 %	59.98 %

The data in Tables 10.2 and 10.3 clearly highlights the advantage of the consensus truth over laryngograph-derived ground truths: between 5.7 % and 9.5 %[1] of corpus ground truth estimates are considered gross pitch errors and VAD errors by the majority of PDAs. When evaluating PDAs against these ground truths, these errors would be attributed to the PDAs. The consensus truth should therefore result in fewer errors in general, and a more truthful representation of PDA accuracy.

The consensus truth is furthermore preferable to any one individual PDA's estimates of clean speech recordings, as it is not susceptible to individual algorithms' biases. Table 10.4 shows the amount of gross pitch errors of each algorithm when judged against each corpus' ground truth as well as the consensus truth. Every PDA shows at least a small amount of GPEs compared to the consensus truth[2], indicating that every PDA's GPEs are different. The consensus is therefore indeed required for suppressing these individual GPEs, and can not be replaced by any single PDA's estimate.

Furthermore, GPEs in Table 10.4 are consistently and significantly lower when judged against the consensus truth instead of the corpora's ground truth. The differences in PDA GPEs are similar to the GPE differences between truths in Table 10.2, indicating once more that the corpus ground truths indeed deviate from what can be estimated from their speech signals. In contrast, the consensus truth is by design more consistent with PDAs, and therefore a better target for PDA evaluations.

10.4 Conclusions

In order to evaluate fundamental frequency estimation algorithms, a reliable ground truth is required. However, such ground truths are only available for select speech corpora and typically rely on single reference PDAs. Additionally, these ground truths are typically derived from laryngograph recordings, which differ slightly from acoustic recordings when used to derive fundamental frequency estimates. The consensus truth offers a solution to both of these problems by being more consistent across corpora, and being calculable from any acoustic speech recording without requiring laryngograph recordings.

The consensus truth was shown to be a superior replacement for laryngograph-derived ground truths and is provided as part of this dissertation for a number of widely-used speech corpora for speech analysis tasks, including some that do not have a fundamental frequency ground truth of their own. The evaluation proved that the consensus truth is sufficiently similar to existing ground truths, while avoiding some of their biases. It is thus a valid replacement for existing ground truths, and a better choice than any single reference PDA for calculating fundamental frequency ground truths of future speech datasets.

Source code and instructions on how to calculate the consensus ground truth for further corpora are provided at this dissertation's companion website at `https://bastibe.github.io/Dissertation-Website/`.

[1] the sum of both kinds of VAD errors and GPEs

[2] except for SWIPE, which is due to a problem examined in Chapter 12

Table 10.4: Fundamental frequency estimation accuracy in GPE for clean speech recordings of PDAs for various corpora and ground truths. Consensus truth GPEs are consistently lower than corpora's ground truths.

Truth	*PTDB-TUG*		*KEELE*		*FDA*	
	Corpus	Consensus	Corpus	Consensus	Corpus	Consensus
AMDF	21.99 %	11.95 %	15.78 %	10.85 %	16.71 %	11.91 %
AUTOC	17.99 %	14.33 %	22.73 %	20.79 %	40.25 %	37.93 %
BANA	16.40 %	7.27 %	11.60 %	6.21 %	8.33 %	4.03 %
CEP	44.25 %	38.70 %	37.69 %	33.57 %	37.70 %	35.64 %
CREPE	3.74 %	1.03 %	1.86 %	0.56 %	0.81 %	0.05 %
DIO	3.44 %	0.43 %	1.35 %	0.40 %	0.66 %	0.11 %
DNN	8.26 %	5.60 %	9.99 %	6.77 %	7.89 %	5.52 %
KALDI	1.77 %	1.33 %	0.43 %	0.15 %	0.43 %	0.16 %
MAPS	1.98 %	0.11 %	1.15 %	0.19 %	1.32 %	0.09 %
MBSC	1.65 %	0.33 %	0.81 %	0.11 %	0.51 %	0.02 %
NLS	23.45 %	17.01 %	15.21 %	11.69 %	12.05 %	8.22 %
PEFAC	19.32 %	16.98 %	12.79 %	9.00 %	4.16 %	2.44 %
PRAAT	4.32 %	2.08 %	2.09 %	0.88 %	1.89 %	0.32 %
RAPT	5.91 %	4.54 %	5.05 %	3.90 %	5.40 %	4.09 %
RNN	8.76 %	6.46 %	9.56 %	6.68 %	7.49 %	5.45 %
SACC	3.61 %	0.50 %	2.86 %	0.87 %	1.25 %	0.03 %
SAFE	3.33 %	0.50 %	2.40 %	0.09 %	2.70 %	0.28 %
SHR	6.97 %	2.09 %	1.48 %	1.12 %	8.68 %	4.51 %
SIFT	31.29 %	22.91 %	18.86 %	14.78 %	20.15 %	17.15 %
SRH	2.92 %	0.66 %	1.63 %	0.66 %	3.46 %	2.16 %
STRAIGHT	3.41 %	0.57 %	1.96 %	0.11 %	1.78 %	0.26 %
SWIPE	1.21 %	0.00 %	0.05 %	0.03 %	0.07 %	0.00 %
YAAPT	19.67 %	14.25 %	15.52 %	10.91 %	21.31 %	17.25 %
YIN	11.54 %	2.15 %	5.76 %	1.52 %	5.60 %	1.44 %

10.5 Acknowledgments

The author would like to thank Christina Imbery, Grace Gahman, John Goodyear, Menno Müller, and Ulrik Kowalk for language editing, proofreading and insightful discussions about this chapter.

Part V

Evaluating Fundamental Frequency Estimation Methods

Where the true purpose of this dissertation is revealed in an unprecedented comparison of algorithms, databases, and truths. This comparison is so large in scope that it is broken up into two chapters:

Chapter 11 introduces the algorithms to be compared in a review of fundamental frequency estimation's varied history over the last thirty years, as well as a quantitative literature survey of their scientific reception and significance. The chapter also rigorously defines the performance measures used for the comparison, and explicitly defines the full breadth of the comparison dataset.

Chapter 12 finally conducts the comparison of algorithms, corpora, and truths that the whole dissertation has been preparing. The unique quantity of data available to this study allows for an equally unique depth of analysis, with error measures both traditional and novel that reveal a number of hitherto unknown properties in all algorithms.

Quite contrary to our initial intentions, we found this meta-analysis of fundamental frequency estimation of greater interest than the algorithms themselves. The fundamental frequency of speech, as we discovered, is not as rigorously defineable as originally thought, and its estimation is therefore an art as well as a science, with no definitive way of defining its beauty.

Chapter 11

A Replication Dataset for Fundamental Frequency Estimation

Abstract

The previous chapters introduced pitch and the fundamental frequency of speech as general concepts. Afterwards, a concrete algorithm for estimating them was outlined, followed by various databases and a ground truth for comparing estimators. Thus, the stage is set to conduct a thorough comparison of estimators.

This chapter introduces the "rules" of the comparison: It establishes the algorithms and implementations used for the comparison, and a set of error measures, databases, and ground truths. The result is a database of algorithms, signals, and fundamental frequency estimates, as well as pre-computed error measures and ground truths that can be used for comparisons with new algorithms so as to replicate existing studies, and conduct comparison studies within the dataset.

11.1 Introduction

The fundamental frequency of speech appears to be a complete and specific area of research, apparently with a clearly defined estimation target, and a number of simple and obvious estimation algorithms. Yet, the research community has published hundreds of papers on this topic over the last sixty or so years, with no end or definitive solution yet in sight.

With such an abundance of knowledge on this topic, it is perplexing that no definitive solution has yet been found. The problem is the endless variability of human perception. Speech is inherently a human signal and any definition of its properties is inherently related to how we humans perceive them. Any purely technical definition and estimation of *fundamental frequency* is unlikely to agree with our perception, and is therefore deemed inadequate for speech, while perhaps technically correct. A more "human" definition must lie somewhere in between *fundamental frequency* and *pitch*, both rigorously definable and close to human perception. But like any human concepts, there is considerable leeway between these ideals, with no obvious ground truth. Where there is no truth, there cannot be a definitive solution, either.

Thus, PDAs must strike a balance between perception and math. Each PDA must define its own model of speech, in which its own implementation can be said to be optimal, or at least *closer* to optimal than previous PDAs. These signal models can be surprisingly varied. In general, they are typically either based in speech production, where regular glottis pulses produce a periodically self-similar signal, or they are modeled on speech perception, where the speech signal is made from

harmonically related phase-locked sinusoids. Some have argued that these models are mathematically related [21], but practical implementations typically are not and result in distinct characteristics.

Furthermore, PDAs can target different applications, which prioritize different algorithm characteristics. For example, some applications such as automotive or aviation, can prioritize certain kinds of environmental noises. A focus on singing voices requires high frequency resolution, but does not need as much robustness to noise. Applications in forensics might favor the exact opposite behavior. Besides these target criteria, some applications have to work with limited computational resources, or within a certain margin of delay from an audio recording.

Comparisons between PDAs, then, must somehow define a common denominator for all these diverging ideas about pitch, even though doing so is necessarily putting more exotic PDAs at a disadvantage. But as publications nowadays require claims of "being the best" at all costs, even a flawed comparison is more important than ever, unfairly or not.

Such claims of a "novel", "best" PDA should always be interpreted as the introduction of a new and exciting definition of a signal model for fundamental frequency, and not so much as a claim of actual supremacy in a comparison study. Conversely, comparison studies should focus on *differences* instead of ordered rankings.

Thus, there is still value in comparison, so long as the testing conditions are clearly defined and varied enough to capture the intricacies of each PDAs' behavior. While there cannot be a "best" PDA in general, the lesser claim of being "best at X" should still be valid and useful. Moreover, the usefulness of such a comparison is probably proportional to the number of "Xs" considered, to draw as complete a picture of each PDA's strengths and weaknesses as possible.

Solving the comparison for a large number of Xs, requires an equally large data set. A dataset as varied as possible, including many different PDAs, speech signals, and noise signals. A dataset, in other words, the likes of which has not been attempted before.

This chapter describes the construction of such a dataset, its components, and how to reproduce it. This "replication dataset" can recreate most published smaller-scale comparison studies, at least for the PDAs included. Moreover, future PDAs could opt to simply compare to the replication dataset instead of running their own comparisons, so long as they agree to use the same error measures and signals for evaluation.

11.2 Algorithm Availability

With over 800 publications available on fundamental frequency estimation algorithms in the last thirty years alone[1], but only limited computation time and limited human resources available, a choice has to be made. Only a small subset of all PDAs can be included in a practical dataset. Necessarily, these must have an implementation available, a reasonable run time, and be of general interest. The latter point might mean being widely cited, of historical significance, or an otherwise noteworthy reference point of the state of the art.

A pre-existing selection of such PDAs can be found in existing comparison studies, and in the evaluations of newly published PDAs if they include a comparison. These PDAs have the benefit of having already been found suitable for comparison studies, and having been integrated into a framework for comparisons at least once.

Integrating a large number of these PDAs into the larger framework for this study, however, still proved extraordinarily difficult. Old PDAs in particular often used severely outdated programming environments that are no longer reproducible on modern operating systems. Even where the programming languages and build tools are still around, years of software updates often lead to obscure bugs

[1] See chapter 9

or crashes. In a few cases, these could be fixed, which will be noted in their detailed discussions later, but in other cases, the problems proved insurmountable and the algorithm had to be dropped.

In general, implementations in high-level programming languages such as Matlab or Python proved easier to work with than lower-level languages such as C or Java. This is mostly due to their use of high-level abstractions such as built-in linear algebra frameworks, whereas these tend to come as independent modules in lower-level languages that must be installed separately and can be problematic in their own right. Also, high-level run-time errors were easier to debug in interpreted languages than the compile-time errors and separate (and archaic) debuggers of lower-level languages, despite the author's extensive experience in both kinds of systems.

Where no implementations of historically significant algorithms were available, re-implementations were attempted. However, even widely cited, basic algorithms were found to lack important parameter values or decision thresholds in their original publications; hence, this endeavor had to be limited to only the most critical PDAs, and even then leave out non-essential parts such as voicing determination systems.

A number of algorithms were additionally found to run acceptably fast only on the small data sets of their original publications, but could not be run on even dozens of speech samples due to memory leaks, or excessive memory use while running[2]. More generally, performance optimizations in scientific algorithms were found to be a constant source of problems, ranging from simple unmet assumptions such as fast GPUs or availability of multiple CPUs, to excessive hard drive caching, to the myriad errors of C-based Mex files[3]. The latter, particularly, wreaked all sorts of unintended havoc including, but not limited to, memory access violations, leaking temporary files, leaking zombie processes, infinite-looping, continuously consuming a system resource until the process crashes, crashing other programs, crashing the operating system, and destroying file systems. Suffice it to say that performance optimizations should be avoided wherever possible in scientific algorithms.

It should be explicitly noted that none of this behavior was the respective authors' intention, but merely a consequence of inexpert programming. The fact that such optimizations actively detract from future use of the algorithm, and therefore hinder a wider adoption and citation, is, however, ultimately damning. Where published scientific programs are concerned, simpler and higher-level programs were found vastly preferable to highly "optimized" low-level ones. This is particularly true as most of the run-time performance differences do not matter with hardware years in the future anyway, and the vast parallelism of a current-day computer cluster makes most polynomial-time performance differences between algorithms unimportant.

Regardless, 25 PDA implementations met the above criteria and will be detailed in the following sections. These range from simple recreations of the first digital PDAs from the 1960s, to state-of-the art machine learning constructs from 2020 and should provide a broad overview of the various developments of pitch estimation research.

Finally, due to the considerable computation time required, the selected algorithms had to be run in a batch processing framework on a compute cluster to gather the dataset. However, the algorithms' volatile and error-prone run-time behavior precluded the use of common batch processing frameworks, as these often cannot deal with crashes or reboots, and can be difficult to integrate with Matlab, particularly. Additionally, it was found essential to run every algorithm in its very own process, as many of them could not be trusted to run correctly in a non-clean environment, where they would re-use temporary data from previous runs or simply crash. These complications required the construction of a bespoke crash-robust batch processing framework[4], an audio-friendly database for storing results

[2]Gigabytes per second of audio data were observed more than once

[3]Binary Matlab extensions written in C.

[4]Runforrest: `https://github.com/bastibe/RunForrest`

and datasets[5], and a custom Matlab-Python interface[6], which were since published as Open Source Software for others independently of this dissertation.

11.3 Selected Algorithms

In total, 25 PDA implementations were collected over the span of a few years. Most of these were originally obtained in late 2017, with some later additions when they became available and some replacements with more recent versions as they were released or found.

The following sections will introduce each algorithm in general terms, with a particular emphasis on their various goals and their historical context. To emphasize the latter, the algorithms are presented here in order of their publication.

To gain an impression of the time of the earliest fundamental frequency estimation algorithms included, they were still before the true integration of computers into the scientific workplace:

> "The development of the computer program described here proceeded in two stages. During the first stage, photographs of the spectrum analyses were taken on motion picture film for a number of spoken words; then, these analog spectra were converted to corresponding digital data by a manual data reduction procedure. Next, a manual program for the extraction of pitch from these data was written which consisted of a set of written instructions that an assistant was to follow in order to determine the pitch frequency from the digitized Fourier analyses. [...]
>
> During the second stage of the development of the program logic for pitch extraction, the manual program was converted to a computer program written in FORTRAN language. [...] The computer program was checked repeatedly against the data and corrected until the computer output was in complete agreement with the pitch obtained from the manual program." [53, p. 2]

It was a different time indeed from today's heavily computational, digital-first design methods.

A number of even earlier analogue PDAs were excluded from this study on the grounds that they left no digital implementations behind, and their accuracy has since been made obsolete by digital methods. A thorough examination of this DSP-prehistory can be found in [59].

The first four PDAs in the following list are included mostly out of historical interest, as they describe early archetypes that later PDAs will build upon. None of these original algorithms' source code could be found, so we partially re-implemented them based on the original publications. The voicing decision criteria in these algorithms were omitted, however, as important parameters and thresholds were not included in the original publications and hence deemed non-essential, as today's ground truths all include a "true" voicing decision that can be used in lieu of algorithmic ones.

11.3.1 *CEP* [105] (1967)

Re-implemented in Python, without voicing decision, pitch tracking, or octave error suppression. See the Appendix for the source code.

Only a few years after the above quote on nascent computer programming was a fully-digital algorithm developed by Noll at the Bell Telephone Laboratories, USA, so as to supply a vocoder with a pitch estimate for the purposes of low-bandwidth speech transmission. The algorithm is based on the idea that harmonic tone complexes have a repetitive spectrum, with harmonic peaks at regular

[5] JBOF: `https://github.com/bastibe/jbof`
[6] Transplant: `https://github.com/bastibe/transplant`

intervals on the frequency axis. In an ideal harmonic tone complex, the peaks are spaced one fundamental frequency apart from one another, and detecting this regular spacing equates to detecting the fundamental frequency.

To estimate this repetitive spacing, the logarithmic spectrum is subjected to a second Fourier Transform, now called the "Cepstrum", which exhibits a single peak at the peak-spacing distance quefrency[7] and corresponds to the fundamental frequency of the signal. The origin of this idea is summarized as follows:

> "In the fall of 1959, Bogert (of Bell Telephone Laboratories) noticed banding in spectrograms of seismic signals. He realized that this banding was caused by 'periodic' ripples in the spectra and that this was characteristic of the spectra of any signal consisting of itself plus an echo. The frequency spacing of these ripples equals the reciprocal of the difference in time arrivals of the two waves. Tukey (of both Princeton University and Bell Telephone Laboratories) suggested that this frequency difference might be obtained by first taking the logarithm of the spectrum, thereby making the ripples nearly cosinusoidal. A spectrum analysis of the log spectrum then could be performed to determine the 'frequency' of the ripple. In early 1960, Bogert programmed Tukey's suggestion on a computer and proceeded to analyze numerous earthquakes and explosions. Tukey, noticing similarities between time series analysis and log-spectrum series analysis, introduced a new set of paraphrased terms. The spectrum of the log spectrum was called the 'cepstrum,' and the frequency of the spectral ripples were referred to as 'quefrency.' Bogert, Tukey, and Healy published their ideas in an article with perhaps one of the weirdest titles ever encountered in the scientific literature: 'The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking.' In the article, they very clearly expressed a pessimistic view for achieving adequate classification of seismic events by cepstral techniques. In fact, no definitive indication of focal depth was found. Their article was issued as an internal Bell Laboratories memorandum before publication in Rosenblatt's book. Schroeder read the memorandum and realized that voiced speech spectra also have ripples, and hence cepstrum analysis might be suitable for vocal-pitch determination. In June 1962, Schroeder suggested cepstrum-pitch determination as an area worthy of further study. At that time, he and Atal had just completed a paper on methods for performing short-time spectrum analyses. Thus, the atmosphere was perfect for the concept of short-time cepstrum analysis that then developed." [105, p. 1]

One can feel the excitement of these early days of digital signal processing, when fast digital computers and linear algebra offered boundless new possibilities, and scores of old ideas were begging to be translated into the digital realm.

While the paper's description of the algorithm is thoroughly modern, this is still in the very early stages of computer programming, as "the program was still very lengthy and required about 0.8 h to compute the cepstra for about 2 sec of speech. Recently, an algorithm has been developed by Cooley and Tukey for performing fast numerical Fourier transformations. This algorithm has been incorporated into the cepstrum program and has resulted in a program about eight times faster than the previous one." [105, p. 8] That new algorithm is the Fast Fourier Transform, as it is known today. Truly, the impact of this invention on digital signal processing cannot be overstated, and it will form the basis of most later PDAs.

The PDA's voicing decision is based on the idea of picking a cepstral peak between 1-15 ms, with a linear weight of 1 at 1 ms and 5 at 15 ms. Afterwards it has an adaptive threshold for picking time-

[7] a cepstral frequency, one of the neologisms invented by Bogert, Healy, and Tukey for cepstral analysis (or "quefrency alanysis"), along with "rahmonics" and "liftering" [109].

continuous cepstral peaks, while excluding octave errors. The replication dataset re-implementation picks the weighted peaks, but includes neither a voicing decision nor suppression of octave errors, as their description lacks certain decision threshold values.

As processing time still represented a large challenge at the time of publication, no comparison or large evaluation was included in the paper, except that "additive white noise is not too degrading if it does not destroy the spectral ripples. Actually, a clearly defined cepstral peak has been obtained for speech signals with a 6-dB signal-to-noise ratio over the 40-msec analysis interval." [105, p. 15]

Later publications generally refer to this algorithm as the "Cepstrum algorithm" or even shorter as "CEP". So popular is this PDA that it was included in the `pitch` function in the signal processing toolbox of Matlab 2018a.

11.3.2 *AUTOC* [140] (1968)

Re-implemented in Python, without voicing decision, pitch tracking, or octave error suppression. See the Appendix for the source code.

Around the same time as the cepstrum pitch estimator detailed above, another common template for fundamental frequency estimation was published, which is today known as the "Autocorrelation algorithm", or "ACF" for short, by Sondi, also of Bell Telephone Laboratory, USA. The idea is that the period of a periodic signal is visible as a strong peak in the signal's autocorrelation function. However, this is technically only true for perfectly harmonic signals with equal-amplitude harmonics. Speech signals therefore need to be spectrally flattened in some way before autocorrelation to produce strong autocorrelation peaks and reliable fundamental frequency estimates.

The paper describes three methods of spectral flattening. After some deliberation, the simplest of these methods was found to be the most effective: The signal is center clipped, which removes all amplitudes up to a moving threshold, leaving only the amplitude tips. This effectively removes most of the formant information, but retains periodicity. The remaining signal is unintelligible, but perfectly suitable for autocorrelation.

Furthermore, "in at least one type of situation, [center clipping] works more reliably than [...] the more elaborate cepstrum pitch extractor. This is the case when a voiced segment of speech becomes almost sinusoidal. (This occurs, for example, if the speech signal is the sound /i/ spoken by a female and high-pass filtered with a cutoff at about 200 or 300 Hz. This is not a very unusual situation if the speech has traveled over an ordinary telephone circuit.) Since the success of [the cepstrum algorithm] depends upon the presence of a large number of harmonics, these types of pitch extractors are prone to error in such cases. The absence of a large number of harmonics clearly is not a serious problem for the center-clipping method" [140, p. 3]. A peak picking and pitch tracking stage then follows, explicitly adhering to the logic of the cepstrum estimator discussed previously, and again only partly implemented in the replication dataset for the same reason.

An informal evaluation was conducted with high-pass filtered and low-pass filtered speech with additional broadband noise. In the same vein as the cepstrum algorithm, the goal was to drive a speech vocoder, where "the resulting resynthesized speech was judged excellent by listeners in informal listening tests. None of the usual troubles of pitch doubling and loss of the trailing portions of voiced intervals was noticeable." [140, p. 5]

At the time, speech vocoders promised high-efficiency signal transmissions on narrow channels and promised more efficient usage of the already congested telephone network.

11.3.3 *SIFT* [93] (1972)

Re-implemented in Python, without voicing decision, pitch tracking, or performance optimizations. See the Appendix for the source code.

The spectral flattening idea discarded in the autocorrelation PDA was later reimagined as the Simple Inverse Filter Tracking (*SIFT*) algorithm by Markel of Speech Communications Research Laboratory, USA. An "inverse filter", or pre-whitening filter, removes the formant structure from a speech signal, whose pitch period can then be estimated as the delay of the first peak > 2 ms in its autocorrelation sequence. The pre-whitening filter is calculated with what would later be called linear predictive coding (LPC), minimizing the influence of linear components, and thereby removing vocal tract resonances.

A large part of the paper is relegated to performance optimizations, such as subsampling before the filter calculations to save computation time, and interpolating the position of the resulting maximum back to the original sampling rate. These optimizations are no longer necessary on today's hardware, and have been omitted in our re-implementation. Again, the voicing decision was omitted as well, which is additionally justified by the authors themselves as "it has been experimentally demonstrated that the difficult problem of detecting voicing during the transition from voiced to unvoiced interval is not completely resolved. [...] (It should be pointed out, however, that whenever the *SIFT* algorithm failed to extract correct voicing, cepstral analysis also failed.)" [93, p. 10]

The paper closes with a performance evaluation on a handful of short utterances against the cepstrum algorithm, mostly to prove that it could maintain similar accuracy while being "an order of magnitude [20×] faster than the cepstral analysis pitch extraction method" [93, p. 8].

11.3.4 *AMDF* [130] (1974)

Re-implemented in Python, without voicing decision or pitch tracking. See the Appendix for the source code.

The Average Magnitude Difference Function (*AMDF*) is a close relative to the autocorrelation function mentioned before. However, it uses the eponymous average magnitude difference instead of the multiplication of variously delayed signals. Thus, periodic signals do not produce maxima, but minima instead. This "anticorrelation" is partly done as a calculation speed improvement over autocorrelation, and partly as minima near zero were found to be easier to detect than maxima at arbitrary magnitudes.

An additional performance improvement truncated the auto-subtraction sequence, such that only short, approximately two-period segments, were compared against each full block of audio data. As the parameters of this procedure were not fully specified, it has been omitted in our re-implementation, which uses full-block comparisons. Again, the paper additionally included a complicated and multi-staged voicing decision, which has been omitted as well.

The evaluation section in the paper compares the *AMDF* against a simplified autocorrelation method without center clipping, and finds it to be similarly accurate, although errors in voicing decisions were found to be of great significance to overall accuracy. Also, "adding noise to the input signal caused pitch errors to be generated. These errors were speaker dependent but appeared to consist mostly of pitch doublings occurring at the onset or central portion of voiced sounds. [...] As the signal-to-noise ratio was varied from 30 dB to 10 dB, the number of errors increased, although not a substantial amount. A more substantial increase in error was found in going from the uncorrupted speech to a high [SNR] (30 dB) than in decreasing the [SNR] appreciably. Some evidence is available which shows that the *AMDF* remains suitable for pitch extraction down to a 0 dB [SNR]" [130, p. 7].

The rest of the paper explains the details of running the *AMDF* in real-time on commercially-available computers of the time. While the estimation accuracy and computation speed of the *AMDF* do not seem particularly impressive today, the general idea will be re-used in various PDAs in years after this publication, which makes it significant as a historical artifact.

11.3.5 The 1980s

There is a long gap in between the previous, "historical" PDAs and the subsequent ones. This is despite the eighties being a fascinating decade for fundamental frequency estimation, with personal computers finally becoming fast enough to do signal processing on affordable hardware. This processing power allowed for more complex algorithms to be implemented, such as statistical estimators [117, 126], PDAs based on psychoacoustic models [33, 153], harmonic combs [94, 58], and phase-based approaches [38, 20]. However, the resulting PDAs proved mere steppingstones towards more widely cited later iterations and hardly left usable implementations behind.

Additionally, the newfound computing power allowed for the first large-scale comparison studies [127, 108, 112, 60]. The first of these, "A Comparative Performance Study of Several Pitch Detection Algorithms", by Rabiner, Cheng, Rosenberg, and McGonegal, is particularly noteworthy as most later comparisons use the error measures and nomenclature defined in this work. These are the gross pitch error (GPE), fine pitch error (FPE), and voicing decision error (VDE), which are still the prevailing standard today. The paper compared eight short utterances with manually labeled pitch using the autocorrelation algorithm, the cepstrum method, *AMDF*, *SIFT*, and two additional PDAs that are not included here. These algorithms were classified as time-domain, frequency-domain, or mixed-domain, which has also endured as archetypes in recent publications, despite a proliferation of "mixed" approaches in recent years.

Follow-up work by the same authors [97] added a formal subjective evaluation of these PDAs, and Oh and Un [108] explicitly extended the evaluation to speech in noise. In general, these comparisons found the *AMDF* and autocorrelation methods to work most reliably both for clean recordings and in noise.

Another seminal work in this time is the book "Pitch Determination of Speech Signals" by Hess [59], which examines the history and major PDAs of the time in great detail, and–to the best of our knowledge–coins the term "PDA" for Pitch Determination Algorithm. It also popularized the idea of using laryngographs as a ground truth for fundamental frequency estimation, which proved tremendously influential in the years afterwards. The book itself neatly bookends the era of sample-by-sample, often analogue, feature detectors with added periodicity estimators, and serves as a gentle introduction to the new, digital, block-based approaches (here called "frames") that are the basis for almost all PDAs discussed here. The novelty of this idea is captured in the following quote:

> "Hence, we do not obtain the boundaries of individual periods, not even the lengths of individual periods, but rather an estimate of the average period length or fundamental frequency within a given frame. To detect periodicity at all, at least two periods must be situated within one frame; otherwise the information of periodicity is lost. Thus a minimum frame length of two maximum-duration pitch periods must be observed. On the other hand the [block-based analysis] principle implies that the signal is quasi stationary, i.e., the extracted parameters can be assumed constant within the frame. Thus the frame length must not be too large, otherwise the natural change of pitch in the signal may become significant and may spoil the intraframe estimate of this parameter. For speech signals, these two conditions are just compatible; they do not yet really conflict. Usual values for the frame length range between 20 and 50 ms, according to the actual value of F0 in the signal under consideration.

> [...] The discussion up to now already suggests that the majority of the [block-based] analysis PDAs go digitally. In fact, the analog [PDAs] of this category - not very numerous anyhow - have mostly been outperformed by their digital 'colleagues.'" [59, p. 357]

The last prediction proved prophetic, as later algorithms almost exclusively work digitally and use some kind of block-based approaches. In fact, the book's broad classification of PDAs into analogue real-time and block-based algorithms is largely meaningless today, as real-time algorithms all but ceased to be developed in the following decades.

With the introduction of data-driven PDAs in the 1980s and thoroughly digital workflows also came the introduction of *noise* as the main obstacle to fundamental frequency estimation. Before, PDAs were predominantly judged by their clean-speech accuracy for the purposes of vocoder-based synthesis, and perhaps when degraded by telephone transmission. But from the 1990s onward, PDAs needed to remain accurate in the presence of noise. This is perhaps no accident, as the "wireless revolution" of all-digital GSM mobile phones standardized and legitimized low-bitrate, vocoder-driven speech transmission at an acceptable speech quality in the early 1990s, thereby obviating the pressing need for further research into vocoders.

11.3.6 *PRAAT* [9] (1993)

Implementation by the original authors at `https://github.com/praat/praat` *in the C programming language.*

With great confidence, Boersma of the University of Amsterdam, Netherlands, introduced *PRAAT*'s signal model as

> "By definition, the best candidate for the acoustic pitch period of a sound can be found from the position of the maximum of the autocorrelation function of the sound, while the degree of periodicity (the harmonics-to-noise ratio) of the sound can be found from the relative height of this maximum." [9, p. 1]

Thus the signal model of the *PRAAT* algorithm is a time-domain definition of self-similar signal periods. The paper however argues that signal blocks are to be Gaussian-windowed before calculating the autocorrelation[8]. Further, autocorrelation must be normalized by the window autocorrelation to counteract the effects of windowing, which "seems to have gone by unnoticed in the literature" [9, p. 4]. Furthermore, the resulting normalized autocorrelation is to be upsampled with a sinc interpolator.

These considerations seem a bit odd, as there is no obvious need for windowing in autocorrelation-based period determination algorithms. In fact, without windowing, no bias correction would be necessary. However, the corrections might be justified as the algorithm uses an FFT-based calculation of autocorrelation, which out of necessity works on windowed signal blocks.

The authors claim that their improved autocorrelation method is still prone to octave errors, even for ideal sinusoids or pulse trains. This is corrected in a post processing stage, which seeks to minimize octave errors by viterbi-searching a pitch track that penalizes large frequency jumps. A formal evaluation shows the PDA's effectiveness for sinusoids and pulse trains in low-pass filtered white noise up to 0 dB SNR, but no comparison or evaluation with speech signals is given beyond "it works equally well for low pitches (the author's creaky voice at 16 Hz, alveolar trill at 23.5 Hz, and bilabial trill at 26.0 Hz), middle pitches (female speaker at 200 Hz), and high pitches (soprano at 1200 Hz, a two-year-old child yelling /i/ at 1800 Hz). The only 'new' tricks are two mathematically

[8] It is Hann-windowed at first, but an appendix adds that Gaussian windows are even better. Which makes sense as they state that side lobes in the window spectrum are problematic for *PRAAT*.

justified tactics: the division by the autocorrelation of the window [...], and the 'sin x / x' interpolation in the lag domain" [9, p. 14].

More important than the algorithm itself, however, is the release of the *Praat [software] for doing phonetics by computer* [10] in 1996 in collaboration with David Weenink. This program includes the PDA and a powerful graphical user interface for displaying and analyzing speech signals. It is likely the availability and continuous development of this software that has garnered *PRAAT* such a lasting support in the speech analysis community, much more than the individual algorithmic components.

The program is still maintained today, available for multiple operating systems, and includes a scripting system that allows for unattended use without the graphical interface. The replication dataset uses *PRAAT* version 6.0.24, retrieved in January 2017.

11.3.7 *RAPT* [150] (1995)

Implementation in the VOICEBOX framework by Brookes [14] at `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`*, written in Matlab, based on a C implementation by the original authors at* `http://www.speech.kth.se/wavesurfer/links.html`*.*

Just like the previous PDA, the *RAPT* PDA by Talkin of the Entropic Research Laboratory, USA, is a periodicity detector:

> "For the purposes of this chapter, [fundamental frequency] is defined as the inverse of the smallest *true* period in the interval being analyzed. This definition provides for the short-time variation in F0 that is observable in human speech. As will be seen later, determination of "true" is the crux of the matter!" [150, p. 2]

This statement is again interpreted as a variant of the autocorrelation algorithm, this time without windowing, but zero-meaned, and normalized by a signal energy estimate. Candidate pitch estimates are determined with a moving threshold, and the precise frequency of the maximum is determined from a parabolic interpolation around the autocorrelation maximum.

As predicted in the introductory quote, this conventional extractor is followed by a rather complex post processing procedure that forms the true innovation of *RAPT*: it includes viterbi-searching for an optimal pitch track with various weighting functions for suppressing octave errors, a low-frequency bias, and a joined voicing decision based on spectral differences, energy contours, and LPC analysis. The precise procedure is too complex to reproduce from the publication alone, but C code is provided by the original authors on their website.

As an evaluation, the PDA "has been used with satisfactory results on speech recordings varying in quality from noisy telephone to quiet laboratory conditions" [150, p. 20].

The popularity of this PDA in later publications and comparison studies likely stems from a re-implementation included in the widely cited VOICEBOX toolbox for Matlab by Brookes in 2006 [14], which was used in the replication dataset as well. This is visible in Figure 11.2 on page 126 as a sharp increase in mentions of *RAPT* around 2006. According to its documentation, the VOICEBOX implementation of *RAPT* is a straight translation of the original author's C source code with only minor modifications.

The *RAPT* algorithm is sometimes referred to as "Get_F0" due to the function name in the original source code.

11.3.8 *YIN* [24] (2002)

Implementation by the original authors at `http://audition.ens.fr/adc/`*, written in Matlab.*

Like the previous two PDAs, the *YIN* PDA by de Cheveigné and Kawahara of Ircam-CNRS, France, is a periodicity estimator, although this time defined rather rigorously:

> "The fundamental frequency [...] of a periodic signal is the inverse of its period, which may be defined as the smallest positive member of the infinite set of time shifts that leave the signal invariant. This definition applies strictly only to a perfectly periodic signal, an uninteresting object (supposing one exists) because it cannot be switched on or off or modulated in any way without losing its perfect periodicity. Interesting signals such as speech or music depart from periodicity in several ways, and the art of fundamental frequency estimation is to deal with them in a useful and consistent way." [24, p. 1]

Instead of relying on the autocorrelation sequence for matching signal periods, they use a square difference function, similar to the *AMDF* PDA, as "despite its appeal and many efforts to improve its performance, the autocorrelation method (and other methods for that matter) makes too many errors for many applications" [24, p. 3]. They explain this as "the [autocorrelation] is quite sensitive to amplitude changes. [...] An increase in signal amplitude with time causes [autocorrelation] peak amplitudes to grow with lag rather than remain constant [...]. This encourages the algorithm to choose a higher-order peak and make a 'too low' error (an amplitude decrease has the opposite effect). The difference function is immune to this particular problem, as amplitude changes cause period-to-period dissimilarity to increase with lag in all cases" [24, p. 3]. The authors additionally point out how the square difference function can be calculated as a sum of autocorrelation terms.

To reduce the influence of low-lag minima, as caused by formants and the zero at origin, the square difference function is further divided by the signal average up to the lag value. The resulting "cumulative mean normalized difference function" "starts at 1 rather than 0, tends to remain large at low lags, and drops below 1 only where [the function] falls below average" [24, p. 4]. Additionally, octave errors are reduced by picking the lowest frequency of multiple candidate minima, and final frequency estimates are improved by parabolic interpolation. A post processing step smooths out estimates slightly to prevent intermittent frequency fluctuations.

At this point, it would arguably be constructive to compare the last three PDAs with regards to their emphasis on estimator accuracy versus post processing complexity. The *PRAAT* and *RAPT* algorithm, and all four of the preceding "historical" PDAs placed significant effort into their post processing as a workaround for shortcomings of the main estimator, whereas *YIN* focuses more so on the estimator, and with less need for post processing. This was done in part to explicitly avoid complexity, "as including [complex post processing] complicates evaluation and credit assignment" [24, p. 11]. With the same reasoning, no voicing decision was included.

In a surprisingly large evaluation section of PDAs across multiple clean-speech corpora, "*YIN* performs best of all methods over all databases. Averaged over databases, error rates are smaller by a factor of about 3 with respect to the best competing method" [24, p. 5]. This includes evaluations of the aforementioned *PRAAT*, Autocorrelation, and the Cepstrum method, although their post processing and voicing decision was disabled and their frequency range was expanded. A laryngograph PDA was used as ground truth. This evaluation was done on multiple databases, including the *KEELE* and *FDA* databases used in the replication dataset.

The Matlab implementation by the original authors relies heavily on compiled C code, which needs to be recompiled for recent versions of Matlab. As the implementation was found to have problems with high sampling rates, signals in the replication dataset were resampled to 16 kHz before passing them to *YIN*.

11.3.9 *SHR* [149] (2002)

Implementation by the original authors at https://mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm, written in Matlab.

The subharmonics-to-harmonics ratio (*SHR*) by Sun of Northwestern University, USA, is a different kind of PDA from the previous ones, in that does not work in the lag domain, and specifically targets a sore spot in fundamental frequency estimation: speech signals with alternate cycles, which appear in creaky voices or some pathological voices and produce prominent subharmonic partials.

> "The alternate cycles make pitch determination difficult. The solutions [to estimating pitch in the presence of subharmonics] of most current algorithms either rely on fine-tuning some threshold parameters based on particular databases or post processing techniques, such as linear/nonlinear smoothing, dynamic programming, etc. In the present paper, an alternative approach is explored." [149, p. 1]

The algorithm posits that likely fundamental frequencies have a large ratio between "harmonic" bins, spectral bins at integer multiples of a candidate pitch, and "subharmonic" bins, at integer multiples of half the candidate pitch.

This is a significant development in a lineage of harmonic-comb/harmonic-sieve PDAs working in the magnitude spectrum [94, 58], last seen in this list with the *CEP* PDA from 1967. These PDAs correlate synthetic comb spectra at various candidate pitches with a short-time spectrum to find likely fundamental frequencies, not entirely unlike the cepstrum method discussed earlier, but without the logarithm and with arbitrary comb shapes. The innovation of *SHR* is the inclusion of a subharmonic comb, and thereby a measure not just of harmonic energy, but simultaneously a rejection of non-harmonic energy.

As a slight contradiction to the above quote, a complex post processing stage selects the most likely pitch of two candidates and makes a voicing decision based on a noise floor estimate, the signal energy, signal correlation, and zero-crossing rate.

The publication evaluates two versions of *SHR* with the *FDA* and *KEELE* database, and compares them against *RAPT*, *PRAAT*, and a variant of the autocorrelation method. Error measures are the usual voicing error, gross pitch error, and fine pitch error. In defense of their premise, they conclude "that [the most difficult speakers] indeed contain more 'irregular' speech cycles and appears to have low and rough voices, whereas [the easier speaker's] speech seems to be more 'regular'" [149, p. 4]. Contrary to its stated purpose, however, *SHR*'s accuracy does not perform relatively better than other PDAs for these voices.

While the originally published website and *SHR*'s source code are no longer available, the original author has published a Matlab implementation of the algorithm on the Mathworks File Exchange. This version appears to include additional post processing procedures not documented in the original publication and is used in the replication dataset.

11.3.10 *YAAPT* [69, 172, 173] (2002-2008)

Implementation by the original authors at http://www.ws.binghamton.edu/zahorian/yaapt.htm, written in Matlab.

YAAPT, or Yet Another Algorithm for Pitch Tracking, developed by Kasi and Zahorian at Old Dominion University, USA, and others[9] has seen multiple publications over time, with seemingly ever-growing complexity. The latest iteration has multiple interlocking pitch trackers, with features in

[9] possibly including Princy Dikshit and Hongbing Hu

both the time and frequency domain. In the words of the authors, "Although methods similar to all the individual components of *YAAPT* have been used to some extent in previous F0 trackers, these components have been implemented and integrated in a unique fashion in the current algorithm." [173, p. 12]

The resulting algorithm is indeed uniquely elaborate, and estimates the fundamental frequency in no less than three separate signal representations: as a *RAPT*-like normalized auto-correlation of a band-passed signal, the same of a squared band-passed signal, and a soft spectral comb of the squared band-passed signal. Each has their own peak-picking method with their own viterbi-search, either on the full signal, or a voiced-only concatenated signal and multiple voicing determination and octave-error-suppression stages using additional full-signal context information. It is safe to say that this is by far the most complex PDA discussed thus far, and well beyond the scope of this section to describe fully.

Unsurprisingly, "*YAAPT* is quite demanding [computationally] due to the variety of signal processing approaches used and then combined in the complete algorithm." [173, p. 12] In fact, the authors themselves state that "it could be questioned whether or not both the temporal and spectral tracks are needed and the extent to which each of these sources of information contributes to the accuracy of the F0 tracking" [173, p. 8]. However, their evaluation of the PDA's various components finds that "the combination of the temporal and spectral tracks results in better performance than using any individual component, illustrating the benefits of using both temporal and spectral information" [173, p. 9].

A wider comparison with other PDAs, using the *KEELE* and *FDA* speech corpus in white noise and babble noise, shows *YAAPT* outperforming *PRAAT*, *RAPT*, and *YIN* in terms of gross pitch errors. However, these results are also compared to previously published comparison studies, and "although test conditions and parameter settings are intended to be identical, clearly, there are differences since the results obtained with *YIN* in this study and those obtained with *YIN* in the previous studies are significantly different" [173, p. 11].

This vividly illustrates the need for the replication dataset developed in the present study.

Source code for *YAAPT* in Matlab is available on Zahorian's website. The replication dataset uses version 4.0 from 2016.

11.3.11 *SWIPE* [18] (2007)

Implementation by the original authors used to be available at `http://www.cise.ufl.edu/~acamacho/english/curriculum.html`*, written in Matlab. Still accessible on the Internet Archive.*

The *SWIPE* PDA by Camacho at University of Florida, USA, is another harmonic comb PDA, but this time not with discrete and sharp comb teeth but a cosine shape instead that smoothly connects positive harmonic peaks with negative subharmonic valleys. Both the harmonics and subharmonics are weighted to match the sensitivities of the human auditory system, and the correlation is carried out in an ERB-warped, Hann-windowed spectrum.

Interestingly, spectral amplitudes are compressed with a square root instead of a logarithm, as "the use of the logarithm of the spectrum in an integral transform is inconvenient because there may be regions of the spectrum with no energy, which would prevent the evaluation of the integral, since the logarithm of zero is minus infinity. But even if there is some small energy in those regions, the large absolute value of the logarithm could make the effect of these low energy regions on the integral larger than the effect of the regions with the most energy, which is certainly inconvenient." [18, p. 51]

The primary problem of harmonic combs, according to the publication, are octave errors. Like any harmonic comb, *SWIPE* still matches a pitch similarly well as double or half the pitch. To reduce this tendency, the *SWIPE*-prime PDA uses a cleverly modified comb with all non-prime comb teeth

removed. Thus, an accidental pitch halving correlates fewer comb teeth than the true pitch.

SWIPE and *SWIPE*-prime were evaluated against *PRAAT*, *CEP*, *RAPT*, *SHR*, *YIN*, and seven other PDAs on three speech databases, including *FDA* and *KEELE*. This evaluation overlaps significantly with *SHR*'s evaluations. Results, however, differ strongly. In particular, *SWIPE*'s comparison consistently favors male voices for all PDAs, where no such bias exists in the other comparison.

Until very recently, source code for the *SWIPE* and *SWIPE*-prime PDA was available on the author's university website, but this is no longer online. However, a printed copy is still appended to the publication and the original website is still available on the Internet Archive. In other publications and the replication dataset, the name "*SWIPE*" typically refers to the *SWIPE*-prime PDA, and the distinction between the two is ignored.

11.3.12 *STRAIGHT* [73] (2008)

Implementation by the original authors at `https://github.com/HidekiKawahara/legacy_straight`*, written in Matlab.*

STRAIGHT is "a speech analysis, modification, and synthesis system" developed by Kawahara at Wakayama University, Japan. As the original publication on the subject explains [71],

> "The central idea of the proposed method considers the periodic excitation of voiced speech to be a sampling operation of a surface $S(\omega, t)$ in a three-dimensional space defined by the axes of time, frequency, and amplitude; these axes represent the global source characteristics and the shapes and movements of articulation organs. In this interpretation, a periodic signal $s(t)$ with a fundamental period τ_0, is thought to provide information about the surface for every τ_0 in the time domain and every $f_0 = {}^1/_{\tau_0}$ in the frequency domain. In other words, voiced sounds are assumed to provide partial information about the surface. The goal of spectral analysis that enables flexible manipulation is to recover the surface $S(\omega, t)$ using this partial information." [71, p. 3]

The publication then extends this fascinating view of speech to a signal model that is neither strictly harmonic nor strictly periodic. Accordingly, fundamental frequency estimation cannot rely on periodicity or harmonicity, and instead measures the instantaneous frequency of the fundamental component. This is one of the most rigorous definition of "fundamental frequency" possible. However, the history of *STRAIGHT* is tumultuous, with quite a number of publications over the years that contain multiple, varying interpretations. For example, a later iteration of the PDA used autocorrelation on various spectrally flattened frequency bands instead [70].

The TANDEM-*STRAIGHT* approach from 2008, which is included in the replication dataset, instead aims to produce an altered spectrum with reduced temporal and spectral modulations, which allows traditional harmonicity-detection algorithms to work on originally non-harmonic signals. This is implemented using a combination of two spectra of the same signal block, with different time windows that compensate for each other's spectral minima. This "temporally stable", "interference-free" *STRAIGHT* spectrum serves as a smooth approximation of the formant envelope without any temporal or harmonic structure. An additional TANDEM spectrum is derived to have a similar envelope, but with sinusoidal modulations in the spectrum for harmonic tone complexes. Dividing these spectra produces a spectrally flat "fluctuation spectrum" that only contains the harmonic structure.

The fundamental frequency is extracted from this fluctuation spectrum with a weighted Fourier transform of the spectrum and a peak picking algorithm, somewhat similar to the cepstral PDA discussed earlier. According to the authors, this PDA operates "pitch-synchronously or pitch-adaptively with temporal resolution comparable to that of the fundamental period. Both TANDEM and

STRAIGHT spectra simultaneously satisfy a finer temporal resolution requirement and essentially yield pitch synchronous analysis without the need for precision in window positioning" [73, p. 2].

Such a high temporal resolution would serve to bridge the gap between the older, pitch-synchronous analogue PDAs, and the new, noise-robust short-time PDAs, and might be desirable for musical applications.

The popularity of this method must be influenced by the large number of publications it has garnered over the years [71, 70, 73, 76, 72, 74, 75]. Although evaluations of some versions of *STRAIGHT* were published, none seem to be using the newest TANDEM-*STRAIGHT* PDA discussed above. The version of the software used in the replication dataset was originally obtained by email request, but has since been moved to a public location on Github. It must be assumed that later mentions of the "*STRAIGHT*" or "TANDEM" PDA typically refer to this implementation instead of the earlier, source-code less variants.

11.3.13 *DIO* [100] (2009)

Implementation by the original authors as part of the WORLD framework used to be at `http://ml.cs.yamanashi.ac.jp/world/`*, written in Matlab, still available in the Internet Archive. A newer version is available at* `http://www.kisc.meiji.ac.jp/~mmorise/world/english/`*, and a C and Python version has since been published at* `https://github.com/mmorise/World`*.*

The "Distributed Inline-filter Operation", or *DIO*, of Morise at Kwansei Gakuin University, Japan, was specifically designed for real-time applications for singing voices. As such, it explicitly optimizes for clean recordings.

The estimator pre-processes the signal with a set of very steep low-pass filters at various cutoff frequencies of human singing voices. Each of these candidates is then weighed by a "fundamentalness" score, which is the variance between four period detectors, one of signal peaks, one of valleys, one of rising zero crossings, and one of falling zero crossings. The longest candidate period with a zero fundamentalness is taken as the fundamental frequency.

This is truly a modern version of the analogue period detectors of yore, although extended by parallel evaluation of a low-pass filterbank. A comparison with literature results from *YIN*, *STRAIGHT*, *AUTOC*, and *CEP* on the clean-speech *FDA* corpus indeed shows somewhat middling accuracy, which is explained as "[*STRAIGHT*] is the best of all methods, but its processing time was much longer than [*DIO*], which performs better than the conventional methods without post-processing" [100, p. 3].

However, *DIO* is part of the WORLD speech synthesis framework, and has garnered a number of citations due to this inclusion. The source code in the replication dataset is the original version in Matlab, which has since been superseded by a newer implementation to C as part of the WORLD framework. It is otherwise referred to as "*DIO*" or "WORLD" in later publications.

11.3.14 *SAFE* [22] (2010)

Implementation by the original authors at `http://www.seas.ucla.edu/spapl/weichu/safe/` *in the C programming language.*

The *SAFE* algorithm by Chu and Alwan of the University of California, USA, or "Statistical Algorithm for F0 Estimation for both clean and noisy speech" marks a phase change in PDA development. From this point onwards, most PDAs consider the estimation of clean speech fundamental frequencies only of academic interest, and identify the real challenge in robustness to noise[10].

[10] *SAFE* is not the first PDA in history to focus on accuracy in noise, merely the first one in this list. But its publication date at the beginning of a new decade is a convenient inflection point in the narrative, and a good proxy for the cultural

The algorithm itself is our first specimen of a new breed of statistically motivated approaches that define harmonics not in terms of spectral amplitudes, but as maxima in a probability space. In this case, the *SAFE* algorithm defines harmonics as peaks in a probabilistic signal-to-noise ratio that is derived from a number of simplifications and assumptions to bend the endless varieties of speech and noise spectra into a mathematically tractable shape.

In essence, an estimate of the background noise is obtained from the first and last blocks in the noisy recording, assuming that those contain no speech and that the noise is quasi-stationary throughout the file. It then calculates an SNR measure between the local spectrum and the noise estimate. Peaks in the difference between a strongly smoothed and a weakly smoothed version of this SNR are selected as harmonics. Finally, a Bayesian estimator is trained on these harmonics and various parameters, and a maximum likelihood estimate for a set of candidate fundamental frequencies is calculated.

A post processing stage viterbi-searches an optimal pitch track and suppresses octave errors. However, it explicitly does not deal with unvoiced frames, and ignores them with the help of a voicing decision ground truth.

The *SAFE* algorithm is then trained on a part of the *KEELE* dataset, and compared to *RAPT*, *PRAAT*, TEMPO, and *YIN* on the *FDA* and *KEELE* dataset in varying levels of white and babble noise from the *NOISEX* corpus. This evaluation concluded that "the *SAFE* algorithm has the lowest GPE when the SNR is at or below 5 dB under white noise, and at or below 10 dB under babble noise. [...] Although there is a mismatch between the *KEELE* [training set] and *FDA* [test set] corpora, *SAFE* still has the lowest GPE on *FDA* under low SNR conditions as it does for the *KEELE* corpus" [22, p. 4].

The *SAFE* algorithm is provided as C source code for a command line application, and was compileable without trouble in 2018. Since it expects audio data with a sampling rate of 16 kHz, data was resampled if necessary before passing it to *SAFE*. The replication dataset uses a version of *SAFE* downloaded in late 2018.

11.3.15 *SRH* [31] (2011)

Part of the COVAREP [27] framework at `https://github.com/covarep/covarep`*, written in Matlab, in collaboration with the original authors.*

The publication for this PDA by Drugman and Alwan of the University of Mons, Belgium starts with "this paper focuses on the problem of pitch tracking in noisy condition [...] using harmonic information in the residual signal" [31, p. 1], and again highlights the changed priorities of the new decade: The performance in noise is the forefront in fundamental frequency estimation.

Like the *SIFT* PDA mentioned earlier, the Summation of Residual Harmonics, or *SRH*, is based on an auto-regressive LPC filter for pre-whitening and the removal of vocal tract effects. Instead of autocorrelation, however, an *SHR*-like harmonic comb with negative comb teeth at subharmonics is used on the residual spectrum to estimate candidate fundamental frequencies, and to make a voicing decision. Octave errors are minimized by limiting fundamental frequency candidates to within plus-or-minus one octave around the mean pitch of the entire recording, figuring that it "can be indeed assumed that a normal speaker will not exceed these limits" [31, p. 2].

The second contribution of this publication is its thorough and exemplary evaluation section that measures not only gross pitch errors, but also voicing decision errors, fine pitch errors, and F0 frame errors (union of GPE and VDE). These error measures are calculated for *RAPT*, *SHR*, *STRAIGHT*, *PRAAT*, and *YIN* with the *FDA* and *KEELE* databases, corrupted with noise signals from the *NOISEX* database.

---

shift.

The results are, "In clean speech, it is seen that [*SRH*] give a performance comparable to other techniques. [...] On the opposite, the advantage of *SRH* is clearly noticed for adverse conditions. In 9 out of the 10 noisy cases (5 noise types and 2 genders), *SRH* provides better results than existing methods" [31, p. 3]. This is hardly surprising, as none of the other methods was explicitly designed for noisy conditions, but nevertheless marks the change in priorities in the state of the art.

As a corollary to the focus on accuracy in noise, voicing activity determination in noise is no longer trivial. The comparisons therefore show a strong implicit (GPE) and explicit (FFE, VDE) emphasis on voicing decision errors, and special reference is made to the quality of *SRH* as a voicing detector. This will be another recurring theme in the succeeding years.

Source code for the *SRH* algorithm is made available as part of the COVAREP repository for collaborative voice analysis for speech technologies. The replication dataset uses the Matlab implementation in v1.4.1 of the COVAREP from October 2015. Another version was published as part of the `pitch` function in the signal processing toolbox of Matlab 2018a. As *SRH* only works for sampling rates of 16 kHz, all recordings in the replication dataset were resampled accordingly before passing them to *SRH*.

11.3.16 *SACC* [86] (2012)

Implementation by the original authors as part of the LabROSA project at `http://labrosa.ee.columbia.edu/projects/SAcC/`*, written in Matlab.*

The Subband Autocorrelation Classification algorithm, *SACC*, by Lee and Ellis of Columbia University, USA, is the first instance in this list of a new kind of PDAs that explicitly relies on machine learning techniques for fundamental frequency estimation, and implicitly eschews the need for interpretable results in exchange for ease of implementation. However, as will become clear later on, this distinction is more fluid than a firm decision, with various levels of machine learning taking over more or less of the algorithmic design.

As far as the *SACC* algorithm is concerned, a multi-layer perceptron classifier is trained on the principal components of the normalized autocorrelations of subbands from an 48-channel auditory filter bank. Thus, elements of traditional PDAs, such as the filter bank and the autocorrelation, are still present. One could argue that it is merely the post processing of these autocorrelations that is relegated to the machine learning system. Be that as it may, the autocorrelations are reduced to ten bins per channel using principal component analysis, and fed to a three-layer perceptron with an 800-node hidden layer for joint fundamental frequency estimation and voicing determination. A viterbi search then finds the optimal fundamental frequency track.

As is now expected in the 2010s, the optimization target of the PDA is noisy speech, explained as: "when acoustic degradations such as frequency band limitation and additive noise are introduced, the problem becomes still more challenging. This work is motivated by the problem of identifying and recognizing speech signals in low-quality radio transmissions, which we simulate, based on measurements of a real narrow-FM radio channel" [86, p. 1].

The machine learning stage, however, requires considerably more data for training than any of the previous PDAs. A training dataset was constructed from multiply resampled versions of each recording, filtered variously, and mixed with noise recordings at multiple SNRs. According to the publication, the same dataset was used for training and for evaluation.

The resulting PDA was evaluated with the *KEELE* and *FDA* corpora in various levels of pink noise, and compared against *YIN*, *SWIPE*, and others. Interestingly, they evaluated voicing errors separately from estimation errors, and found that voicing false positives[11] dominated at high SNRs, while false

[11] estimated voiced, although truly unvoiced

negatives[12] were more prevalent at low SNR. With the growing importance of noisy conditions and voicing determination, these kinds of statistical evaluations increasingly become critical for a thorough examination of PDA accuracy.

Source code for *SACC* in heavily-optimized Matlab with various functions implemented in C for speed is provided by the original authors on their website. The software includes various pre-trained models, and the replication dataset uses the default model trained on the RATS [159] dataset for a sampling rate of 16 kHz. Accordingly, all samples in the replication dataset were resampled to 16 kHz before applying *SACC*.

11.3.17 *BANA* [56] (2012)

Implementation by the original authors at `http://www2.ece.rochester.edu/projects/wcng/code.html`*, written in Matlab.*

The *BANA* PDA, named after its two main authors, Ba and Na (et al.) of University of Rochester, USA, is a modern re-combination of two older approaches, the cepstrum PDA from 1967, and the period histogram from Schroeder in 1968 [135].

First, the signal is band-pass filtered between 50 Hz to 3000 Hz, and the five lowest and most prominent spectral peaks in this frequency range are selected as harmonics.

Much like the period histogram PDA, a histogram of candidate fundamental frequencies is created from peak frequencies at conspicuous multiples of each other. For example, if two neighboring peak frequencies are a multiple of two of each other, this might identify the first one as the fundamental; a multiple of 1.5 would occur between the second and third harmonic and indicate a fundamental of half the first peak frequency. Additional candidate fundamental frequencies are added from a cepstral analysis, and from the frequency of the first peak. These candidates are collected in a histogram, and a confidence score is calculated from the number of candidates in close proximity to one another.

A pitch track is generated by viterbi-searching the histograms of each block and penalizing both frequency jumps and confidence jumps. Although the PDA claims to be optimized for high levels of noise, no particular algorithmic steps are taken to consider noise.

The evaluation section of the paper compares *BANA* to *PRAAT*, *YIN*, *CEP*, the harmonic product spectrum[13] against a hand-labeled ground truth of an unspecified number of speech recordings in various SNRs of *NOISEX* noise.

It is surprising that this combination of very old and outdated fundamental frequency algorithms works as well as it is shown to do. Perhaps this is another instance of a simple estimator, vastly improved by powerful post processing that elevates it beyond the accuracy of its ancestry.

11.3.18 *MBSC* [151] (2013)

Implementation by the original authors at `http://www.seas.ucla.edu/spapl/shareware.html`*, written in Matlab.*

In contrast to most of the previous PDAs, which mostly adhere to one or two guiding principles, the *MBSC* algorithm, or multi-band summary correlogram by Tan and Alwan of the University of California, USA, combines a large number of techniques into easily the most complex PDA in this list.

The time signal is split into four overlapping 1-kHz filter bands between 0 and 3.5 kHz to capture at least two harmonics per filter, as "a signal contain[ing] more than 1 harmonic of the target voiced

[12] estimated unvoiced, although truly voiced

[13] possibly confused with the period histogram, which was published in the same referenced paper and would be more appropriate

speech, its envelope would typically oscillate at an amplitude modulation frequency corresponding to the inter-harmonic separation" [151, p. 3].

From these filter bands a Hilbert envelope is calculated and mean-normalized. An additional non-envelope band is added for the lowpass band. Each band's spectrum is comb-filtered with a set of raised-cosine spectral combs and matching subharmonic combs at various candidate pitches. Additionally, a normalized autocorrelation function is calculated for each harmonic comb band.

In several stages, candidate pitches are removed; if their harmonic-to-subharmonic ratio is not a local maximum, if their harmonic-to-subharmonic ratio is < 1, if the maximum is not a power-of-two of other maxima, if the corresponding autocorrelation maximum does not agree with other candidates. Finally, the two best candidates are selected per filter band.

The autocorrelation functions of the harmonically filtered envelopes of the remaining candidates are combined across candidates and across the four filter bands, weighted by their harmonic-to-subharmonic ratio and cross-band agreement. This final multi-band summary correlogram forms an "improved" autocorrelation sequence that yields pitch estimates from weighing, peak-picking, parabolic interpolation, and viterbi-searching, like many other autocorrelation-based PDAs.

> "Together, the proposed signal processing schemes (subband multi-channel comb-filtering, [harmonic-to-subharmonic based] channel-selection-and-weighting, stream-reliability-weighting) help to enhance the maximum *MBSC* peak at the most likely pitch period, which in turn improves the accuracy of pitch estimation, as well as [voicing] detection. The variability of the maximum *MBSC* peak amplitude with SNRs is reduced, such that robust [voicing] detection is achieved by simply applying a constant threshold on this single feature, followed by median filtering – without requiring additional features" [151, p. 8]

This is an extreme example of shifting the balance between estimator complexity and post processing complexity far towards the former. It must be said that this approach does not lend itself to theoretical reasoning about the algorithm's components, or their individual merits, although this might be a moot point with the recent shift towards entirely opaque machine learning models.

A thorough evaluation section compares the PDA accuracy and its voicing decision on the *FDA* corpus with *RAPT*, *YIN*, *SHR*, *SWIPE*, and a PDA called WWB [169][14] and noise from the *NOISEX* corpus. The *KEELE* corpus was used to train *MBSC* and the other PDAs' voicing thresholds, and some PDAs were altered to ensure all-voiced output for one evaluation and a wider pitch range. All signals were resampled to 8 kHz, and optionally filtered to a telephone-like bandwidth.

Although this comparison makes a few assumptions that might not agree with some of the PDAs' signal models, it is very thorough and includes many signal conditions. *MBSC* is shown to perform particularly well in moderate to heavy noise. Interestingly, no comparison was conducted with the authors' own *SAFE* PDA, published three years prior.

11.3.19 *PEFAC* [45] (2014)

Implementation by the original authors as part of the VOICEBOX framework at `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`*, written in Matlab.*

PEFAC is the "Pitch Estimation Filter with Amplitude Compression", "a pitch estimation algorithm robust to high levels of noise" by Gonzalez and Brookes of the Imperial College, UK. Straight from the title, this sets the stage for another modern, noise-focused algorithm.

[14] WWB had to be heavily modified to work in a comparison setting, which is why it wasn't included in the replication dataset despite being available and popular.

The introduction has a broad summary of various algorithmic approaches. On the topic of the popular class of autocorrelation-based PDAs, it says that "Autocorrelation-based pitch detectors perform well in moderate noise levels since the [autocorrelation function] of an aperiodic noise source typically falls off rapidly with lag. At negative SNRs, however, a voiced speech signal whose energy is dominated by low-order harmonics will not generate a distinct peak in the [autocorrelation function] and [...] reliable pitch estimation becomes impossible" [45, p. 2].

Instead, *PEFAC* is loosely a harmonic comb PDA, which is claimed to be advantageous in noise "since most of the energy of a voiced speech signal is normally concentrated into a small number of harmonic peaks, these remain detectable even at poor SNRs" [45, p. 2]. The comb in question is a mean-normalized reciprocal of a raised-cosine shape with ten comb teeth, i.e. a soft comb with sharp positive teeth and shallow negative gaps. It is correlated with a shaped spectrum in the log-frequency domain, thus attenuating the influence of higher harmonics.

A spectral shaping stage forces the spectrum to conform to a long-term average speech spectrum shape, which reduces low-frequency and narrow-band noises that would otherwise warp the spectral shape. "The motivation for this is that if the shape of the average power spectrum of clean speech is known a priori, deviations from this shape indicate either a non-uniform channel response or the presence of noise" [45, p. 3].

The fundamental frequency is estimated by viterbi-searching up to three candidates per frame, while maximizing spectral peak amplitude, maintaining a stable fundamental frequency, and staying close to the median pitch.

An independent voicing decision is made on the log-mean power of the shaped spectrum of each frame, and the ratio of harmonic-peak power to the frame's total power using Gaussian mixture models.

The Gaussian mixture models and viterbi parameter weights were trained on a *NOISEX* and *TIMIT* training set, with a consensus truth from *PRAAT*, *YIN*, *RAPT*, and manually corrected where needed. An evaluation was carried out with two databases, including the *TIMIT* test set, in various noises from *NOISEX*, where *PEFAC* is compared to *YIN*, *RAPT*, and another PDA, using an uncommonly-tight $\pm 5\%$ GPE limit. Particularly in white noise and at very low SNRs, *PEFAC* is shown to work very well.

The popularity of *PEFAC* was no doubt helped by its inclusion in the popular VOICEBOX framework, and, as of 2018, its availability as part of the `pitch` function in the signal processing toolbox of Matlab.

11.3.20 *DNN/RNN* [50] (2014)

Implementation by the original authors at `http://web.cse.ohio-state.edu/pnl/software.html`*, written in Matlab, included both in its* DNN *and* RNN *versions.*

As a direct extension of *PEFAC*, the *DNN/RNN* algorithm by Han and Wang of Ohio State University, USA, uses the very same spectral shaping and logarithmic-frequency soft comb, but replaces the peak picking post processing with a deep (DNN) or recurrent (RNN) neural network. This is justified as "Although the [*PEFAC*] feature vector is designed to deal with noisy speech, [its] rule-based pitch candidate selection may lose useful information because it simply ignores non-peak spectral information. In our study, we treat [the *PEFAC* feature vector] as the extracted feature and employ supervised learning to estimate pitch probability, i.e. to learn the mapping from the features to the pitch frequencies. We expect supervised learning to yield better results" [50, p. 2].

The DNN is constructed with an input layer of three blocks, three hidden layers with 1600 sigmoids each, and an output layer with 68 softmax units for frequencies between 60 and 400 Hz and one unvoiced state. The RNN has two hidden layers with 256 sigmoids each, only the second of which is

recurrent. As with other machine learning publications, there is little justification given for the choice of these parameters, and some training parameters even remain entirely unspecified, such as learning rates and initialization schemes.

The networks are trained with the *TIMIT* corpus with a *PRAAT*-generated ground truth and *NOISEX* noises at SNRs around 0 dB. The resulting "posterior" is transformed into a likelihood by a separately learned prior and an empirical correction factor, and viterbi-searched and smoothed for a fundamental frequency estimate. The combination of a viterbi-search and an RNN is curious, however, as "the output of the RNN is the posterior probability given an observation of a sequence rather than a single frame, which does not exactly satisfy the assumption of the [hidden Markov model] and the Viterbi algorithm, but we ignore this for simplicity" [50, p. 4].

Essentially, this amounts to *PEFAC*, but with more complex post processing.

The resulting algorithm was evaluated on the *TIMIT* database and *NOISEX* noises, as well as *FDA* and another noise database, with the same $\pm 5\%$ GPE and VDE as in the *PEFAC* paper. Its accuracy and voicing decision were compared against *PEFAC*, *SACC* and two other PDAs. Where PDAs could be trained, they were trained on the same data as *DNN/RNN*. In the results, *DNN/RNN* is particularly strong at negative SNRs, with little benefit to differentiate between *DNN* and *RNN*, and is reported to generalize well to untrained speech and noise corpora as well as reverberant conditions.

Since the neural networks were seemingly trained on 16 kHz data, all audio material in the replication dataset was resampled accordingly.

11.3.21 *KALDI* [42] (2014)

Source code "based on" the publication at `https://github.com/LvHang/pitch`*, in the C programming language, by some of the original authors.*

The *KALDI* PDA by Gahremani at al. from John Hopkins University, USA, part of the automatic speech recognition (ASR) toolkit of the same name, is "an algorithm that produces pitch and probability-of-voicing estimates for use as features in automatic speech recognition systems. These features give large performance improvements on tonal languages for ASR systems, and even substantial improvements for non-tonal languages" [42, p. 1].

This PDA is explicitly based on *RAPT*, but expanded in several ways. The signal is low-pass filtered, energy-normalized, and zero-meaned before calculating a normalized autocorrelation much like in *RAPT*. Instead of a threshold, a viterbi search chooses an optimal pitch track through the entire autocorrelation space of each signal.

The end goal of the algorithm is to drive an ASR system. In addition to the fundamental frequency estimate, a voicing decision is calculated from an approximation of a log-likelihood ratio from each frame's normalized autocorrelation. Two additional voicing determination features are forwarded to the ASR system as well.

An evaluation was carried out on the clean-speech *KEELE* corpus in comparison with *YIN*, *RAPT*, *SACC*, *SWIPE*, *YAAPT*, and one more PDA in terms of GPEs. All further evaluations use the word error rates typical in ASR research, which find the *KALDI* PDA better suited for this task than *SACC*, *YIN*, or *RAPT*, especially for pitch-sensitive tonal languages.

11.3.22 *NLS* [104, 103] (2016)

Implementation by the original authors at `https://github.com/jkjaer/fastF0Nls`*, written in Matlab. A different version also by the original authors used to be available at* `http://vbn.aau.dk/en/publications/fast-fundamental-frequency-estimation-making-a-statistically-efficient-estimator-computationally-`

efficient(c9604a90-5140-40fa-b973-7feea1fa3ea7).html and is also included in the dataset.

The non-linear least squares algorithm by Nielsen et al. of Aalborg University, Denmark assumes a harmonic spectrum in slowly-varying noise, and derives a maximum likelihood estimator, which "is the most accurate estimator in statistical terms. When the noise is assumed to be white and Gaussian, the [maximum likelihood] estimator is identical to the [NLS] estimator" [103, p. 1]. The algorithm is quoted to be based on earlier astronomical work from 1991 [126].

This is a member of a class of fundamental frequency algorithms that has not been discussed yet, where a parametric estimator is derived directly from a signal model, an even stronger assumption than general parametric estimators such as the aforementioned *SAFE*. As the *PEFAC* paper states, "The advantages of the parametric approach to pitch estimation are that the assumptions about the signal are explicit, the limitations of an algorithm are often predictable, the performance can be optimal in a well defined sense and in some cases a Cramér-Rao lower bound can be calculated or estimated. The disadvantage of the approach is that the performance may degrade when the (often quite strong) modeling assumptions are not satisfied." [45, p. 1]

The resulting PDA closely resembles a harmonic comb, albeit with dynamically adjusted teeth weight, which is "statistically efficient asymptotically. That is, it has the optimal estimation accuracy when enough data are available" [103, p. 3]. The math for this operation is quite complex, and involves a number of simplifying assumptions to run at acceptable speeds.

An evaluation was carried out with harmonic tone complexes in white noise, exactly satisfying the signal model. In this case, the *NLS* algorithm "can attain the Cramér-Rao lower bound and is the most accurate method for low-fundamental frequencies" [103, p. 10]. Results are compared to *YIN*, which is shown to be less accurate, particularly at negative SNRs.

It should be noted that the publication is unusually coy about practical matters and mentions *human speech* only in passing. The source code repository, however, clearly targets this application, although with the following caveats: "Please note that the code only contains a pitch estimator and NOT a pitch tracker. The difference is that a tracker contains a smoothing step on top of the estimator. The smooting step is there to minimise the risk of, e.g., octave errors (aka pitch halving) by smoothing out the estimates produced by the estimator which typically analyse the data on a segment-by-segment basis. Of course, our estimator can also be used inside a pitch tracker. For the best performance, we recommend that the smoothing step by Tabrikian et al. is used." and "For voiced speech, where the lowest fundamental frequency is typically bigger than 80 Hz, [...] the estimator typically works well down to a segment length of 12.5 ms. [...] In our experience, the estimator does typically not break down if the noise is not white and Gaussian. However, if the noise is coloured and has most of the energy at the lower frequencies, then the estimator can suffer from problems with octave errors. In this case, we recommend that some kind of pre-whitening is applied to the data prior to estimating the fundamental frequency."[15]

The replication dataset includes two implementations of the *NLS* algorithm, both implemented in Matlab. These were published in 2015 and 2016 by the same authors, yet seem to have been developed separately.

11.3.23 *CREPE* [79] (2018)

Implementation by the original authors at https://github.com/marl/crepe in Python.

The convolutional representation for pitch estimation, *CREPE*, from Kim et al. at New York University, USA, is a "data-driven pitch tracking algorithm, [...] which is based on a deep convolutional

[15] from https://github.com/jkjaer/fastF0Nls, downloaded Jun 2020

neural network that operates directly on the time-domain waveform." This approach is motivated by reframing the rising popularity of machine learning techniques in signal processing as a removal of a defect:

> "A notable trend in [older PDAs] is that the derivation of a better pitch detection system solely depends on cleverly devising a robust candidate-generating function and/or sophisticated post-processing steps, i.e. heuristics, and none of them are directly learned from data, except for manual hyperparameter tuning." [79, p. 1]

The neural network in question has an input layer of 1024 audio samples, six hidden layers, one 2048-neuron latent representation, and one 360 sigmoid output layer that corresponds to pitch estimates between 32.7 Hz and 1975.5 Hz in cent steps. The final fundamental frequency estimate is given as the weighted average of the output layer.

This neural network was developed, trained, and tested with partitions of 6.16 h of synthetic music made from "a fixed sum of a small number of sinusoids, meaning the dataset is highly homogeneous in timbre" as well as 15.56 h of more complex synthesized music "with a perfect f0 annotation that maintains the timbre and dynamics of the original track" [79, p. 3]. These signals contain 25 different synthesized instruments and singers.

An evaluation measured the algorithm's accuracy in the musical equivalent to GPE against a variant of *YIN* and *SWIPE* in various realistic and synthetic noises. These evaluations show that "*CREPE* performs better in all cases where the SNR is below 10 dB while the performance varies depending on the spectral properties of the noise when the noise level is higher, which indicates that our approach can be reliable under a reasonable amount of additive noise." Of particular note is *CREPE*'s very small fine pitch error, which "suggests that *CREPE* is especially preferable when even minor deviations from the true pitch should be avoided as best as possible" [79, p. 3].

While most of these examples cater more towards musical applications than speech, frequent references are made to speech in the publication and source code repository. The published model provided with the source code is in part trained on singing voices.

11.3.24 *MAPS* (Chapter 8)

Original implementation, as detailed in chapter 8, and available online at `https://bastibe.github.io/Dissertation-Website/maps/index.html` *in various programming languages.*

The magnitude and phase spectrogram based fundamental frequency estimator, *MAPS*, by Bechtold et al. of Oldenburg University, Germany, was developed as part of the present dissertation, and is a harmonicity detector in the frequency domain. It is comprised of two parts, a harmonic comb in the magnitude spectrum, and a harmonic sawtooth in the time derivative of the phase spectrum. This dual signal model is motivated as "It uses more of the available information in the signal, and it can use phase spectral information to account for octave ambiguities in the magnitude spectrum. Finally, it can use the magnitude spectral information to help the phase spectrum discern salient parts from non-salient ones" (p. 64).

The magnitude harmonic comb is a variant of *PEFAC*'s soft comb, in that it is a mean-normalized soft comb with sharp positive teeth and shallow negative gaps. However, the comb teeth correspond to window function spectra instead of reciprocal raised cosines. This is also applied in linear instead of logarithmic frequency, necessitating a separate frequency weighting to attenuate the effect of higher harmonics.

The phase sawtooth follows the shape of the instantaneous frequency deviation, a variant of the time-derivative of the phase spectrum. Since the phase spectrum is naturally confined to values

around zero, it can be subtracted from the signal phase spectrum, instead of the correlation used in the magnitude.

These two estimators, the magnitude correlation and the instantaneous frequency difference, are combined in a Bayesian voicing determination framework trained on speech from a *PTDB-TUG* corpus training set and noise from the *QUT-NOISE* corpus. The result "does not estimate the probability of general voice activity within a frame, but the specific confidence that a pitch can be estimated accurately at the current frame" (p. 71).

An evaluation compared *MAPS* to *PEFAC*, *RAPT*, and *YIN* with a *PTDB-TUG* corpus test set and noise from *QUT-NOISE*. Aside from a GPE and FPE evaluation, considerable effort was made to highlight the algorithm's voicing decision, which is unusually conservative, but "its positive VAD decisions almost always result in accurate and precise estimates, as evidenced by the negligible false positive rate and very low GPEs. Furthermore, this remains true even at low SNRs, even though false negatives clearly deteriorate at low SNRs" (p. 74).

Since the algorithm was trained at 48 kHz, all audio data is resampled accordingly in the replication dataset. Source code for the PDA is available in Julia, Python, and Matlab, and the Python version was chosen for the replication dataset, as it was the original, and fastest, implementation.

11.4 Literary Survey

The 25 PDAs presented in the previous chapter span more than half a century, and a wide variety of different methods, from analogue-inspired period detectors in the time domain, to autocorrelation-based methods with various pre- and post processing methods, to harmonicity estimators in the frequency domain, to statistical methods, and opaque machine learning PDAs. As more processing power became available over the years, complexity rose in lockstep, and ever more elaborate schemes were attempted.

All of these methods aim to estimate the pitch of the human voice, nebulously defined either from a time-domain periodicity measure, or frequency-domain harmonicity, or later, a database of recordings with pre-computed ground truths. As these definitions of pitch or fundamental frequency differ, so does their performance for different signal conditions. Hence, comparisons between PDAs were conducted to assess their differences.

Figure 11.1 shows the results of clean-speech comparisons published in the above 22 publications. This includes every tabulated Gross Pitch Error measure of clean-speech *FDA* or *KEELE* recordings that occurred in at least two papers. On the one hand, a good consensus has emerged with regards to common databases and error measures. On the other hand, implementations clearly differ wildly, as do results.

The *CEP* and *RAPT* PDAs appear particularly variant in Figure 11.1, perhaps because their apparent simplicity invited frequent re-implementations with subtly varied details. More complex algorithms had to rely on implementations by their original authors, which made results more consistent. Nevertheless, there are always outliers beyond any reasonable standard deviation. This highlights once again the need for a consistent framework for comparisons, such as the replication dataset.

These 25 PDAs in the replication dataset make up only a very limited subset of the full breadth of fundamental frequency estimation research. In the last thirty years from 1990-2020, a literary search turned up 851 publications on the subject, as detailed in Section 9.2.

Figure 11.2 shows mentions of the 25 PDAs above in these publications over the years. Over this time period, the yearly volume of publications on PDAs has roughly quadrupled, with a sharp increase around 2005, no doubt spurred by the new availability of serious computing power and the Internet. This allowed much broader access to other groups' source code and evaluation databases, as well as the processing power to conduct comparisons and evaluations on meaningfully large datasets.

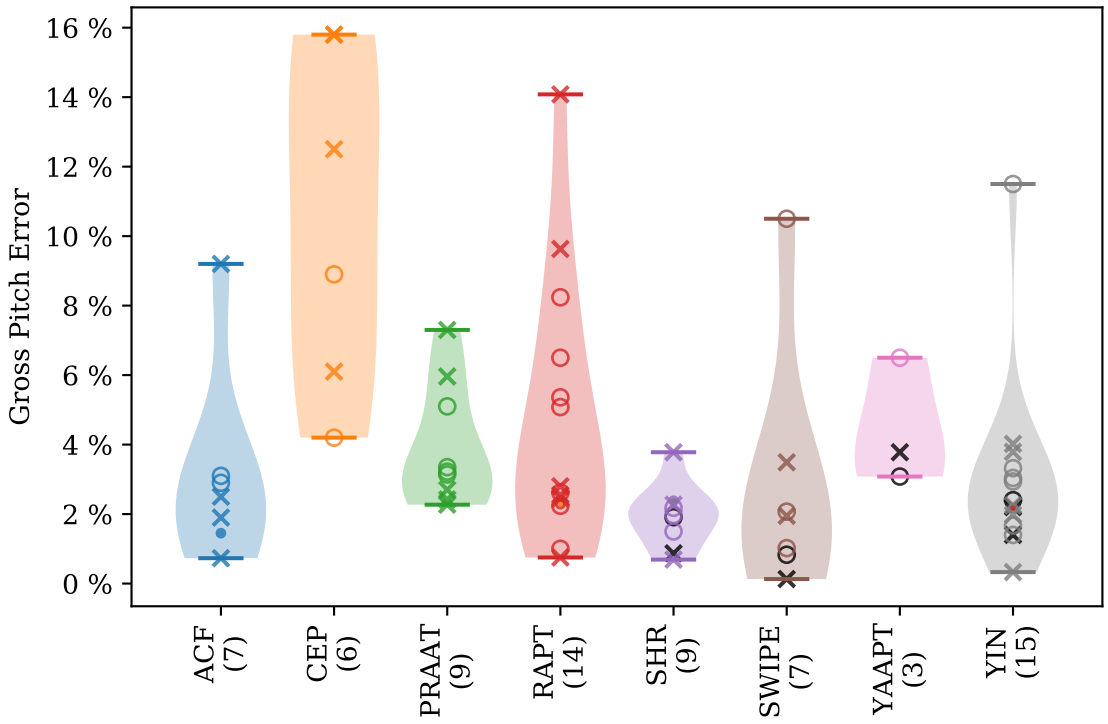


Figure 11.1: Violin plot of GPE of PDAs for clean speech on the *KEELE* (○), *FDA* (×), or both (·) corpora, according to the PDAs' publications. Black symbols are from the PDAs' own publication, if it included a comparison.

Consequently, the number of mentions of published PDAs rose even more significantly in this time frame than the total volume of publications.

Of particular popularity were *CEP*, *PRAAT*, *YIN*, and *STRAIGHT*; *AMDF* and *RAPT*, and later *PEFAC* then showed intermittent popularity, according to Figure 11.2. Interestingly, this list is different from anecdotal evidence of frequently *compared* PDAs, probably due to source code availability issues.

Figure 11.3 graphs mentions of the same PDAs across publications. Indeed, the majority of mentions are in a small number of journals, most notably the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), the INTERSPEECH conference, the European Signal Processing Conference (EUSIPCO), the IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)[16], and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), in order of overall popularity.

Interestingly, the relative popularity of *YIN*, *RAPT*, *PRAAT*, and *STRAIGHT* seem to be mostly driven by the ICASSP, INTERSPEECH, and TASLP publications. These trends are indications of a cultural norm in these journals. This norm defines *YIN*, *RAPT*, *PRAAT*, and possibly *STRAIGHT* and *PEFAC* as archetypes of fundamental frequency estimation that must be mentioned or compared against when publishing new PDAs.

In contrast, far fewer mentions of these PDAs originate from the EUSIPCO and WASPAA conferences, which thus buck this cultural norm, despite being similarly popular for publications on fundamental frequency in general (see table 9.2 on page 82).

These 851 publications include works by 1731 authors from all over the world. To gain a sense of

[16]including the older IEEE Transactions on Speech and Audio Processing

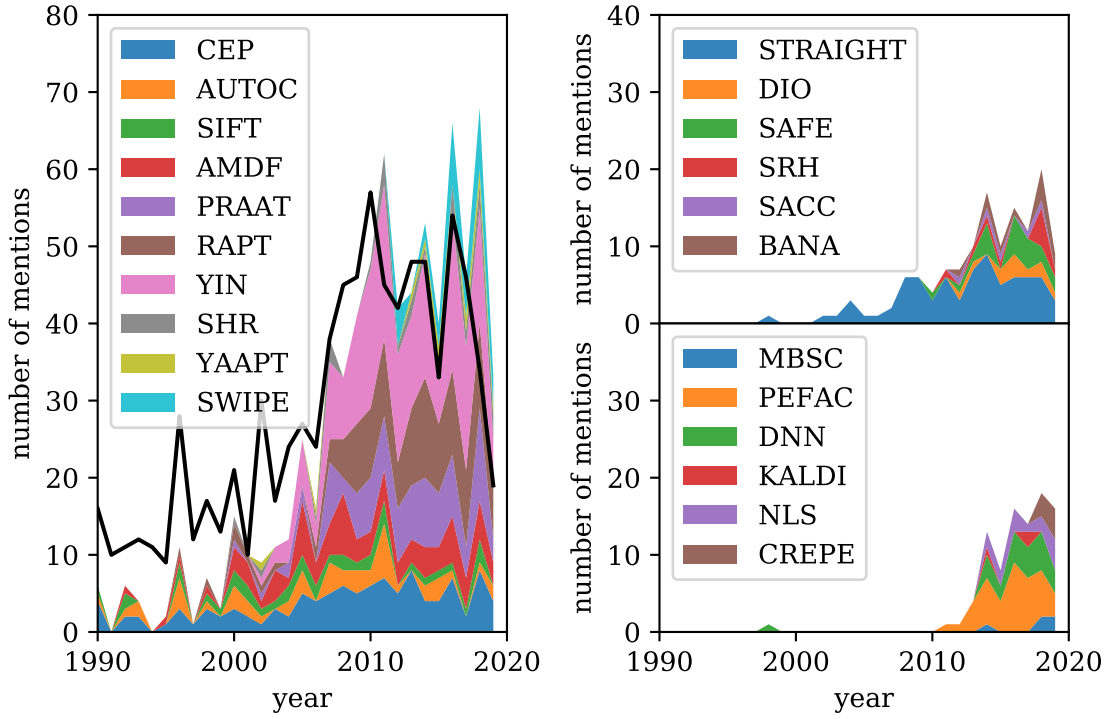


Figure 11.2: Mentions of each PDA in the replication dataset in papers on fundamental frequency estimation from 1990 to 2020. Only counts papers that contain both a variant of the PDA's name, and the author's last name. Black line indicates total number of papers per year. Sum of PDAs can be higher than total number, since paper can mention more than one PDA.

their scientific interactions, Figure 11.4 shows a network graph of the most significant 280 authors, accounting for 234 publications and their scientific relationships to one another. Authors were identified by their first and last name[17], as recorded in the papers' publication metadata, which may have duplicated some authors if they used different spellings in different publications.

Particularly interesting are dense clusters of prolific authors, which indicate close-knit groups with frequent intra-group collaborations. The largest of these is the group involving Jesper Jensen, Mads Christensen, and Andreas Jakobsson of Aalborg University and Lund University, which have released 48 publications in their group since 2006, involving 43 coauthors.

Other prolific groups with at least ten publications are centered around Keikichi Hirose of the University of Tokyo, Japan with 14 publications and 18 coauthors, the group of C. Shahnaz and M. Ahmad of Concordia University in Quebec, Canada with 16 publications and 10 coauthors, and DeLiang Wang of Ohio State University, USA with 14 publications and 10 coauthors.

The density of clusters in Figure 11.4 is proportional to the number of papers published together. It is heartening to see that some research groups indeed collaborate on doing science, particularly in Asian countries. Most groups are centered around one or two professors, and involve many one-off collaborators, presumably students. A few groups also habitually coauthor within the group.

In general, however, the clusters are separate from one another, and show little cross-group collaboration. Perhaps surprisingly, many groups' work spans multiple decades. Fundamental frequency estimation thus remains a fruitful field of study despite more than fifty years of research already having

[17] more accurately, the first sequence of non-whitespace letters in the first name, and their full last name, to be at least somewhat resilient to different spellings, while still properly differentiating between similarly named authors.

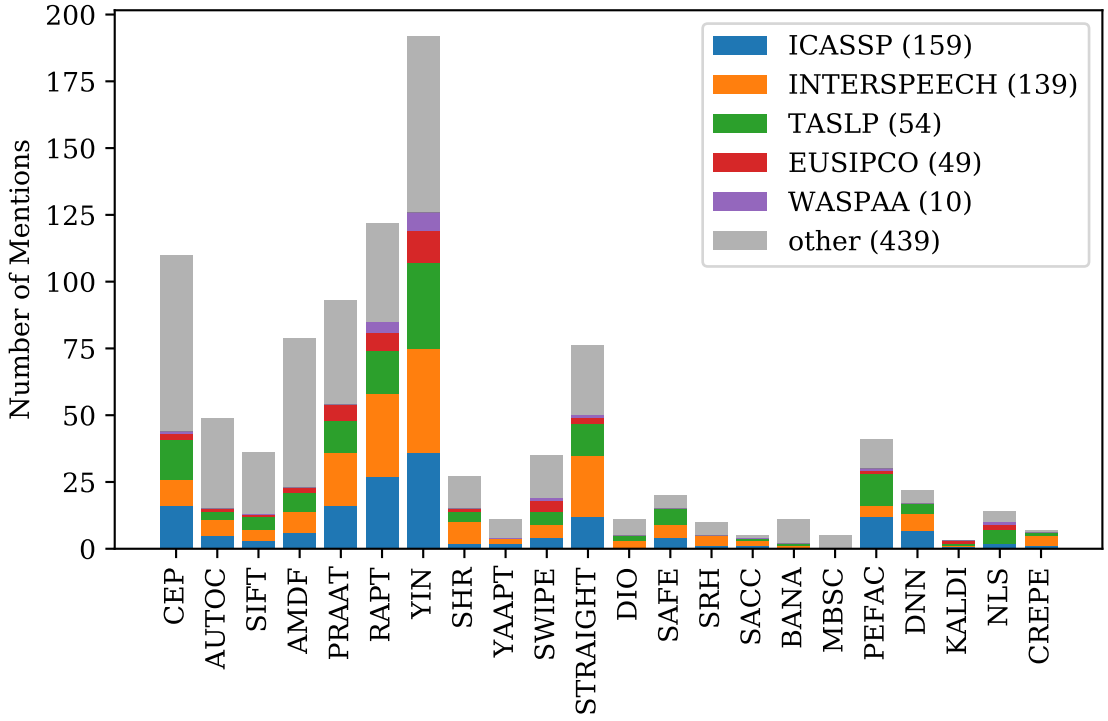


Figure 11.3: Mentions of each PDA in the replication dataset per journals in papers on fundamental frequency estimation from 1990 to 2020. Only counts papers that contain both a variant of the PDA's name, and the author's last name. Colored journals have at least 10 publications.

been published.

With such an abundance of PDAs published over the years, many people have endeavored to categorize them by some measures. For example, the book *Pitch Determination of Speech Signals* by Hess in 1983 [59] includes an impressive map that classifies algorithms into a hierarchy of algorithms, with the biggest split between (analog) real-time algorithms and (digital) short-term PDAs. Short-term PDAs are further subdivided into correlation-based time-domain and frequency-domain PDAs. The latter split between algorithms is still popular today [148, 41], where one class of PDAs searches for patterns in the time domain and another seeks them in the frequency domain. Alternative classifications split PDAs by whether they fit a complete model of speech and noise to a signal, or maximize some fitness measure of speech only [21, 45]. Conversely, most current classifications follow depending on whether that model was crafted by humans or inferred by a machine learning process [79].

Nowadays, as PDAs are becoming more and more complex, it could be argued that such classifications are becoming meaningless. Even classical time-domain PDAs, such as *DIO*, *SACC*, and *KALDI*, now include frequency-domain features such as viterbi-searches or filterbanks. Simultaneously, opaque deep-learning-based methods might calculate cross-domain features internally that were not even explicitly designed at all. The only tentative classification still applicable is whether the PDA's understanding of pitch is based on a production-centric periodic signal model, or a perception-motivated harmonic model. But even that is more of a characterization of the authors' beliefs than of algorithmic parameters.

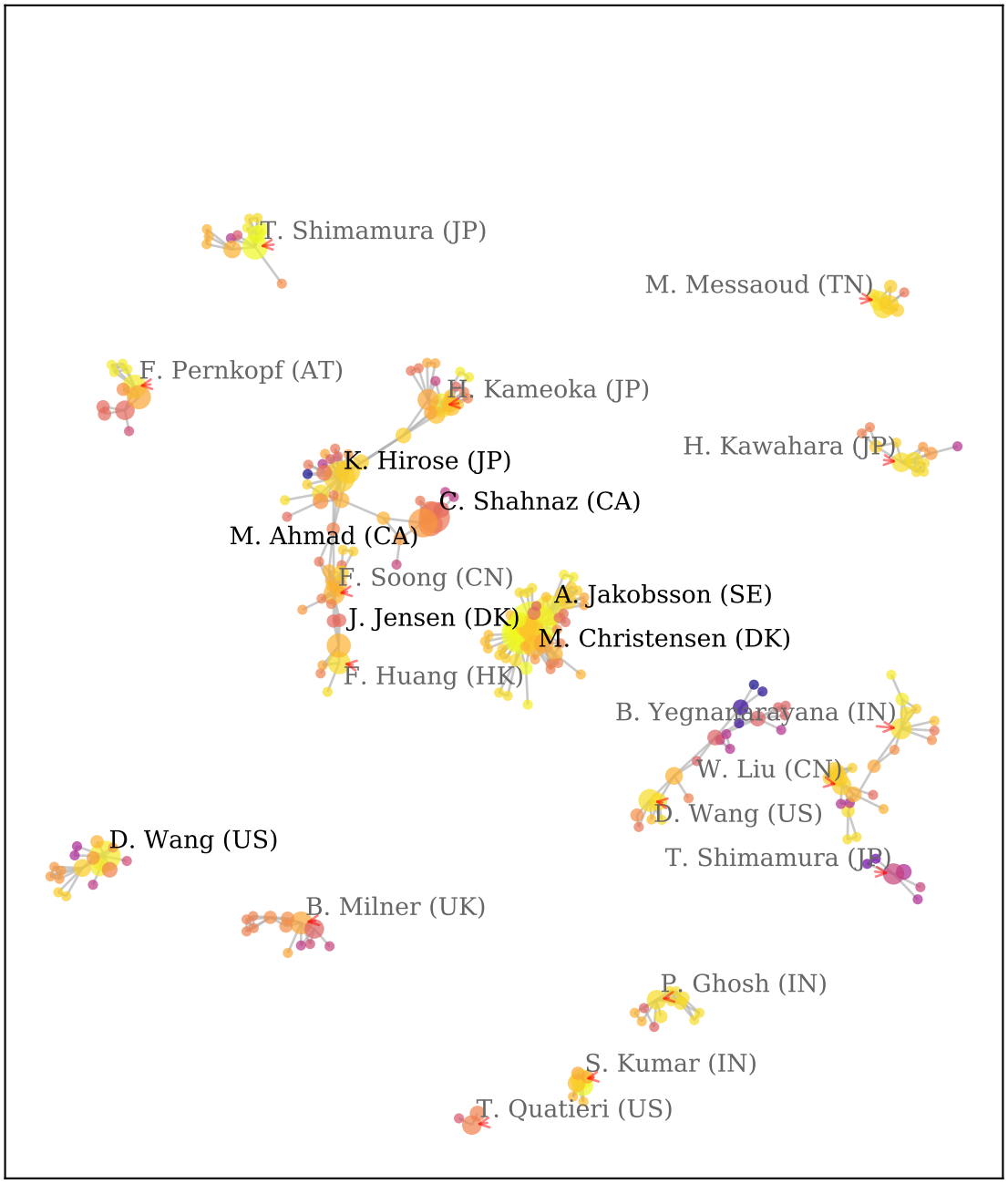


Figure 11.4: Network of significant authors and their papers on fundamental frequency estimation. Only networks with at least five publications are included. Each bubble represents an author with the size proportional to the number of papers on PDAs. Distance between authors is roughly inversely proportional to the number of papers published together. Bubble color indicates age of the latest paper from 1990 (blue) to 2020 (yellow). Author names in black have ten or more papers. Grey author names have five or more papers, and are included only where there was space on the graph for them.

11.5 Dataset Definition

The last chapters detailed the landscape of PDAs, and their historical and current context. To make sense of this large plurality of methods, a framework of comparison is required. Thus, the stage is now set to actually assemble the replication dataset from PDAs, corpora, and performance measures.

The preceding sections established the following PDAs as worthwhile for comparison: *CEP*, *AUTOC*, *SIFT*, *AMDF*, *PRAAT*, *RAPT*, *YIN*, *SHR*, *YAAPT*, *SWIPE*, *STRAIGHT*, DOI, *SAFE*, *SRH*, *SACC*, *BANA*, *MBSC*, *PEFAC*, *DNN*, *KALDI*, *NLS*, *CREPE*, *MAPS*. Chapter 9 detailed the following speech corpora: *CMU-ARCTIC*, *FDA*, *KEELE*-mod, *MOCHA-TIMIT*, *PTDB-TUG*, *TIMIT*; and the noise corpora *NOISEX*, *QUT-NOISE*. To compare PDA estimates of these data, performance measures were calculated against the corpus ground truths, the consensus truth from Chapter 10, and each PDA's own clean estimate.

No exhaustive comparison of all combinations of speech and noise can be possible, as recordings can be mixed with an infinite variety of sections for each noise recording, and at an infinite number of SNRs. Thus, some kind of selection had to be made, both to keep the dataset at a manageable size, and to keep the computational requirements feasible.

Noise recordings and speech corpora are already built to be somewhat homogeneous and can therefore be randomly sampled. SNRs are mostly interesting in the range between mostly clean (20 dB SNR) to mostly noise (-20 dB SNR), and typically remain at a steady upper or lower bound beyond. As results typically vary relatively smoothly with SNR, the SNR range can be evaluated in somewhat coarse 5 dB steps.

In total, the replication dataset contains the results of running 25 PDAs on 6 speech corpora, 35 noise recordings from two noise corpora, at 9 different SNRs, each repeated 20 times with different speech recordings and noise sections. This constituted a total of 945000 experiments.

Additionally, an upper bound on estimation accuracy was established with 20 repetitions of synthetic speech-like harmonic tone complexes of 9 fundamental frequencies between 80 and 260 Hz in white noise in 9 SNRs for another 40500 experiments.

The resulting fundamental frequency tracks alone combine to about 6 GB of data, and the performance measures an additional 600 MB.

11.5.1 Experiments

To run these experiments, speech signals and sections of noise signals were mixed at a given SNR level. For this to work, signals needed to be of compatible length, sampling rate, and channel count, which meant cutting, resampling, and downmixing.

Noise or speech recordings with more than one channel were downmixed by using only the first channel and ignoring all other channels.

The noise signal section was randomly chosen to start within 0-3 min for the *NOISEX* corpus, or 5-25 min for the *QUT-NOISE* corpus. As speech signals were guaranteed to be shorter than 30 s, this ensured that noise recordings of sufficient length were able to be extracted. In the case of *QUT-NOISE*, the first and last minutes of noise recordings were ignored, as they frequently contained calibration signals that would disturb experimental results. Thus, regretfully, these specifications ignored parts of the noise recordings to simplify computations.

Between signal and noise, the speech signal was deemed more important. Thus, if sampling rates between signal and noise differed, the noise signal was resampled to the sampling rate of the speech signal, and not vice versa, using the resampy[18] library.

[18] `https://github.com/bmcfee/resampy`, version >0.2

SNR

Mixing speech and noise at a specific SNR was done by adjusting the noise power such that speech sections were at the selected SNR. The procedure used in the replication dataset accomplished this by applying a gain factor on the noise signal, but explicitly did not touch the speech signal in order to keep it as true to the original recording as possible.

To calculate the gain factor on the noise signal, an estimate of the speech level and noise level were obtained in two different ways:

Since the noise recordings were comparatively stationary, the noise level was simply the root mean square (RMS) level of the entire noise section.

The speech signal however contained pauses and short stretches of silence at the beginning and end of each recording. Thus the level was instead calculated from active speech segments only: For each 20-ms block in the speech signal, a logarithmic RMS level in dB was calculated, and a threshold set as the mean between the 5th and 95th percentile of these RMS levels to classify speech from silence. The overall speech RMS level is then calculated only from the speech signal blocks that exceeded the threshold.

This procedure is somewhat arbitrary. Different thresholds or block lengths could have been chosen, which would have resulted in slightly different level estimates and bias the SNR somewhat differently. However, there does not seem to be a consensus on these matters, and many publications indeed do not even specify their particular method for estimating the SNR; hence a simple method was given preference.

Additionally, the replication dataset needed to be entirely reproducible by new PDAs; thus, a simple procedure was chosen instead of a possibly more "correct" one so as to aid future scientists in producing comparable results.

Summary: Experimental Parameters

In summary, each experiment had the following parameters:

- PDA name
- speech recording name
- speech corpus name
- noise recording name
- noise corpus name
- noise segment start time in seconds
- SNR in dB
- repetition index
- sampling rate in Hz

From these parameters, the exact test signal was able to be reconstructed from the corpora, thereby enabling experimental reproducibility and validation.

Whenever a PDA had algorithmic parameters, they were either left at their default values if present, left at the default values provided by the PDA's examples or documentation, or specified in the PDA's description at the beginning of this chapter.

The source code for running these experiments, including source code for all the PDAs, have been included on this dissertations' website at `https://bastibe.github.io/Dissertation-Website/`

Synthetic Signals

In addition to the speech recordings detailed above, the replication dataset also includes a shorter evaluation of synthetic harmonic tone complexes, so as to gain insight into the PDAs' performance measures in ideal conditions.

For this purpose, speech-like harmonic tone complexes with ten harmonics and variable fundamental frequency were mixed with white noise. These harmonic tone complexes were perfectly harmonic and perfectly periodic, and thereby satisfied the signal models of most PDAs, while still mildly resembling voiced speech.

Fundamental frequencies were modulated at 1 Hz, with an amplitude of $\sqrt{2}$ around the base $f_{0\mathrm{B}}$, such that

$$f_0(t) = f_{0\mathrm{B}}\sqrt{2}\,(1 + \sin(2\pi t)) \tag{11.1}$$

The harmonic tone complex was then constructed from ten modulated harmonics of this fundamental frequency track:

$$\mathrm{HTC}(t) = \sum_{p=1}^{10} \cos\left(\int_{\tau=0}^{t} 2\pi\,\frac{p f_0(\tau)}{f_s}\,\mathrm{d}\tau\right) \tag{11.2}$$

Finally, the result was low-pass filtered with a first-order butterworth filter at 2000 Hz to approximate the spectral characteristics of human speech.

The replication dataset includes five seconds of these signals for fundamental frequencies ranging from 80 Hz to 260 Hz in 20 Hz increments, and SNRs from -20 dB to 20 dB in 5 Hz increments.

11.5.2 Evaluation

In addition to the fundamental frequency estimates of each test signal and PDA, the replication dataset contains pre-calculated performance measures for the experiments. The following sections provide a precise definition of each performance measure, and how to calculate it.

These includes common performance measures such as the gross pitch error and fine pitch errors, as well as evaluations not commonly seen in publications, like octave pitch errors, gross remaining errors, and fine remaining bias.

All performance measures were carried out with five different ground truths:

- Speech corpus ground truth (if available) with estimated VAD (if available)
- Speech corpus ground truth (if available) with speech corpus VAD (if available)
- Consensus truth with estimated VAD (if available)
- Consensus truth with consensus VAD
- Clean estimate as ground truth with clean estimated VAD

The last one of these uses each PDA's own estimate without noise as ground truth for a noisy-speech estimate. It thus provides a baseline of how robust a PDA is to noise, without considering other estimation errors.

Preprocessing

PDAs estimate fundamental frequency differently in terms of their time bases, in how they report missing data such as VAD negatives, and in terms of how (and whether) they report their VAD estimates.

As such, as a baseline, all of these estimates had to reflect a comparable format. For the purpose of the replication dataset, this meant interpolating fundamental frequency estimates to a common time base. To avoid interpolation errors, the replication dataset works on each ground truths' time base and chooses the nearest estimate for each instance.

Since a number of PDAs report VAD-negative pitches as 0 Hz, all other PDAs' VAD-negative pitches were also set to 0 Hz. This is unfortunate, as it discards possibly recoverable data in some cases; however, not doing so would disadvantage the PDAs without VAD-negative estimates, and thus create an unfair comparison. In general, VAD estimates were treated as unvoiced if their voicing predictor is < 0.5, and voiced otherwise.

If a PDA did not have a VAD, the consensus truth's VAD was substituted. This provided a slight advantage to these PDAs when evaluates against the consensus truth, as they were incapable of VAD errors in this case.

Gross Pitch Error (GPE)

The most important performance measure for fundamental frequency estimation is the gross pitch error, the percentage of pitches where the estimated pitch $f_{\text{est}}(t)$ deviates from the true pitch $f_{\text{true}}(t)$ by more than 20 %:

$$\text{GPE} = \frac{\sum_t [\![\, \Delta f(t) \geq 0.2 \wedge v(t) \,]\!]_{\text{I}}}{\sum_t [\![\, v(t) \,]\!]_{\text{I}}} \tag{11.3}$$

with the quotient between estimate and truth denoted as

$$\Delta f(t) = \left| \frac{f_{\text{est}}(t)}{f_{\text{true}}(t)} - 1 \right| \tag{11.4}$$

and normalized to frames that are voiced both according to the PDA and the ground truth:

$$v(t) = f_{\text{true}}(t) \neq 0 \wedge f_{\text{est}}(t) \neq 0 \tag{11.5}$$

where t is the arbitrary time index of the ground truth, and $[\![\cdot]\!]_{\text{I}}$ is the Iverson Bracket, which is 0 or 1 depending on the logical proposition inside. As mentioned before, VAD decisions are at this point encoded into the frequency estimates, which is either positive for VAD-positive, or zero otherwise.

The value of $\pm 20\%$ is somewhat arbitrary, and GPEs have been variously defined with $\pm 10\%$ or even $\pm 5\%$. However, $\pm 20\%$ seems to be the most common definition and suitable for a human or prosodic understanding of pitch, where fine details are less important than the overall pitch shape. Musical applications might want to choose a stricter standard, such as half semitones ($\pm 3\%$).

Furthermore, the GPE is only defined for frames that are voiced both in the ground truth and in the estimator. This is necessary, as some estimators and truths do not provide estimates for unvoiced frames. In the above equations, VAD negatives are assumed to be $f_{\text{est}} = 0$.

The normalization to all correctly voiced frames is again somewhat arbitrary and has been defined differently in a few cases. Truthfully, however, these details are often not defined at all and implicitly bias the data. Options include all ground-truth voiced frames, or all speech frames. The normalization to all correctly voiced frames is a reasonable option, as it gives a good estimate of the PDA's best-case accuracy, when the frequency estimate agrees with the VAD.

Fine Pitch Error (FPE)

The fine pitch error further investigates grossly correct frames, and is the mean error of grossly correct estimates,

$$\mathrm{FPE} = \mathrm{mean}\left(\Delta f(t) \middle| \Delta f(t) < 0.2 \wedge v(t)\right). \tag{11.6}$$

Again, there are varying definitions with, for example, a standard deviation instead of the mean error, or a frequency difference instead of the quotient. Their meaning, however, remains very similar, and the mean error is at least easily interpretable with respect to the GPE, having a maximum of 10 % for entirely random pitches.

High/Low Octave Pitch Error (OPE)

Octave errors are a subset of gross errors that happen to be at an integer multiple of the true pitch. They deserve specific mention, as they are a common error mode of PDAs, where the correct periodicity or harmonicity is found, but a subharmonic is mistaken for the fundamental:

$$\mathrm{OPE}_{\mathrm{high}} = \frac{\sum_t [\![\, |\Delta f(t) - \lfloor \Delta f(t) \rceil| < 0.1 \wedge \frac{f_{\mathrm{est}}(t)}{f_{\mathrm{true}}(t)} \geq 1.2 \wedge v(t) \,]\!]_{\mathrm{I}}}{\sum_t [\![\, v(t) \,]\!]_{\mathrm{I}}} \tag{11.7}$$

$$\mathrm{OPE}_{\mathrm{low}} = \frac{\sum_t [\![\, |\Delta f(t) - \lfloor \Delta f(t) \rceil| < 0.1 \wedge \frac{f_{\mathrm{est}}(t)}{f_{\mathrm{true}}(t)} \leq 0.8 \wedge v(t) \,]\!]_{\mathrm{I}}}{\sum_t [\![\, v(t) \,]\!]_{\mathrm{I}}} \tag{11.8}$$

$$\mathrm{OPE}_{\mathrm{both}} = \mathrm{OPE}_{\mathrm{high}} + \mathrm{OPE}_{\mathrm{low}}, \tag{11.9}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. The first term in the numerator selects pitches within $\pm 10\%$ of every integer multiple of the true pitch, the second term excludes the true pitch itself, and the third term selects only voiced t. Octave errors are a very specific kind of error and thus use a narrower $\pm 10\%$ definition than the $\pm 20\%$ "gross" margin of the GPE.

Gross Remaining Error (GRE)

Any gross error that is not an octave error is a remaining error

$$\mathrm{GRE} = \frac{\sum_t [\![\, |\Delta f(t) - \lfloor \Delta f(t) \rceil| \geq 0.1 \wedge \Delta f(t) \geq 0.2 \wedge v(t) \,]\!]_{\mathrm{I}}}{\sum_t [\![\, v(t) \,]\!]_{\mathrm{I}}}. \tag{11.10}$$

These errors are misclassifications that do not bear any resemblance to the true fundamental frequency. In general, $\mathrm{GRE} + \mathrm{OPE}_{\mathrm{both}} = \mathrm{GPE}$.

Fine Remaining Bias (FRB)

If a PDA can't estimate the true pitch or its harmonic, its guess is often not quite random. Most PDAs have a built-in bias towards higher or lower pitches, which is particularly visible in the absence of a clear pitch. Hence, the FRB is the median of gross remaining errors

$$\mathrm{FRB} = \mathrm{median}\left(\frac{f_{\mathrm{est}}(t)}{f_{\mathrm{true}}(t)} \middle| |\Delta f(t) - \lfloor \Delta f(t) \rceil| > 0.1 \wedge \Delta f(t) > 0.2 \wedge v(t)\right). \tag{11.11}$$

These biases are often indicative of small implicit tendencies towards over- or under-estimating pitches in general.

Voicing Decision Errors

A different class of performance metrics is not concerned with the value of the estimated fundamental frequencies, but with the voicing decision of the PDA, if available. These are standardized measures, such as the true positive rate, false positive rate, or false negative rate. All of these are based on simple counts of voicing decisions:

$$\text{True Positives:} \quad \text{TP} = \sum_t [\![\, f_{\text{true}}(t) \neq 0 \wedge f_{\text{est}}(t) \neq 0 \,]\!]_{\text{I}} \tag{11.12}$$

$$\text{True Negatives:} \quad \text{TN} = \sum_t [\![\, f_{\text{true}}(t) = 0 \wedge f_{\text{est}}(t) = 0 \,]\!]_{\text{I}} \tag{11.13}$$

$$\text{False Positives:} \quad \text{FP} = \sum_t [\![\, f_{\text{true}}(t) = 0 \wedge f_{\text{est}}(t) \neq 0 \,]\!]_{\text{I}} \tag{11.14}$$

$$\text{False Negatives:} \quad \text{FN} = \sum_t [\![\, f_{\text{true}}(t) \neq 0 \wedge f_{\text{est}}(t) = 0 \,]\!]_{\text{I}} \tag{11.15}$$

From these, more comparable error rates are derived:

$$\text{True Positive Rate:} \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11.16}$$

$$\text{False Positive Rate:} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{11.17}$$

$$\text{False Negative Rate:} \quad \text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{11.18}$$

Another way of summarizing these features is *precision* and *recall*, which are "what proportion of the voiced estimates are truly voiced?" and "what proportion of the truly voiced frames are estimated as voiced?", respectively. These give a more accurate characterization of the voicing estimator's characteristics:

$$\text{Precision:} \quad \text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11.19}$$

$$\text{Recall:} \quad \text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11.20}$$

A handy single-number summary of these measures is the F-score, which summarizes the trade-off between precision and recall as

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{PRE} \cdot \text{REC}}{(\beta^2 \cdot \text{PRE}) + \text{REC}}, \tag{11.21}$$

where β indicates that recall is considered β times more important than precision.

Summary: Performance Measures

The replication dataset includes the following performance measures:

- Gross Pitch Error, the proportion of estimates within $\pm 20\%$ of the ground truth
- Fine Pitch Error, the mean error of grossly correct estimates

- High/Low Octave Pitch Error, the proportion of grossly incorrect estimates at integer multiples of the ground truth
- Gross Remaining Error, the non-octave gross errors
- Fine Remaining Bias, the median of non-octave grossly incorrect estimates
- True Positive Rate, the proportion of all voiced frames that are classified as voiced
- False Positive Rate, the proportion of all unvoiced frames that are classified as voiced
- False Negative Rate, the proportion of all voiced frames that are classified as unvoiced
- F_1, a measure for the equally-weighted combination of precision and recall

Of these, the gross and fine pitch errors are very common in the literature, while the others are expansions on existing performance measures, or entirely new additions. In particular, many publications include a *voicing decision error* metric that is differentiated into false negatives and positives in the replication dataset.

11.6 Computational Considerations

The replication dataset had to run each PDA on a large number of audio signals of varying lengths. Along with the experimental parameters and results, the computation time of every experiment was recorded in the dataset. This information can be used to gain insight into the computational performance characteristics of the PDAs.

Most PDAs used in this study were implemented in the programming language Matlab, except for *SAFE*, *PRAAT* and *KALDI*, which used C, and *SIFT* and *MAPS*, which used Python. *SACC* and *YIN* additionally used custom MEX-files (compiled C code for use in Matlab).

All experiments were run on a 16-core AMD Ryzen 1950X in 2019 using Matlab 2019a and Python 3.6 on Linux. All PDAs were run in a purpose-built multiprocessing framework, running 16 processes in parallel. For the purposes of these experiments, multithreading and the JVM in Matlab were disabled using the `-singleCompThread` and `-nojvm` command line arguments. The total computation time for all experiments was roughly 500 processor core days.

Figure 11.5 shows the computation time each PDA took for audio recordings of various durations. Each PDA's computation time can be split into a fixed startup/shutdown time, and a run time that scales with the duration of the recording. From the graph, the various PDAs can be broadly classified into one of three categories:

- *Near-constant-time* algorithms that show little scaling with duration (*DIO*, *KALDI*, *MAPS*, *PRAAT*, *SHR*, *SIFT*, *SRH*, *YAAPT*, *YIN*). These PDAs are employable in most situations, and computation time can be traded for latency if need be.

- *Near-real-time* algorithms that run close to real time (*BANA*, *CREPE*, *NLS2*, *RAPT*, *RNN*, *SACC*, *SAFE*, *SWIPE*). The slope of the run time is roughly parallel to the real-time line. Whether these PDAs can be used for an application depends on the target hardware.

- *Offline* algorithms that require significantly longer than real time (*MBSC*, *STRAIGHT*). The run-time slope of these PDAs is significantly steeper than real-time, or exponential. These kinds of algorithms can not be used in real-time.

Depending on the application and the target hardware, these characteristics for computation time can be highly relevant and may effectively make certain PDAs unusable in some cases. In fact, a few additional PDAs were available, but could not be included in the comparison, as they consumed

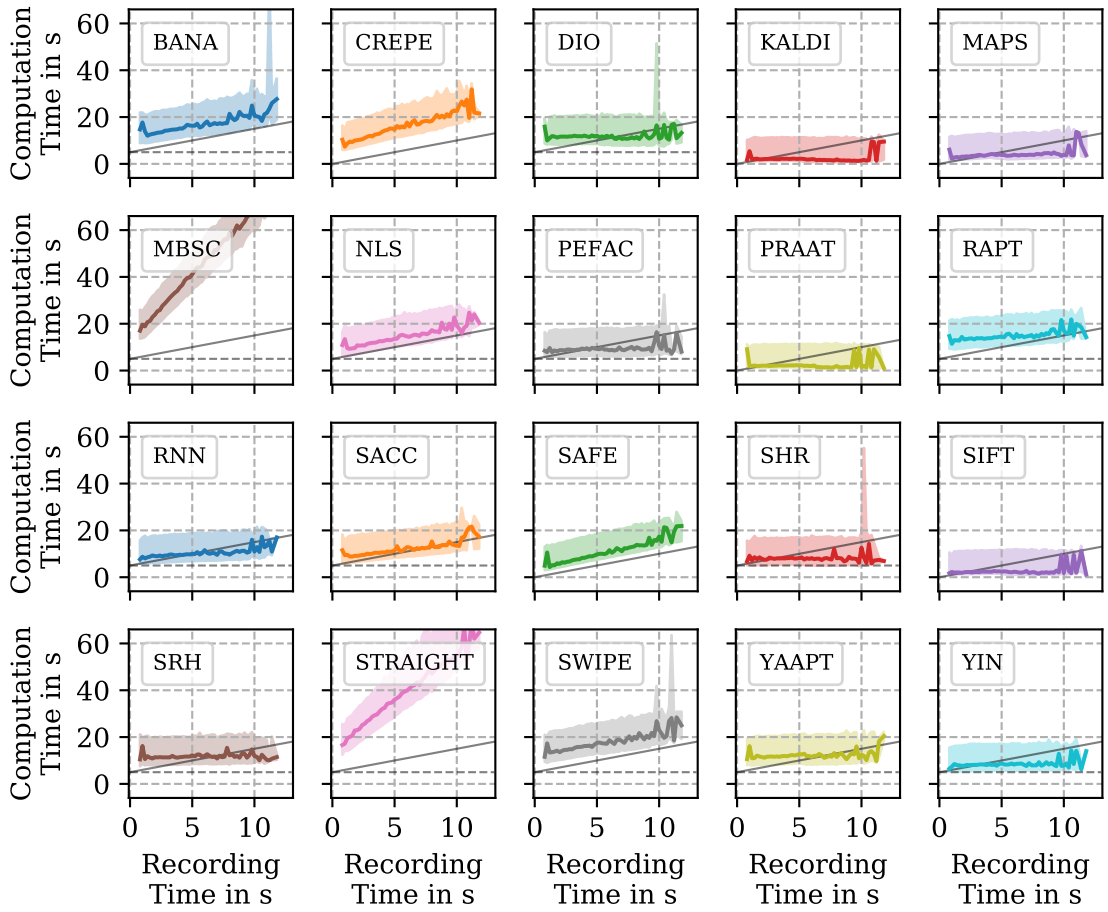


Figure 11.5: Time it takes to calculate the fundamental frequency for audio recordings of various lengths. Solid colored lines marks median, and shaded areas the 5 % and 95 % percentile of calculation times. All times include startup and shutdown times of their programming environment. If the PDA was implemented in Matlab, an additional grid line marks 5 s as the approximate Matlab startup time. A solid diagonal line marks real time (1 s per 1 s). All times in seconds.

too much computation time or memory to be practical. It should be noted that all included PDAs scaled linearly with recording duration, and merely differ in the scaling constant. Thus, the scaling categories above are approximations only useful for current, desktop-class computers as of 2019, and PDAs can change category on faster or slower computers.

11.7 Replication of Publications

One purpose of the replication dataset is to be able to replicate the results from existing publications. However, as figure 11.1 showed, comparable comparison studies are rare, so perfect replication is unlikely.

Figure 11.6 shows the same graph as figure 11.1 before but having been calculated from the replication dataset. *AUTOC*, *CEP*, and *SWIPE* were excluded from this graph, for reasons that will become apparent later. Even though some trends are replicated, the majority of results is very different from the literature.

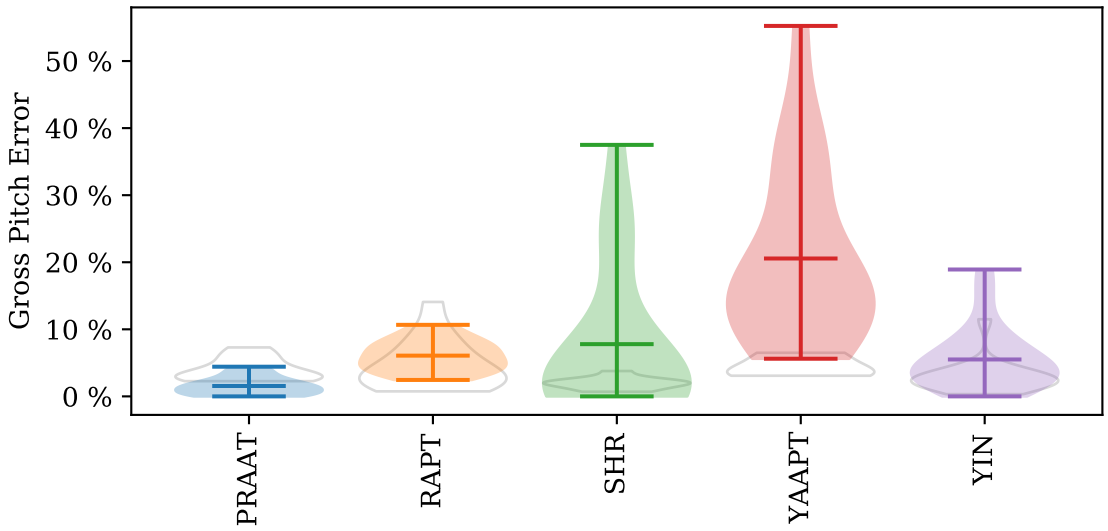


Figure 11.6: Partial recreation of the literature quotes from Figure 11.1 using the replication dataset. Grey outlines show literature quotes from Figure 11.1. Uses 20 dB SNR of white noise instead of clean recordings, as the replication dataset does not include truly clean recordings.

These differences are likely not failures of the individual publication or the replication dataset. Instead, they are simply differences in implementations of PDAs, signals, and performance measures. This undoubtedly highlights the need for standardization of these parameters, without which no meaningful conclusions can be drawn from comparing evaluation results of different publications.

Additionally, we looked at the results from two recent comparison studies, and recreated their results using the replication dataset. The first study that was replicated was [148], which used the *FDA* and *KEELE* corpus in some *NOISEX* noises to compare *AMDF*, *AUTOC*, *BANA*, *CEP*, *MBSC*, *PEFAC*, *SWIPE*, *YAAPT*, and *YIN* in terms of GPE. Figure 11.7 recreates graphs 1–6 from [148]. In both graphs, results for *BANA*, *YIN*, and *YAAPT* are similar to the publication, *CEP* and *PEFAC* are at least of similar shape in the same order of magnitude; but *AMDF*, *AUTOC*, *SWIPE*, and *MBSC* are very dissimilar. Incidentally, the latter category is also the worst performers in [148], which might imply an implementation issue in the publication.

The second study that was replicated was [146], which compared *RAPT*, *YIN*, and *PRAAT* on clean recordings of the *PTDB-TUG* corpus. Their results, as well as the results from the replication dataset, are shown in Table 11.1. Results, however, are only comparable for *RAPT*.

Table 11.1: Recreation of Table 2 from [146], as well as results from the replication dataset.

PDA	GPE [146]	GPE repl.	FPE [146]	FPE repl.
PRAAT	2.09	4.86	1.97	3.94
RAPT	4.67	4.20	2.63	2.00
YIN	1.39	11.39	1.86	3.53

Neither of the two comparison studies rigorously defined the preparation of their audio signals or the calculation of their error measures. This, again, highlights the need for the replication database that includes both pre-computed, verifiable fundamental frequency estimations of various audio signals,

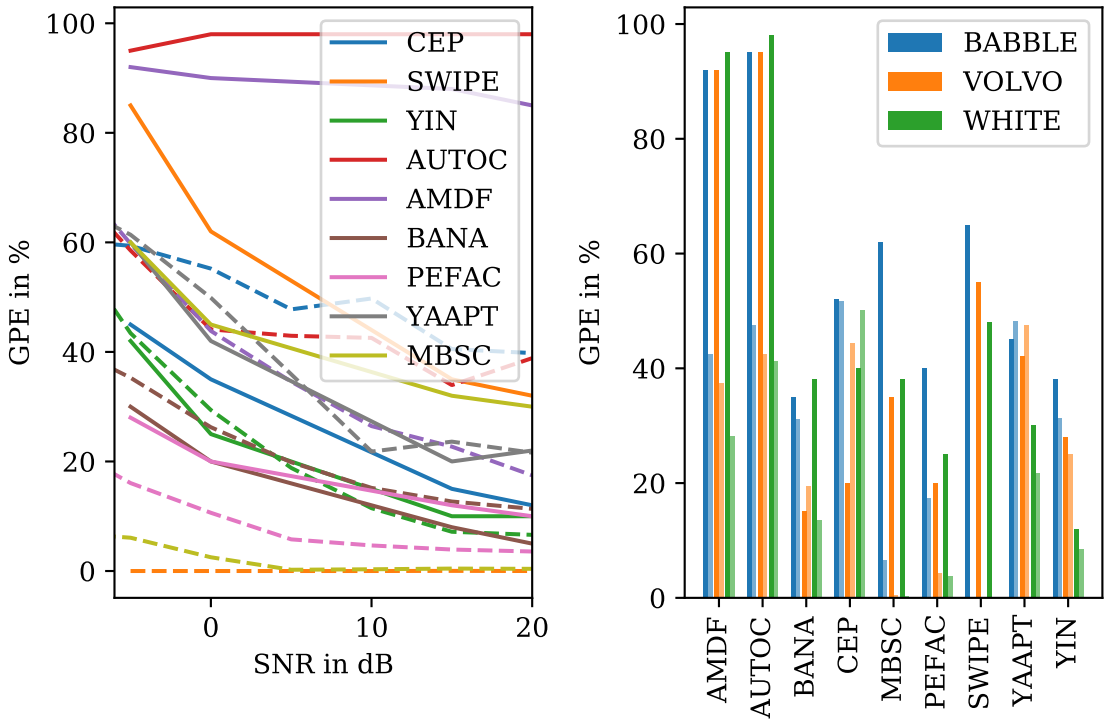


Figure 11.7: Partial recreation of comparison studies' results using the replication dataset. The left graph recreates Figures 1-3 from [148] in solid lines, with replication dataset results as dashed lines. The right graph recreates Figures 4-6 from [148] in the left, dark bars, with replication dataset results as the adjacent, lighter bars.

as well as verifiable, pre-computed performance measures. Without such a reference, cross-publication comparisons seem essentially impossible, as these two studies and Figure 11.1 illustrated.

11.8 Conclusions

As both the comparison of evaluations in Figure 11.1 and the comparison between comparisons in Figure 11.7 and Table 11.1 showed, there is very little consensus with respect to performance measure implementations, PDA implementations, and preprocessing methods. At least, however, speech and noise corpora and performance measure descriptions were widely compatible.

It is unlikely that yet another comparison study such as the present one would change this deplorable situation. Yet, a common standard is sorely needed. Otherwise, ever more sophisticated PDAs would require ever more complex comparisons as well, which could quickly become impractical computationally, as the complexities of data and software involved compound.

On the topic of computational complexity, a number of PDAs could not be included in our study because their code could not be made to run on our compute cluster due to performance optimizations. For scientific code, simplicity and explanatory power are still far more important than computational considerations.

The replication dataset represents not only our attempt at providing the largest comparison dataset yet, but more importantly to serve as a common ground from which future comparisons may benefit. In particular, our rigorous definitions of error measures and signal conditions, as well as our published source code, should enable a truly reproducible evaluation environment, which has so far been absent

from publications.

Ideally, future comparisons would not need to re-calculate the entire comparison dataset but could instead refer to the performance measures in the replication dataset that have already been calculated. To ensure compatible results for their own PDAs, future researchers could rely on the open and accessible replication source code for adding their own comparisons as well.

While this scenario might be far-fetched, the replication dataset is nevertheless the most complete and expansive dataset of speech and noise corpora and PDAs yet created, and will serve at least as a robust foundation for an in-depth comparison of PDAs and databases in the next chapter.

Chapter 12

A Comparison of Methods

Abstract

A plethora of algorithms have been proposed for estimating the fundamental frequency of voiced speech. These pitch determination algorithms (PDAs) and their estimations are typically validated with acoustic speech and noise recordings from one of a number of speech corpora and acoustic noise databases. While a number of comparison studies between PDAs have been undertaken, few have tried to quantify differences between corpora, or ascertained the suitability of various algorithms for different voices. Such deeper comparisons are difficult not only scientifically, but also in terms of engineering due to the large amount of data required to yield meaningful results. Undeterred, this chapter evaluates a large number of fundamental frequency estimation algorithms, speech and noise databases, and performance measures. As a result, algorithms are characterized in unprecedented detail, which reveals hitherto unknown biases and limitations for all investigated PDAs.

12.1 Introduction

The analysis of human speech and its pitch is a vibrant area of research, and instrumental to a wide variety of applications from speech recognition and transcription, speaker or language identification, to more abstract technologies such as speech compression and transmission. Depending on their specific use case, PDAs are optimized for different data sets and applications and vary in behavior and complexity. However, after publication, they tend to be used with audio recordings that were not part of the training data set and/or in circumstances that might deviate from the authors' intents.

It is thus of great interest to compare PDA accuracy and evaluate their suitability for various speech analysis tasks. In order to evaluate the accuracy of PDAs, a large and diverse data set is required that spans both, the original training data and "realistic" unseen data. PDA estimates must then be compared against a known ground truth. Chapter 9 detailed a number of well-known speech databases for pitch determination, some of them with a dedicated fundamental frequency ground truth. Chapter 10 furthermore introduced a new consensus truth that is better optimized for evaluating PDA accuracy, and available for more databases. Table 12.1 summarizes these speech databases, and Table 12.2 summarizes the noise corpora used in this study.

All of these corpora contain recordings of short sentences, except for *KEELE*, which includes multiple sentences per recording. To make the *KEELE* corpus more comparable, this chapter will henceforth always use *KEELE*-mod, introduced in Chapter 9.4, which is the same audio data, cut into shorter segments.

Most PDAs include some kind of implicit or explicit training for a specific source of truth and should be expected to perform best if evaluated against similar ground truths. These might be a

Table 12.1: Common speech corpora for fundamental frequency estimation.

Corpus	Samples	Speakers	Audio	Speech	f_0	Lar.	f_s	License
FDA [5, 4]	100	2	0:06 h	0:04 h	✓	✓	20 kHz	N/A
KEELE [122]	10	10	0:06 h	0:04 h	✓	✓	20 kHz	NC†
PTDB-TUG [119]	4718	20	9:36 h	3:28 h	✓	✓	48 kHz	ODBL‡
TIMIT [40]	6300	630	5:23 h	4:00 h			16 kHz	nonfree
CMU-ARCTIC [80]	15603	18	13:53 h	10:35 h		✓	16 kHz	OSS*
MOCHA-TIMIT [168]	4028	2	4:38 h	2:28 h		✓	16 kHz	NC†

† free for noncommercial use

‡ `https://opendatacommons.org/licenses/odbl/1.0/`

* BSD-style free software

Table 12.2: Common acoustic noise databases for fundamental frequency estimation. If significant, includes the number of references in comparison studies between 2015 and 2019.

Corpus	Samples	Audio	Samplerate	ref	License
QUT-NOISE [26]	20	13:39 h	48 kHz		CC-BY-SA
NOISEX [155]	15	0:16 h	20 kHz	10	N/A†

† The database is no longer available online. However, it is based on the RSG.10 database, which is available at `http://www.steeneken.nl/7-noise-data-base/`.

laryngograph-derived fundamental frequency ground truth included in some speech corpora, or a reference PDA that was used in training. Table 12.3 lists the PDAs used in this study, as well as the databases used for training them.

The comparisons in this chapter were carried out on the replication dataset detailed in Chapter 11, which also includes a full description of the performance measures used for comparison. While this chapter shows mostly comparisons between PDAs, the appendix on page 185 includes PDA profiles that summarizes each PDA on its own, which can be used as a quick reference while reading this chapter.

The rest of this chapter is organized as follows: Section 12.2 analyzes the data from the replication dataset and compares the PDAs' performance measures and resulting characteristics. This includes a large number of subsections for various performance metrics and statistical analyses. Finally, Section 12.3 concludes the paper with a general summary of the findings.

12.2 Evaluation

The replication dataset contains enough variety to assess the qualities and behaviors of the various PDAs in many kinds of different signal conditions. The following sections will look at different aspects of this by grouping the replication dataset by various parameters and averaging their performance measures within each group to show how PDA performance is affected by the parameter.

Additionally, some evaluations require new performance metrics that are not part of the replication dataset's performance measures. These were calculated on the raw estimations included in the replication dataset, and again grouped and averaged for presentation.

We generally attempted to keep the positions and colors of the PDA constant across graphs as much as possible. For the same reason, the 25 PDAs in the replication dataset were pared down

Table 12.3: PDAs used for comparison, and the corpora used by the original authors in training or evaluation.

PDA	training corpora
AMDF [130]	
AUTOC [140]	
BANA [56]	*NOISEX*
CEP [105]	
CREPE [79]	
DIO [100]	*FDA*
DNN [50]	*NOISEX*, *TIMIT*
KALDI [42]	*KEELE*
MBSC [151]	*FDA*, *KEELE*, *NOISEX*
MAPS (chapter 8)	*PTDB-TUG*, *QUT-NOISE*
NLS [104]	
PEFAC [45]	*NOISEX*, *TIMIT*
PRAAT [9]	
RAPT [150]	
SAFE [22]	*FDA*, *KEELE*, *NOISEX*
SACC [86]	*FDA*, *KEELE*
SHR [149]	*FDA*, *KEELE*
SIFT [93]	
SRH [31]	*FDA*, *KEELE*, *NOISEX*
STRAIGHT [73]	
SWIPE [18]	*KEELE*, *FDA*
YAAPT [173]	*KEELE*, *FDA*
YIN [24]	*KEELE*, *FDA*

to merely 20 PDAs, so as to ensure a reasonable graph size of a 4-by-5 grid. The removed PDAs were *CEP*, *AUTOC*, and *AMDF*, which were found to perform very poorly in noise, and therefore expendable in the comparison. Of the two versions of *NLS*, the later iteration was found to be a large improvement and included instead of the earlier one. Of the pair of *DNN*/*RNN*, *DNN* was excluded in favor of *RNN*, as they were found to be all but identical. These PDAs are included in the PDA profiles in the appendix, however.

All PDAs were used with their default parameter values. There is no doubt that many of them could be adjusted to better fit particular situations through their parameters, but this would exceed the scope of even this large comparison study. For the same reason, our own algorithm, *MAPS*, is included with the pitch confidence decision matrix from Chapter 8.2.3 instead of a newly trained matrix for the greater variety of speech amd noise signals used in this comparison.

As each PDA was developed for a specific use case, and a specific signal model, their behavior varies greatly under different signal conditions. Many of these differences only became visible in a large and diverse data set such as the replication dataset. In total, this evaluation is a summary of about one million experiments, consuming a total calculation time of roughly a full year on a single-core computer from 2019. To the best of our knowledge, this is the largest and most complete comparison study of PDAs and corpora yet conducted.

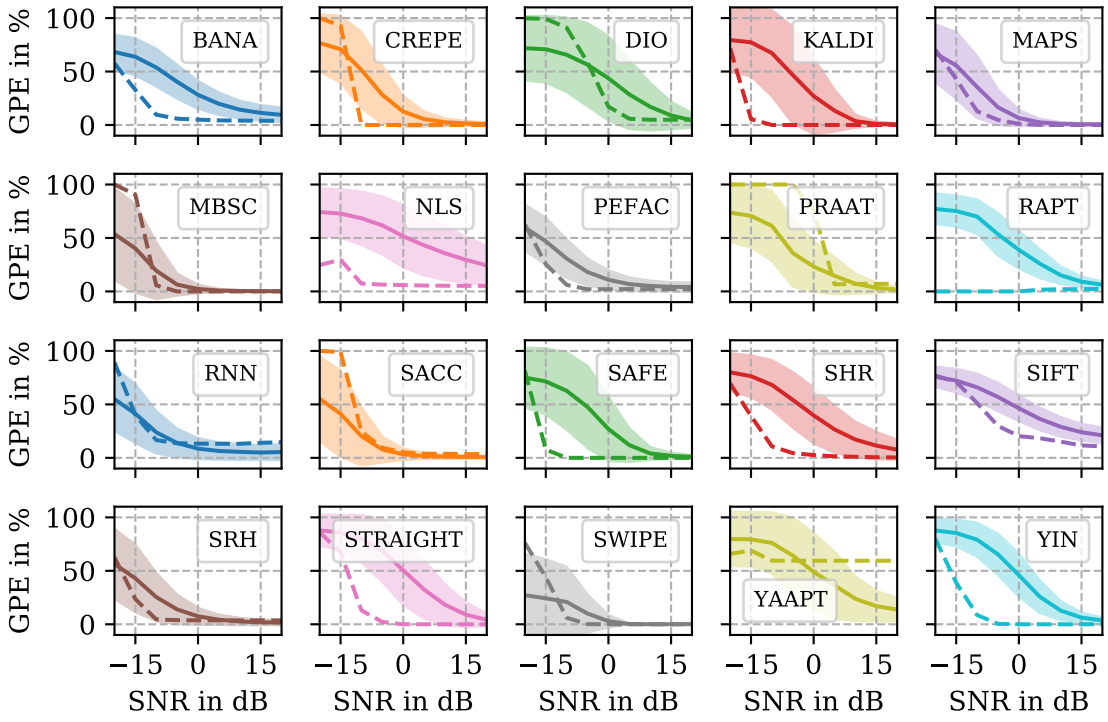


Figure 12.1: GPE vs. SNR of all PDAs from the entire realistic data set for the consensus truth with mean GPE as a solid line, MAD as shaded area, and synthetic baseline as dashed line. Only for frames that are voiced according to the ground truth and the PDA.

Estimation Accuracy (GPE)

Figure 12.1 shows the most common measure for fundamental frequency estimation accuracy: the gross pitch error for various SNRs. In general, PDAs are most accurate if the speech signal is not corrupted by noise and SNRs are high. Towards lower SNRs, more and more regions of the speech signal are corrupted and estimation accuracy suffers.

Most PDAs show a characteristic "knee" somewhere around 0 dB SNR, where noise corruption becomes significant, and estimation accuracy begins to deteriorate. At higher SNRs than this transition area, error rates are typically constant and near zero. On the lower-SNR side, error rates rise. In the following text, frequent reference will be made to this transition area, as many differences between PDAs are most obvious in this area. "Positive SNRs" and "negative SNRs" will frequently be used as a shorthand for areas before or after this transition area.

From GPE curves, we can derive further performance criteria, such as minimum error rates at positive SNRs and the steepness and threshold of the error slope in the transition area. A minimum error rate close to zero at positive SNRs implies an *unbiased* estimator. A *noise robust* estimator can retain its peak performance up to low SNRs, and has a late and shallow transition slope. Depending on the application, error rates below 10-30 % are usually deemed usable. Based on these considerations, *CREPE*, *KALDI*, *MAPS*, *MBSC*, *PRAAT*, *SACC*, *SAFE*, *SRH*, *SWIPE*, and *YIN* are unbiased at positive SNRs, and *CREPE*, *MAPS*, *MBSC*, *PEFAC*, *RNN*, *SACC*, *SRH*, and *SWIPE* are particularly noise-robust.

However, the gross pitch error is a joint evaluation of the PDAs' voice activity determination and estimation accuracy, as only *voiced* frames are taken into account in its calculation. For the purposes

of GPEs, a frame is considered *voiced* if both the ground truth and the PDA label it thus. Therefore, low error rates mean both precise pitch estimates, as well as accurately discarding unusable frames, particularly at negative SNRs, where many frames are irretrievably masked by noise. The trade-off between recalling enough voiced frames and selecting precisely only voiced frames will be evaluated in more detail later.

Beyond the mean GPE, the shaded areas in Figure 12.1 around the mean visualize the mean absolute deviation (MAD)[1] of GPE per trial across the entire data set at the given SNR and illustrates the variability of the estimates across signal conditions. High MAD indicate that the estimation accuracy depends strongly on the kind of signal and noise. For many applications, a low MAD at positive SNRs is highly desirable, as found in *CREPE*, *KALDI*, *MAPS*, *MBSC*, *SACC*, *SAFE*, *SRH*, *SWIPE*, and *YIN*. This will be investigated further later with explicit differences between corpora and noise signals.

The dashed line in Figure 12.1 shows the mean GPE for the synthetic experiment, with modulated harmonic tone complexes in white noise. Most PDAs use them as their internal signal model, and produce very high accuracies for these kinds of signals. Since white noise masks every frequency equally, the drop-off from total accuracy to no estimation is usually very steep, at the precise point where the noise overwhelms the signal in the PDA's feature set and usually does not correspond exactly to the transition area in the realistic data set. If the synthetic accuracy is worse than the realistic accuracy, the algorithm likely does not use harmonic tone complexes as its signal model. This is certainly true for *MBSC* and *SACC*, which both employ autocorrelation on band-pass filters that might expect more temporal structure than is provided by sinusoids. *SWIPE* also shows this behavior, however, perhaps due to its stronger-than-usual subharmonic suppression.

In this first evaluation, algorithms with near-zero error rates at positive SNRs, and transitions around 0 dB SNR can be considered highly accurate, such as *CREPE*, *KALDI*, *MAPS*, *MBSC*, *PEFAC*, *RNN*, *SACC*, *SAFE*, *SRH*, and *SWIPE*. It should be noted however, that these criteria only apply to applications with high levels of noise. For example, *PRAAT* and *RAPT* and *YIN* might be perfectly adequate if little noise is expected.

Estimation Precision (FPE)

While the gross pitch error in the previous section describes the *accuracy* of a PDA, a measure for how close its estimates are to the truth, this section's fine pitch error is a measure for a PDA's *precision*, or random variability. Figure 12.2 shows the FPE of all PDAs against SNR.

The fine pitch error offers a measure of how out-of-tune the average grossly correct estimates of a PDA are in a musical sense. However, prosodic variations in pitch are rather coarse, and there is no specific meaning ascribed to the precise pitch of speech in European languages. The FPE should therefore be interpreted with some reservations, as its significance is limited, at least to Western languages. Nevertheless, its importance to other applications such as musical tune, as well as the analysis of some tonal languages, is without question.

For this reason, most PDAs estimate pitch on a rather coarse scale, both in time and frequency[2]. Some PDAs improve precision with a parabolic interpolation stage of frequencies after the main estimator. No such sharpening measures are typically employed in the time domain, however, beyond the implicit time interpolation of center-weighted window functions. Accordingly, neither PDAs nor the ground truths in our speech corpora can be assumed to be optimized for minimizing fine pitch errors.

[1] the mean absolute deviation around the mean, or $\frac{1}{N}\sum_{n=1}^{N}|x_n - \bar{x}|$ is similar to the standard deviation, but easier to interpret in terms of the data coordinates, as it does not include squaring. The MAD is a measure of dispersion around the mean, thus GPE+MAD may exceed 100 %.

[2] at least by default. Some PDAs include the resolution of their frequency grid as an adjustable parameter.

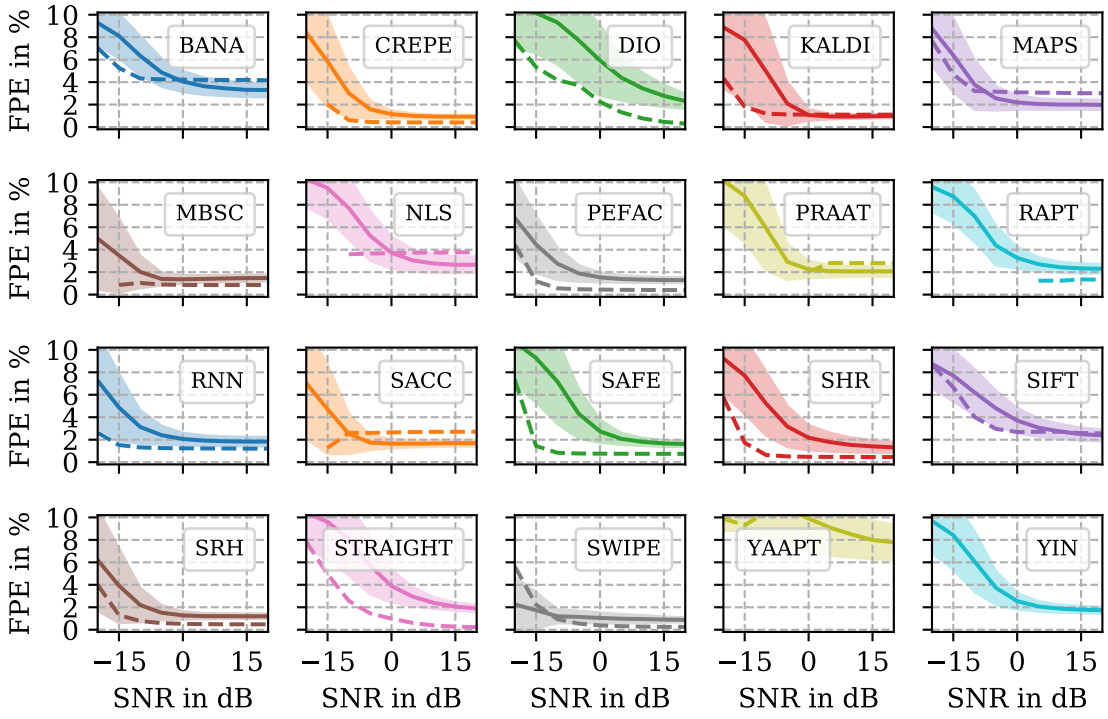


Figure 12.2: FPE vs. SNR of all PDAs from the entire realistic data set for the consensus truth with mean FPE as a solid line, MAD as shaded area, and synthetic baseline as dashed line. Only for frames that are voiced according to the ground truth and the PDA.

It is no surprise then that Figure 12.2 shows significant differences between PDAs. In general, FPEs follow a similar shape as GPEs, and rise from a stable minimum at positive SNRs to a maximum at negative SNRs. However, neither is a near-zero GPE any indication for a near-zero FPE, nor are the transition areas necessarily similar. As FPEs are the mean absolute deviation of grossly-correct frames, pitch estimates are guaranteed to lie between ± 20 % of the true pitch, and entirely-random estimates would therefore result in 10 % FPE.

In contrast to GPEs, where most PDAs clearly trend towards zero, FPEs typically reach a minimum well above zero, with *CREPE*, *KALDI*, *MBSC*, *PEFAC*, *SHR*, *SRH*, and *SWIPE* below 2 %; *DIO*, *MAPS*, *NLS*, *PRAAT*, *RAPT*, *RNN*, *SACC*, *SAFE*, *SIFT*, *STRAIGHT*, and *YIN* around 2 %; and *BANA* and *YAAPT* between 2–10 %. The one PDA optimized specifically for singing voices, *CREPE*, achieves one of the lowest FPE, which is no doubt desirable in that particular application[3].

Synthetic results similarly trend to a stable minimum above zero. For many PDAs, this minimum is somewhat lower than the minimum for realistic speech recordings, and transitions at lower SNR. Three PDAs in particular show the opposite however: *MAPS*, *PRAAT*, and *SACC*. This might be caused by the unrealistically broad fundamental frequency distribution in the synthetic dataset, which are uniformly distributed between 80–240 Hz (plus modulation), whereas human speech pitches cluster around 120 Hz and 180 Hz, according to Figure 10.2.

For most PDAs, the transition area of FPEs starts at lower SNRs than for GPEs. This is of particular interest, as the first effects of rising noise seem to be occasional gross errors, but the precision of the remaining estimates is unaffected at first. Perhaps this is an indication that these

[3] *DIO* also references singing voices, but its FPE minimum is not within the ± 20 dB SNR plotted here, and therefore couldn't be judged.

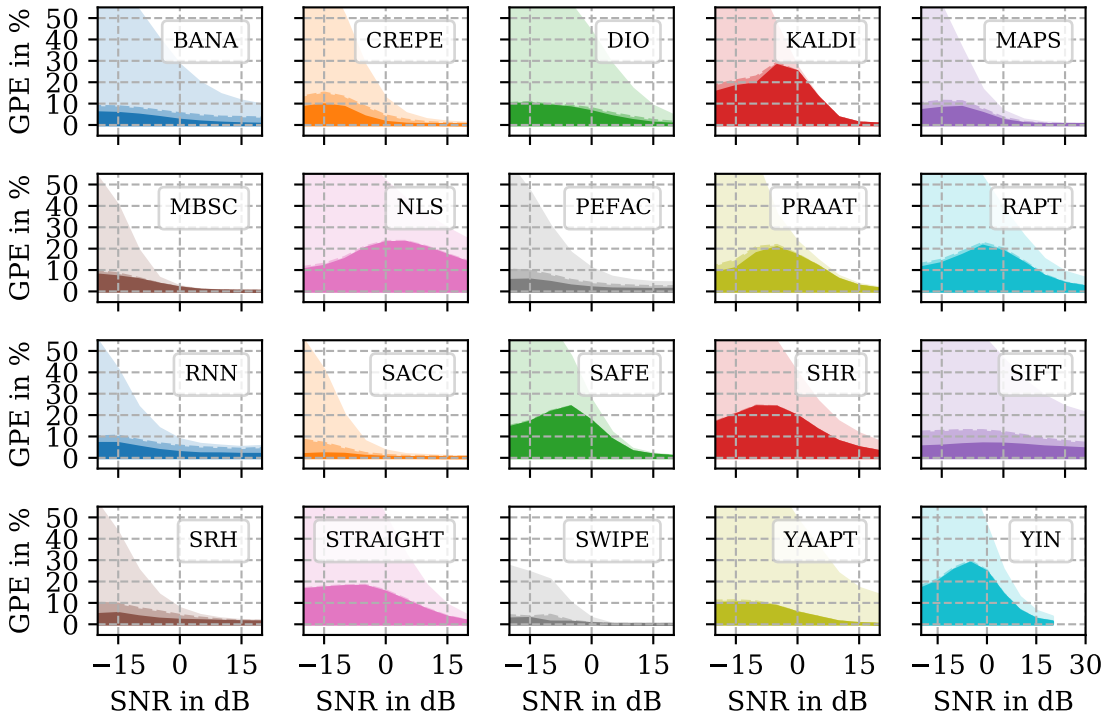


Figure 12.3: Octave pitch errors vs. SNR from all PDAs on the entire realistic data set for the consensus truth. The shaded area are mean GPEs, as in Figure 12.1. The dark areas mark GPEs that are low octave errors, and the middle areas mark GPEs that are high octave errors. Only for frames that are voiced according to the ground truth and the PDA.

PDAs' VADs are more heavily influenced by noise than their pitch estimator, which results in unestimable VAD false positives, while true positives remain clearly discernible. This effect is strongest for *KALDI*, *NLS*, *PRAAT*, *RAPT*, *SHR*, and *YIN*. Only *BANA* and *YAAPT* do not seem to follow this trend at all.

In summary, it must once again be stressed that the importance of fine pitch errors is entirely defined by the intended application. Speech synthesis and transmission might afford significantly larger leeway in FPEs than musical analyses or automatic speech recognition of tonal languages. Within the crop of PDAs in this study, however, the bias is probably towards the former, with very little significance put on pitch estimation precision beyond its usefulness to prosodic characterization of European languages.

Octave Pitch Errors

Since voiced speech produces a repetitive structure in both the time and frequency domain, PDAs are susceptible to octave errors, where an integer multiple of the repetition period is mistaken for the true period. Figure 12.3 shows GPEs that are octave errors, as part of the GPEs shown in Figure 12.1.

Octave errors are particularly frequent in the transitory region where GPEs just start to rise. In this region, speech features are still detectable, but may be partly obscured by noise. Since many noises are low-pass noises, they mask the fundamental frequency more strongly than its harmonics, making octave errors more likely. In many cases, most or all GPEs in the transitory region are octave errors. Particularly, *KALDI*, *NLS*, *PRAAT*, *RAPT*, *SAFE*, and *SHR* GPEs are dominated by low

octave errors up to 0 dB SNR, and only start to diverge at negative SNR, whereas *BANA*, *CREPE*, *DIO*, *MAPS*, *MBSC*, *PEFAC*, *RNN*, *SACC*, *SIFT*, *SRH*, and *SWIPE* only show a minor contribution of octave errors to total GPE.

Most octave errors were low octave errors. Comparing the dark shaded area with the medium shade shows high octave errors to be an exceedingly rare occurrence. Only few algorithms, such as *BANA*, *CREPE*, *DNN*, and *SIFT* showed high octave errors at positive SNRs at all. For all other PDAs, high octave errors were insignificant, and only occured at negative SNRs where they were probably a random occurrence. The exception was *SACC*, which indeed had more high octave errors than low ones, perhaps owing to its bandpass filters being unaffected by out-of-band noise corruption and no octave error suppression system in the machine learning stage.

At negative SNRs, speech features are masked entirely, and most pitch determination becomes impossible. Thus octave errors, too, became less frequent below the transitory region. At this point, most GPEs are likely pitch estimates of speech-like structures in the background noise or of non-voiced speech segments. The latter can be caused from VADs that classify segments as voiced where no fundamental frequency can be determined. This is particularly probable for PDAs where different measures were used for VAD and pitch determination. PDAs with a joint VAD and pitch estimation, such as *SACC*, *MAPS*, and *RNN*, likely derive some of their low GPEs in negative SNRs from not making these kinds of errors.

It is often in the error cases where an algorithm reveals its internal biases, which are otherwise hidden between accurate estimates. GPEs that are not octave errors are shown in Figure 12.4. The figure shows all PDAs' errors centered around the true pitch. While the remaining GPEs of *BANA*, *CREPE*, *MAPS*, and *RNN* were roughly symmetrical around the true pitch, other PDAs, particularly *DIO*, *KALDI*, *NLS*, *PRAAT*, *RAPT*, *SHR*, and *YIN*, showed a bias towards higher or lower frequencies. Such a bias can be caused by a general preference for under- or overestimating pitches, or in some cases by a default pitch that is assumed if no viable estimate can be determined.

Interestingly, some PDAs showed essentially a constant bias over all SNRs, while others varied with SNR, indicating that biases change with noise levels. For example, *PEFAC*, *RAPT*, and *SAFE*, tended to under-estimate low-SNR pitches, but over-estimate cleaner signals, in gross estimation errors. This might indicate that lower SNRs give rise to new kinds of errors that were not visible at higher SNRs, perhaps noisy-but-voiced frames, whereas higher SNRs were dominated by purely voiced VAD false positives.

VAD Errors

Many PDAs feature their own VAD that labels frames *voiced* if there is significant voicing activity, or *unvoiced* otherwise. It should be noted that VAD in the context of fundamental frequency estimation refers specifically to *voiced* speech, whereas other areas of research occasionally use the same term for any kind of speech activity.

Each VAD must balance false positives and false negatives, or recall and precision. Depending on the intended application, precision, recall, or a balance thereof might be preferable. For example, a speech analysis application might require *precise* estimates as a basis for formant estimation or speaker identification, which are rarely wrong, but exclude ambiguous cases. These applications can often tolerate missing data but not incorrect estimates. In contrast, speech recognition might require good *recall* to capture every syllable fully and might discard incorrect estimates on its own. Figure 12.5 shows the PDAs' false positive and false negative rates across SNRs. Note that these kinds of errors were invisible in the GPE graphs, as GPEs only look at true positives.

The graph shows that for most PDAs, SNR mostly affected false negative rates, while false positives remained relatively constant. This should be expected, as rising noise levels necessarily cast more and more frames as noisy, while fewer frames remain intelligible for pitch determination.

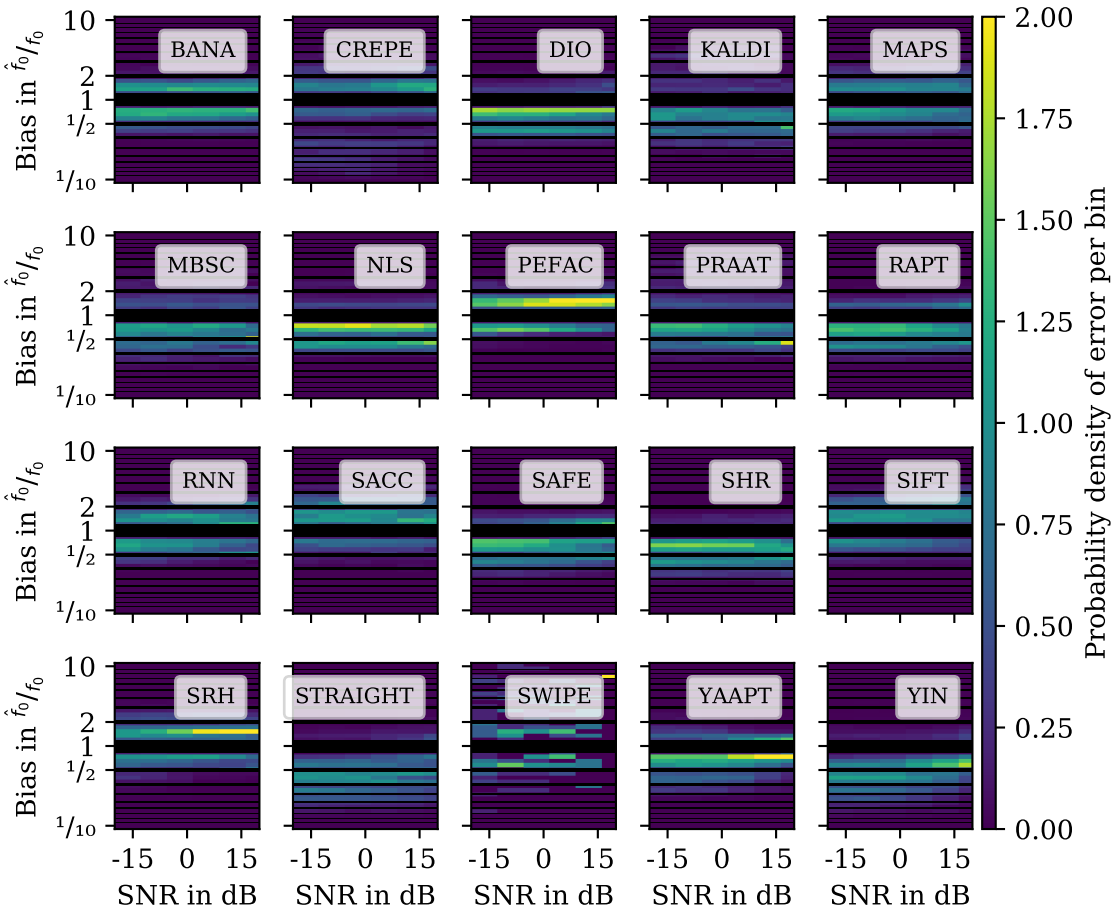


Figure 12.4: Density histogram of remaining estimation bias of GPE errors vs. SNR of all PDAs from the entire realistic data set for the consensus truth. Remaining bias is the frequency factor between the estimated pitch and the true pitch for gross pitch errors that are not octave errors. Only for frames that are voiced according to the ground truth and the PDA.

Interestingly, false negative rates do not seem to be influenced by the transition area, but instead rose steadily with SNR without a central pivot point around 0 dB. VAD performance thus showed a reduction in recall, even in the area of constant, near-zero GPEs. This implies that most PDAs' VADs are less robust to noise than their pitch estimators, and that real-word estimation performance degrades with SNR even if GPEs do not.

In general, the PDAs' curves are mostly parallel, and show little overlap, in Figure 12.5. This suggests that the VADs differ mainly in their threshold, and not in their quality, particularly at low SNRs. The choice of threshold is then a design decision depending on the intended application, with *PEFAC* and *STRAIGHT* being particularly high-recall, and *KALDI*, *MBSC*, and *SWIPE* very precise.

It might seem strange that the results for *MAPS* are very different from Figure 8.11 in Chapter 8. Indeed, *MAPS*' training on *PTDB-TUG* made its VAD particularly precise on that dataset. No similar tendency could be found in any other PDA, however, perhaps highlighting the unique nature of *MAPS*' joint VAD and pitch estimator.

If the application is able to deal with missing data, but not incorrect estimates, low false positive rates and high precision are required. If it is more important to catch all pitched data while tolerating some incorrect estimates, a low false negative rate and high recall are required. It is amusing to note

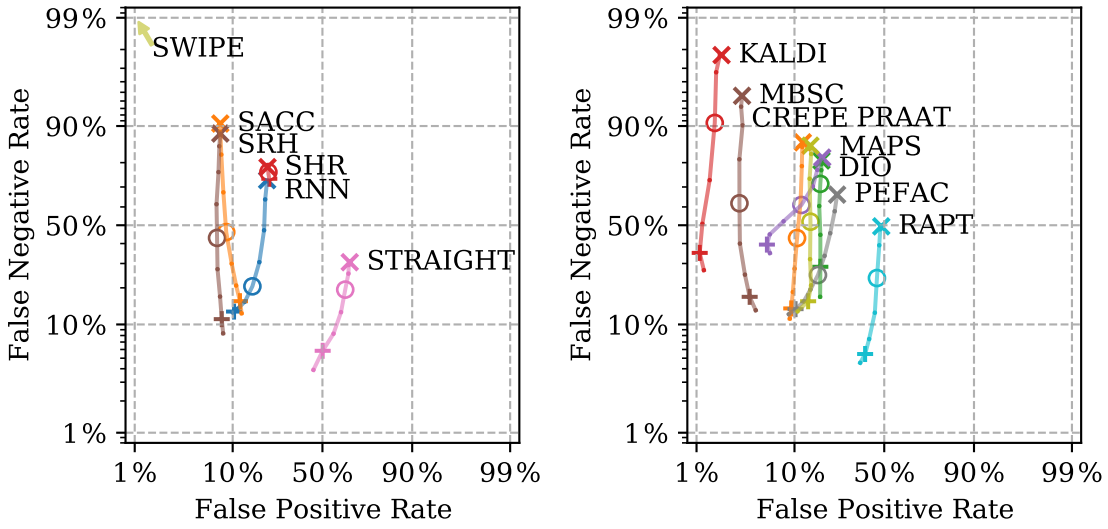


Figure 12.5: VAD false negatives vs. false positives of all PDA across SNRs from the entire realistic data set for the consensus truth. × marks -20 dB SNR, ○ marks 0 dB SNR, and + marks 20 dB SNR, with subtle dots every 5 dB. Only PDAs with a VAD are shown. Axes are logit warped.

that both of these approaches might be called *robust* in some circumstances, while meaning their exact opposite.

Error Summary

Thus far, PDAs were shown to produce gross pitch errors, VAD false positives, and VAD false negatives. Gross pitch errors are important as they indicate incorrect estimates. VAD false positives are ignored in GPEs, as there is no ground truth against which their estimates could be evaluated. In a practical application, however, they are essentially indistinguishable from GPEs, in that they are incorrect estimates. VAD false negatives are missing estimates, but at least they are not wrong. Missing data is generally easier to deal with than incorrect estimates, if enough estimates remain to gain an impression of the signal's fundamental frequency.

Figure 12.6 summarizes all of these error measures as an overview of the findings so far. At positive SNRs, a trade-off between false positives and false negatives is visible. Some PDAs, such as *KALDI*, *MAPS*, *MBSC*, and *SWIPE*, show negligible false positive rates and very low GPEs but relatively high false negative rates. Others trade lower false negatives for higher false positives and slightly higher GPEs, for example *CREPE*, *DIO*, *PEFAC*, *PRAAT*, *RAPT*, *RNN*, *SACC*, *SRH*, and *STRAIGHT*. As previously discussed, this trade-off depends on the intended application. It is unfortunate, however, that the all-important GPE measure is affected by false positives, but not false negatives, creating an incentive to reduce the former, and ignore the latter.

At negative SNRs, if false negatives rise too high, there might not be any frames left for evaluation, as shown by the zeroes in the graph. Naturally, this is particularly prevalent in PDAs with high false negatives, such as *DIO*, *KALDI*, *MAPS*, *MBSC*, *PRAAT*, *RAPT*, *SACC*, and *SWIPE*. Extreme examples of this were *KALDI*, *SACC*, and particularly *SWIPE*, where most recordings at negative SNRs yielded no estimates whatsoever. These cases are hard to evaluate and are probably frequently overlooked as they are not evident in GPEs or FPEs, the two most common error measures in comparison studies.

Taken together and applied to real-world applications, however, the estimation and detection errors

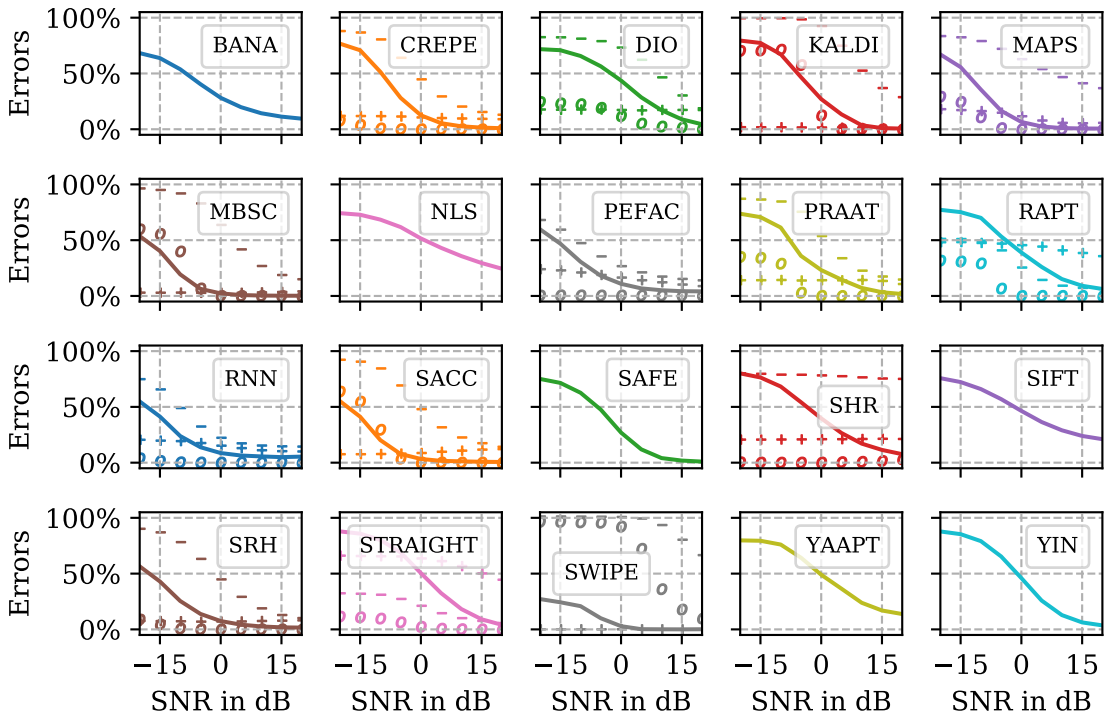


Figure 12.6: Error measures and missing data against SNR. The solid lines are GPEs, normalized to the number of voiced frames both in the estimator and the ground truth as done before. Zeros are the percentage of signals without any estimates, normalized to all signals. Pluses are the percentage of VAD false positive, relative to the number of negatives in the ground truth. Minuses are the percentage of VAD false negatives, relative to the number of positives in the ground truth. If the PDA had no VAD, no zeros, pluses, or minuses are drawn.

combine, and often leave very few accurate estimates at negative SNRs. Thus, even though GPEs imply otherwise, very few of these PDAs are suitable for negative SNRs beyond their transition area, and a better assessment of their accuracy might be the location of the transition area, as opposed to the remaining accuracy in terms of GPE or VAD errors.

In this sense, PDAs can be grouped in four categories: *BANA*, *NLS*, *SAFE*, *SIFT*, *YAAPT*, and *YIN* do not have a VAD but should work acceptably at high positive SNRs. *DIO*, *PRAAT*, *RAPT*, *STRAIGHT*, and *SWIPE* should work reliably down to about 15 dB SNR; *CREPE*, *MAPS*, *MBSC*, *RNN*, *SACC*, and *SRH* down to 0 dB SNR; and finally, *MBSC* and *SACC* possibly slightly below 0 dB. Below this threshold SNR, either GPEs or VAD errors rise extremely quickly and results should be expected to become useless at rapid rates.

VAD Dependence

At first glance, it seems estimation accuracy in terms of GPEs should not be influenced by the PDAs' VADs, as neither VAD false positives nor VAD false negatives were counted in the GPE calculation. However, highly *precise* VADs, which exclude numerous false negative frames from the GPE, could potentially improve accuracy by excluding ambiguous or difficult frames, whereas high *recall* VADs would include many false positives, thus lowering accuracy.

Figure 12.7 shows the PDAs' GPEs for *voiced* frames being determined by both the ground truth

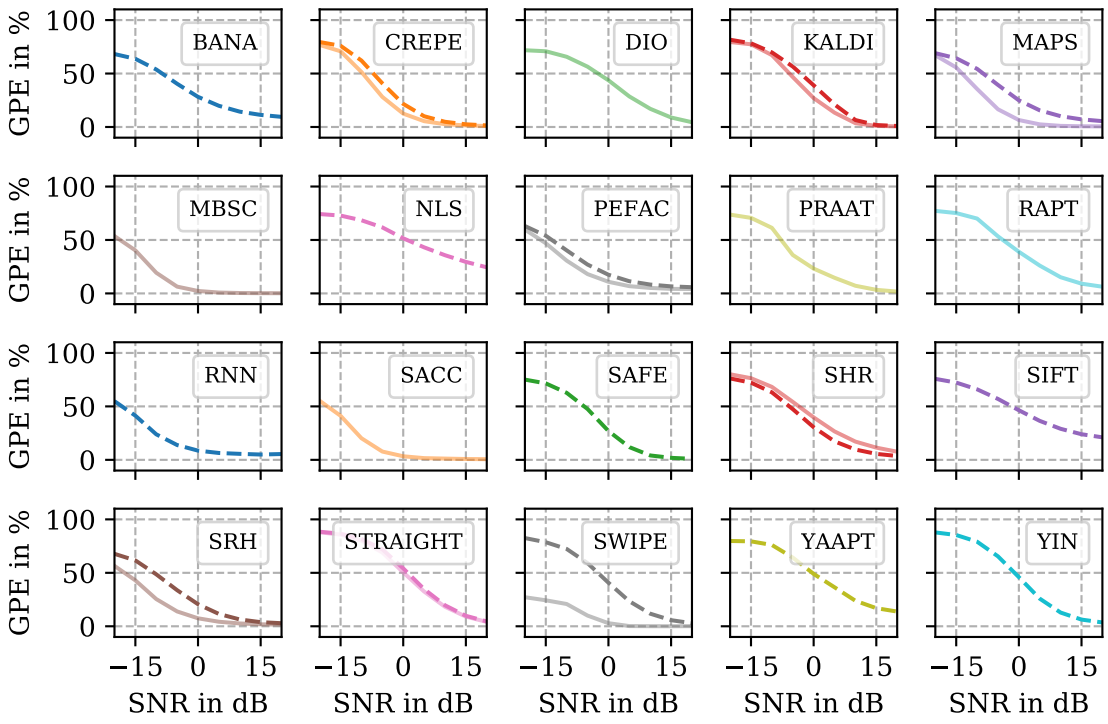


Figure 12.7: GPE vs. SNR of all PDAs from the realistic data set for the consensus truth. Solid lines consider *voiced* frames by both the ground truth and the PDAs' VADs ("mixed" condition), while dashed lines consider frames *voiced* only based on the ground truth ("true" condition). PDAs without VAD have neither mixed condition nor solid lines. PDAs that zero out their unvoiced frames have no true condition and no dashed lines.

and the PDAs' VADs (the default), or only by the ground truth. As expected, PDAs with a *precise* VAD according to Figure 12.5, such as *MAPS*, *SRH*, and *SWIPE*, tend to be more accurate using their own VAD than the ground truth. The effect is less strong for less precise PDAs. The opposite is true for *SHR*, where the ground truth VAD actually improves GPEs.

Thus, including a VAD improves GPEs for most PDAs by more rigorously selecting unambiguous estimates, thereby increasing both precision and accuracy. However, some PDAs choose to bake their VAD into the estimates and zero out all unvoiced pitches. This practice should be discouraged, as it restricts the use cases for the algorithm unnecessarily by forcing the PDA's VAD on every application, regardless of its suitability to the task. This is particularly important as VADs can be more signal-dependent than the often more rigorously defined pitch estimation procedures. Instead, it is preferable to include a pitch estimate and a VAD estimate, and to leave the masking of unvoiced frames to the user.

Significance

PDAs are generally built for a particular purpose, and have particular strengths and weaknesses. Yet, they are employed and evaluated with the same error measures, as if they could be used interchangeably. To test their similarity to one another, Figure 12.8 shows the mean calculated from multiple t-tests of pairs of PDAs' GPE scores. A dark color indicates that the GPE scores were not significantly different for many SNR conditions ($p > 0.05$), which is an indicator that these PDAs could be used

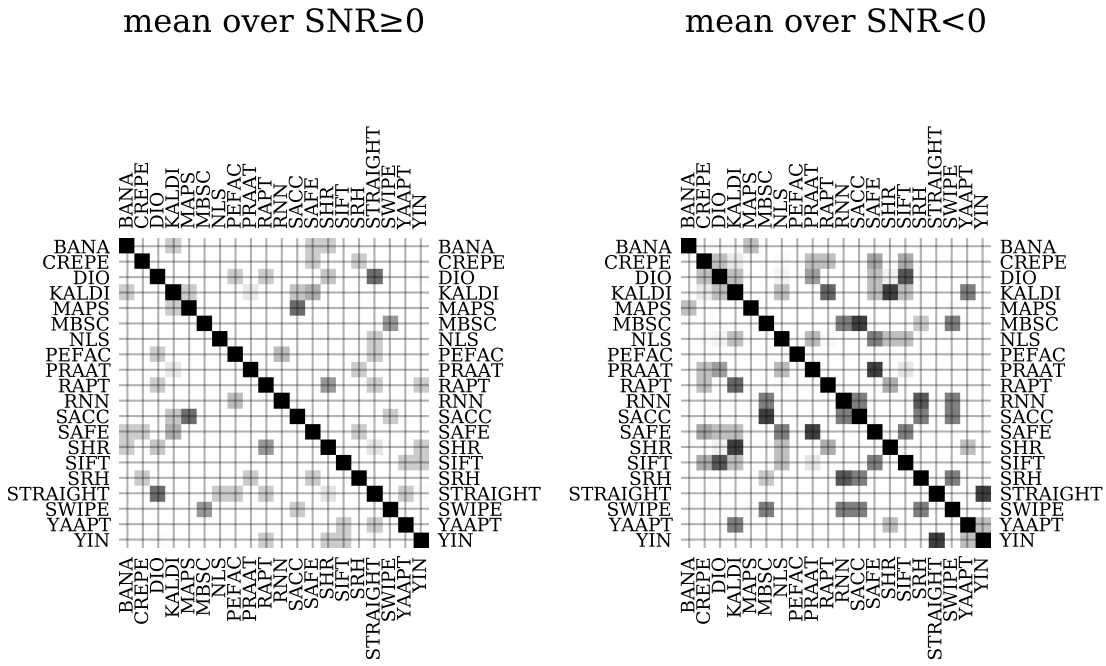


Figure 12.8: Significance of differences as the mean of multiple t-tests on the GPE scores of pairs of algorithms per SNR, averaged over all SNRs. White denotes all SNR conditions significantly different, black means no significantly different SNR conditions.

interchangeably according to the GPE.

The figure shows several groups of PDAs which are essentially indistinguishable from one another. In these cases, the numerical performance differences that remain between these PDAs should be disregarded as measurement noise. At positive SNRs, the differences are generally more significant than at lower SNRs.

At low SNRs, the GPE scores of some PDAs varied greatly with different signal and noise conditions, to the point where a simple mean GPE becomes meaningless. This is visible as a large number of interchangeable PDAs at negative SNRs, for e.g. *SAFE*, *NLS*, and *SIFT*.

At high SNRs, differences became more significant, with fewer interchangeable pairs of PDAs. Comparing significantly similar pairs to Figure 12.1, two distinct patterns emerge: Some algorithm pairs, such as *STRAIGHT* and *DIO*, or *RAPT* and *SHR*, attained similar GPE scores, but with a wide standard error. It is likely that the standard error contributes most strongly to their GPE averages being indistinguishable.

In the opposite case of *MBSC* and *SWIPE*, or *SACC* and *MAPS*, the GPE standard error dropped to zero quickly, but so did the GPE itself. These four PDAs achieve such small error rates at positive SNRs that their results become indistinguishable from GPE scores alone. At even higher SNRs (not shown), these cases of perfect estimation become more frequent, and therefore meaningless for comparison.

Interestingly, none of these highly similar pairings reflected much similarity in implementation. It thus seems that there is no clearly preferable way of doing fundamental frequency estimation in a particular domain or using a particular technique. Instead, even wildly different approaches seem to lead to indistinguishable results. Merely a weak assertion might be that they are all multi-domain

algorithms that make their estimates from both time- and frequency information.

This similarity between high-performance PDAs highlights how fundamental frequency estimation at positive SNRs is a well-studied discipline with plenty of close-to-perfect solutions available. New developments, therefore, should be careful to qualify their algorithms' performance for specific applications, such as high levels of noise, pathological voices, or specific kinds of noises. Merely striving for ever fewer errors at positive SNRs of calm speech in acoustic noise seems unlikely to result in significant improvements any longer.

Speech Corpus Dependence

Every PDA has its own signal model, and was developed with reference to a set of target signals. This includes explicit machine learning processes with training data sets, or it might be implicit, such as an example recording frequently used during development. Either way, a PDA should be assumed to work best for signals that resemble its developmental references.

Figure 12.9 shows the difference in GPE for each PDA and various speech corpora. The graph shows negative numbers for lower/better GPEs and positive numbers for higher/worse GPEs. Some PDAs show a marked preference for one corpus or another. Particularly if these preferences extend into positive SNRs, it is reasonable to assume that these PDAs were trained on the preferred corpus. Table 12.3 on page 142 listed all known training corpora used for various PDAs. In the following descriptions, PDAs known to be trained on the described corpus are highlighted with a *.

The strongest divisor, the *TIMIT* corpus, resulted in significantly worse GPEs for *BANA*, *CREPE*, *DIO*, *SRH*, *STRAIGHT*, and *YIN*, and better GPEs for *MBSC*, *NLS*, *SAFE*, *SHR*, and *YAAPT*, whereas *MOCHA-TIMIT* generally had the opposite effect. *CMU-ARCTIC* improved *DIO**, *NLS*, *SAFE**, *STRAIGHT*, and *YAAPT*, while *FDA** revealed the opposite. *PTDB-TUG** had a positive effect on *MAPS** and *SHR*, and a negative one on *PEFAC* and *YAAPT*. *KEELE*-Mod* was positive for *STRAIGHT*, *SHR**, and *YAAPT**. In general, these latter two corpora, *PTDB-TUG* and *KEELE*-Mod, were the most balanced from those investigated here.

Differences in Figure 12.9 are often particularly apparent in the transition area. This area is especially susceptible to ambiguities, as feature data become noisy and estimators more likely overlook fine details. Thus, the subtle differences between corpora are bound to become most visible in this area. At negative SNRs, details are obscured by noise anyway, and performance differences between corpora generally disappear.

In the transition area, the difference between corpora can be highly significant for some PDAs. From the graph, *KALDI*, *RAPT*, *RNN*, *SACC*, and *SIFT* can be said to be mostly invariant to the differences between the speech corpora. In some cases, the differences only affect a single corpus, such as *MAPS* and *PEFAC* for *PTDB-TUG*, or *SRH* and *YIN* for *TIMIT*. In other cases, every corpus is different, particularly for *CREPE*, *DIO*, *SAFE*, *SHR*, *STRAIGHT*, *SWIPE* and *YAAPT*. As the most extreme example, *CREPE*, *DIO*, and *STRAIGHT* lost more than 20% GPE in their transition area and the *TIMIT* corpus.

It is tempting to assume that such signal specificity should be most prevalent in neural-network-based approaches. However, the data showed no particular difference between algorithms with deep neural networks, such as *CREPE*, *DNN*, *SACC*, or learned decision matrices in *MAPS* and *PEFAC*, or even the more conventional hard-coded decision trees in the other PDAs. Conversely, manual parameter optimization seems to impart a similar amount of bias to a PDA as learning parameters from data.

These differences are no doubt caused in part by different base difficulties of the corpora themselves. To illustrate, the upper graph in Figure 12.10 shows a summary of each corpus' estimation difficulty. While some corpora, such as *CMU-ARCTIC*, seem somewhat less difficult than others, such as *MOCHA-TIMIT* or *FDA*, these differences are minor and do not explain the deviations seen in

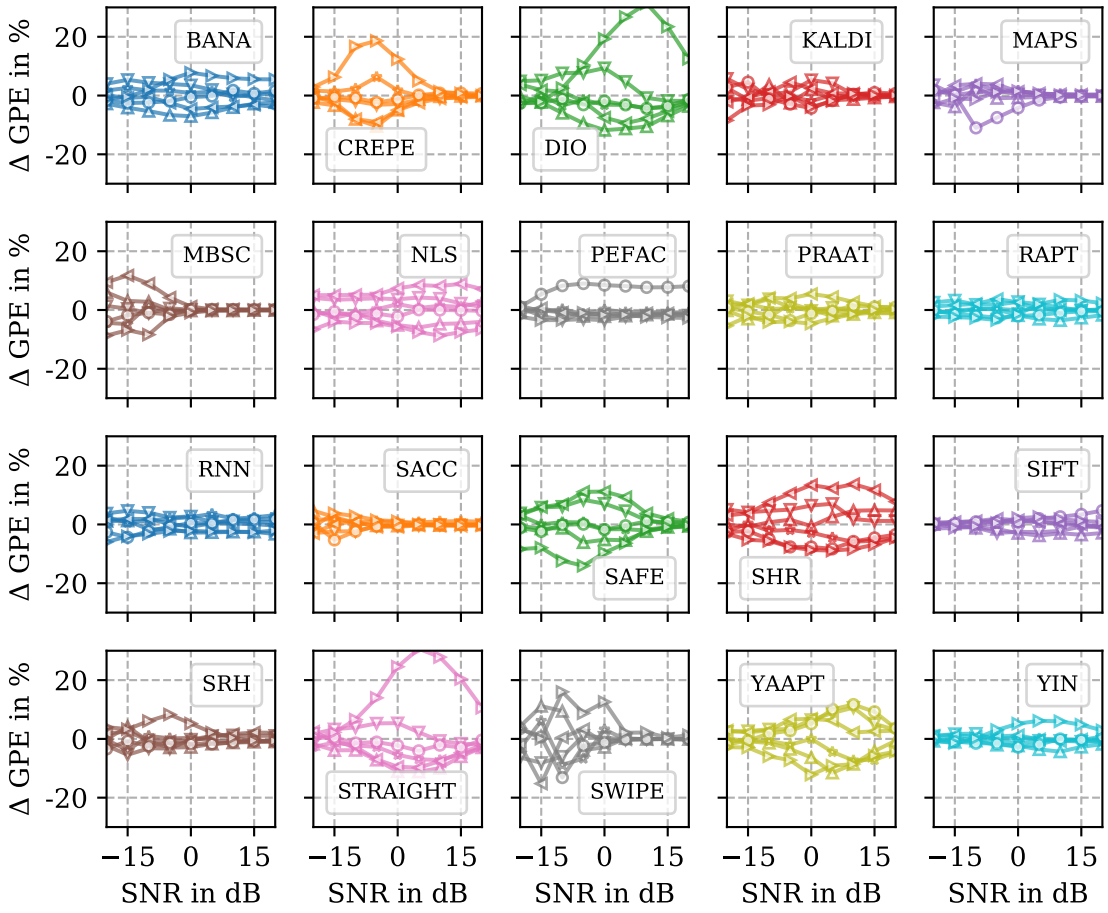


Figure 12.9: GPE delta for varying speech corpora in comparison to the mean over all corpora. △ is *CMU-ARCTIC*, ◁ is *MOCHA-TIMIT*, ▽ is *FDA*, ▷ is *TIMIT*, ○ is *PTDB-TUG*, and ⋆ is *KEELE*-Mod.

Figure 12.9.

Yet, what causes these differences? The bottom graph in Figure 12.10 shows the distribution of fundamental frequencies in each corpus. Most corpora have a balanced distribution of high (female) voices and low (male) voices. *FDA* differs from this with a peculiarly high female voice, but this does not seem to cause any difference in error measures. *TIMIT*, however, seems harder to estimate at lower SNRs than the other corpora, which could be explained by it having many more recordings of male voices than female, which are more easily masked by low-frequency noises. *MOCHA-TIMIT* is the opposite, and indeed did lead to the opposite behavior for some PDAs.

To summarize, there are significant differences in estimation accuracy between various corpora, and in the robustness of PDAs to differences between corpora. The different fundamental frequency distributions of the corpora undoubtedly played a role in this, as did their general clarity of pronunciation. The influence of fundamental frequency on PDA accuracy will be revisited in a later section. As a good general guideline, the *KEELE*-Mod and *PTDB-TUG* corpora seem to lead to relatively little accuracy variance, and *KALDI*, *RAPT*, *RNN*, *SACC*, and *SIFT* are most robust against speech signal variety.

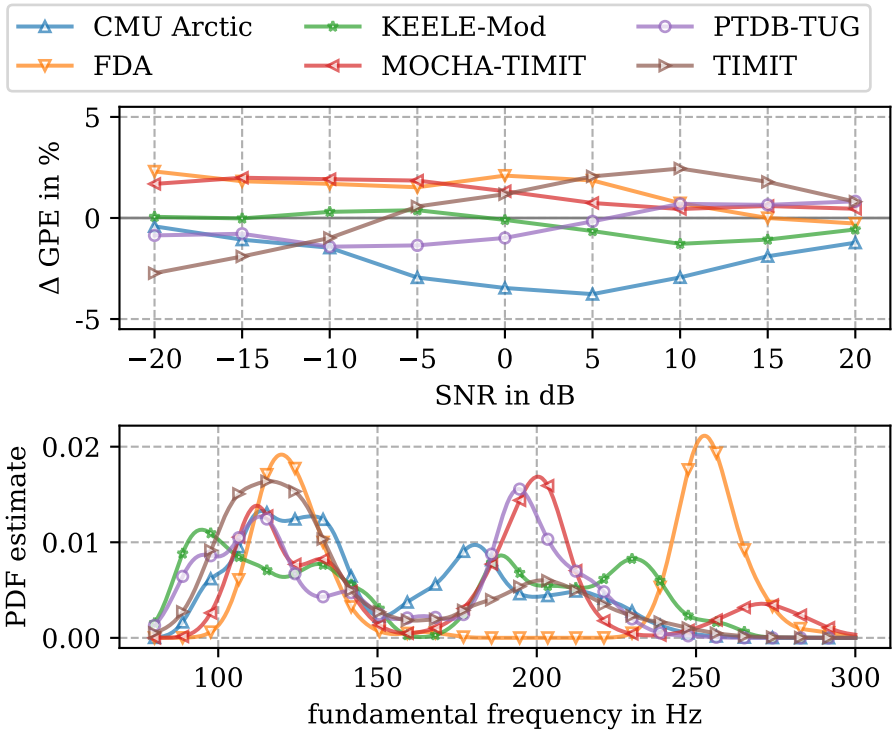


Figure 12.10: General overview of speech corpora. Upper graph is the mean GPE delta between the corpus and the mean of all corpora; the lower graph is a Gaussian kernel density estimate of the fundamental frequency distribution of each corpus.

Noise Corpus Dependence

Similar to the previous section on speech corpora, each PDA was explicitly or implicitly trained for a particular set of noises. Figure 12.11 illustrates the difference in GPE scores for each PDA and the two noise corpora *NOISEX* and *QUT-NOISE*.

For most PDAs, *NOISEX* seems to lead to more accurate estimates. Notably, *NOISEX* includes only very steady noises, such as constant engine noise or babble noise, whereas *QUT-NOISE*'s noise recordings are more natural in that they include noises with more level variations, such as recordings of traffic at an intersection or of a cafeteria. While these variations were equalized somewhat due to the SNR calculation being specific to short segments of speech and noise, algorithms unaccustomed to varying noise levels might still perform worse for *QUT-NOISE*. On the other hand, a more variant noise would also provide more possibilities for listening in momentarily quiet parts than steady noises without such gaps.

The only algorithm that actually performs better in *QUT-NOISE* than *NOISEX* is *MAPS*, which was in fact trained on *QUT-NOISE*.

Similar to the different speech corpora, accuracy differences often showed a maximum in the PDAs' transition areas. This time, *CREPE*, *MAPS*, *PEFAC*, *RNN*, *SACC*, and *SIFT* were mostly invariant to the differences between the noise corpora, while *DIO*, *KALDI*, *PRAAT*, *SAFE*, *STRAIGHT*, and *SWIPE* reflected a particularly strong influence.

In combination with the speech corpora dependence in the previous section, *RNN*, *SACC*, and *SIFT* were outstandingly robust against signal changes. This might make these PDAs of particular interest for applications with strongly varying or unpredictable signal and noise conditions. It additionally

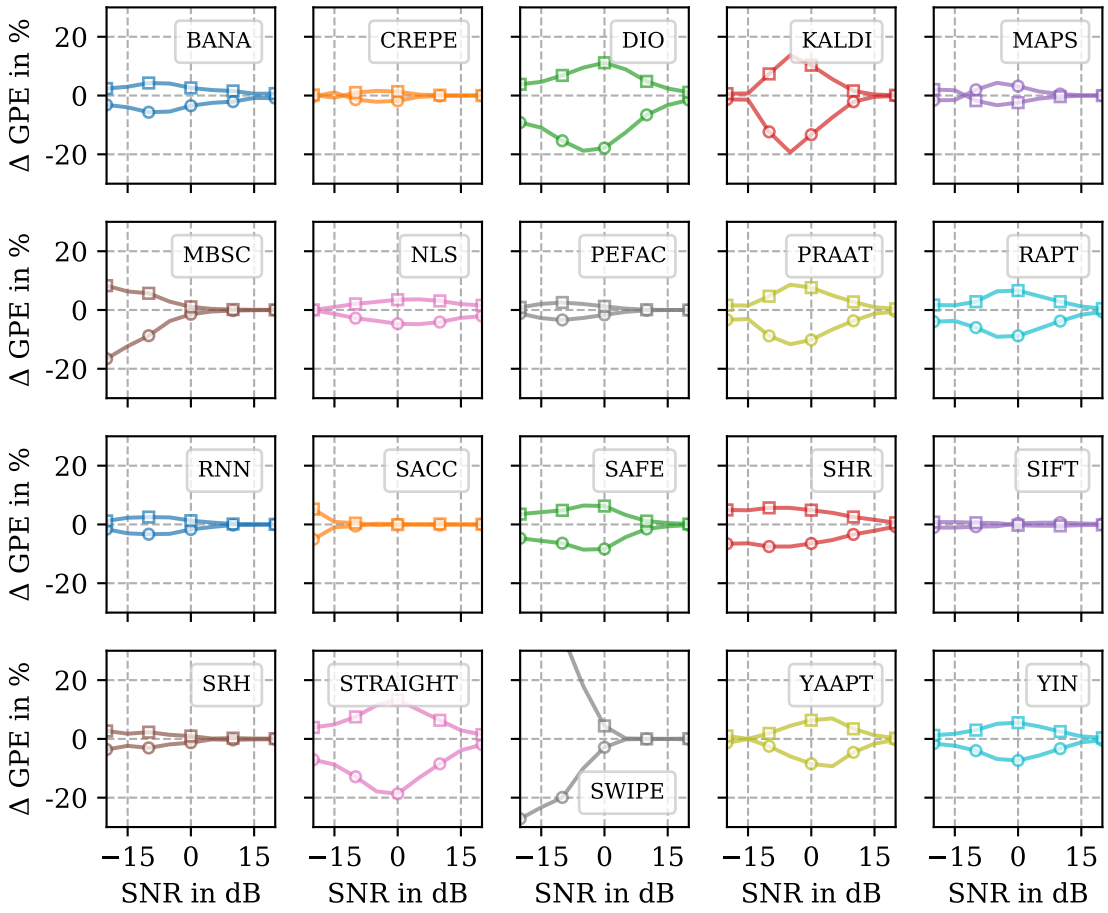


Figure 12.11: GPE difference between individual noise corpora and the overall mean. ○ is *NOISEX*, and □ is *QUT-NOISE*.

makes these PDAs ideal for comparison studies between PDAs, as they are least likely to skew results due to evaluation signal variability. Interestingly, two of these, *RNN* and *SACC*, employ neural networks in their algorithms, vindicating their stigma of being prone to over-fitting.

It is hard to overstate the significance of these results. In many cases, the difference between a well-matched speech recording and background noise for a PDA might yield a GPE difference of easily more than 20 %, or a noise robustness difference of ten or more dB SNR. In fact, these signal dependencies indicate that most comparison studies between PDAs should be reported as *suitability* for particular signals and noises, as opposed to "comparable" accuracy rankings.

Noise Type Dependence

In addition to noise corpora, each PDA may respond differently to the various types of noises contained therein. In general, the fundamental frequency of speech should be easier to detect in near-white, constant-spectrum noises such as car noise or traffic noise, and harder to detect in varying, tonal, or speech-like noises such as babble noise or cafeteria noise.

Figure 12.12 shows how the PDA accuracy changes with various noise types. The graphs show that some PDAs are more influenced by differences between noises than others. Some PDAs can reach very small GPEs across all SNRs for best-case noises, even rivalling the synthetic results from Figure 12.1.

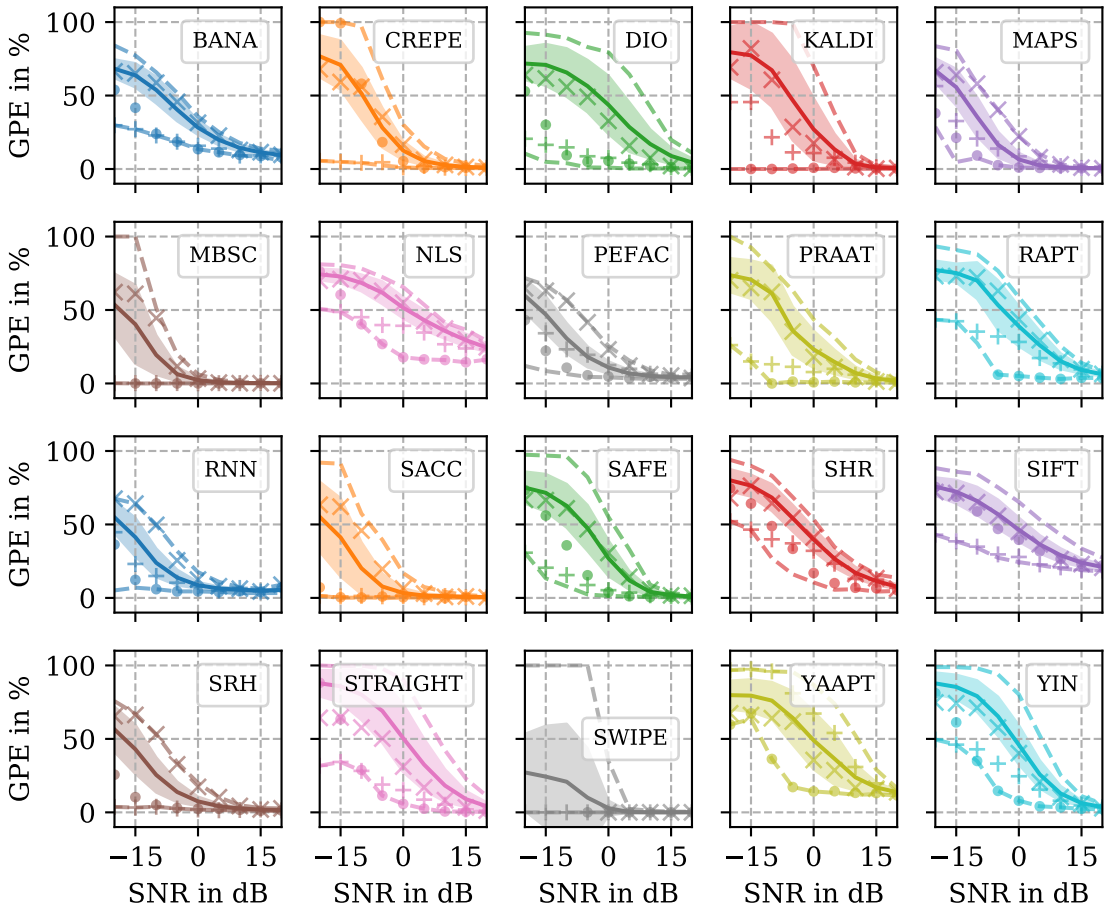


Figure 12.12: GPE vs. SNR of all PDAs from the entire realistic data set for the consensus truth. GPE mean are solid lines, mean absolute errors of the GPEs of different noise types are shaded areas, and best/worst noise GPE are dashed lines. Only for frames that are voiced according to the ground truth and the PDA. × mark babble/cafeteria noises, + machine gun noise, and circles mark white noise.

The opposite is true as well, with some PDAs being exceptionally strongly influenced by worst-case noises.

Additionally, the graph marks out three special kinds of noises: babble/cafeteria noise, which is very variable and has a similar long-time spectrum as speech, white noise, and machine gun noise, which alternates between very loud low-pass shots, and silence. Babble/cafeteria noise is generally one of the more difficult noises, particularly for *BANA*, *MAPS*, *MBSC*, *PEFAC*, *RNN*, *SACC*, and *SRH*. Interestingly, this includes most of the data-driven PDAs, whose decision matrices seem to be easily fooled by speech-like disturbances. One must imagine that this is particularly true for their VADs, which probably produce more false positives than usual in this case.

In contrast, white noise is generally one of the easiest noises to deal with, as it is spectrally and temporally unchanging and easy to discern from speech. The only PDAs that do not follow this pattern are *CREPE* at negative SNRs, and *SHR* and *SIFT*. In these particular cases, however, it is the machine gun noise that has even lower GPEs, likely due to its frequent quiet gaps between shots. This might imply very quick signal adaptation, which allows for good listening in gaps, but easy disturbance if there are none. Conversely, PDAs such as *YAAPT* and *RAPT* are particularly

strongly disturbed by machine gun noise, however, probably because of large smoothing constants that fill in the vital gaps.

Thus, depending on the application and kinds of noises expected, it might be preferable to select a PDA with low variability in response to noises, or one especially accurate for one particular kind of noise that is expected to occur in the intended application. It might also be preferable to select a PDA suitable for dealing with steady noises, or one that is better adapted to variable noises with gaps.

The difference in accuracy between a well-adapted PDA to a particular kind of noise and one unsuited can be extremely large. Taken together with the previous section on speech corpus dependence, these differences outweigh most other evaluations in this chapter.

Ground Truth Dependence

Some speech corpora provide their own fundamental frequency ground truth, some do not. To compensate for this, the consensus ground truth was calculated for all speech corpora used in this study. The choice of ground truth can potentially have a great impact on estimation accuracy, by excluding ambiguous or difficult frames, or simply deciding ambiguous cases similarly as a PDA.

An assumption about the origin of pitch is implicit in each ground truth: In case of a laryngograph-derived ground truth the argument follows a production model, where pitch is caused by vibrating vocal cords. The consensus truth, in contrast, interprets pitch from a perception point of view, where pitch is a property of an audio recording.

Figure 12.13 shows GPE scores for all PDAs for the corpus ground truth, the consensus truth, and additionally for the PDAs' own estimates of clean speech signals as ground truth. At truly clean SNRs (not shown), the GPE against the PDAs' own estimates always reaches zero.

The graph generally shows minimal differences between the consensus truth and the corpus' truth. This is as expected, as the fundamental frequency estimates should not differ much between these two methods. However, where there are differences, the corpus ground truth generally shows slightly more gross pitch errors. This is another indication that the corpus ground truths include some estimation errors of their own, which were erroneously attributed to the PDAs, as discussed earlier in Chapter 10.

This indicates that laryngographs are indeed a somewhat poor source of ground truth, as hypothesized in Chapter 10, at least for the purposes of evaluating PDAs. It remains an unanswered question, and a philosophical one at that, whether speech production or speech perception should serve as the source of truth for pitch estimation. But at least the data shows that PDAs generally estimated a perceptive pitch.

In addition, the graph shows GPEs with respect to the PDAs' own estimates of the clean speech recordings. While the GPEs of this "ground truth" did reach zero GPE at very high SNRs, the differences to the other two truths were surprisingly small. The only exceptions were *NLS* and *BANA*, which in the case of *NLS*, was probably due to its strong tendency towards lower octave errors. *BANA*s estimates, in contrast, seemingly diverged for clean and noisy recordings, with the noisy ones being more accurate. This might be a case of training or over-fitting to noisy speech, and a corresponding low suitability for clean speech.

In general, however, the estimation accuracy differences between the various ground truths was small. This is important as PDAs historically have been evaluated with a large variety of databases and ground truths, and at least this part of the evaluation procedure thus indeed proved unproblematic. In particular, it means that methods like the consensus truth can be used without issue to make new speech datasets available to fundamental frequency estimation research, and correspondingly enlarge the repertoire for future PDAs.

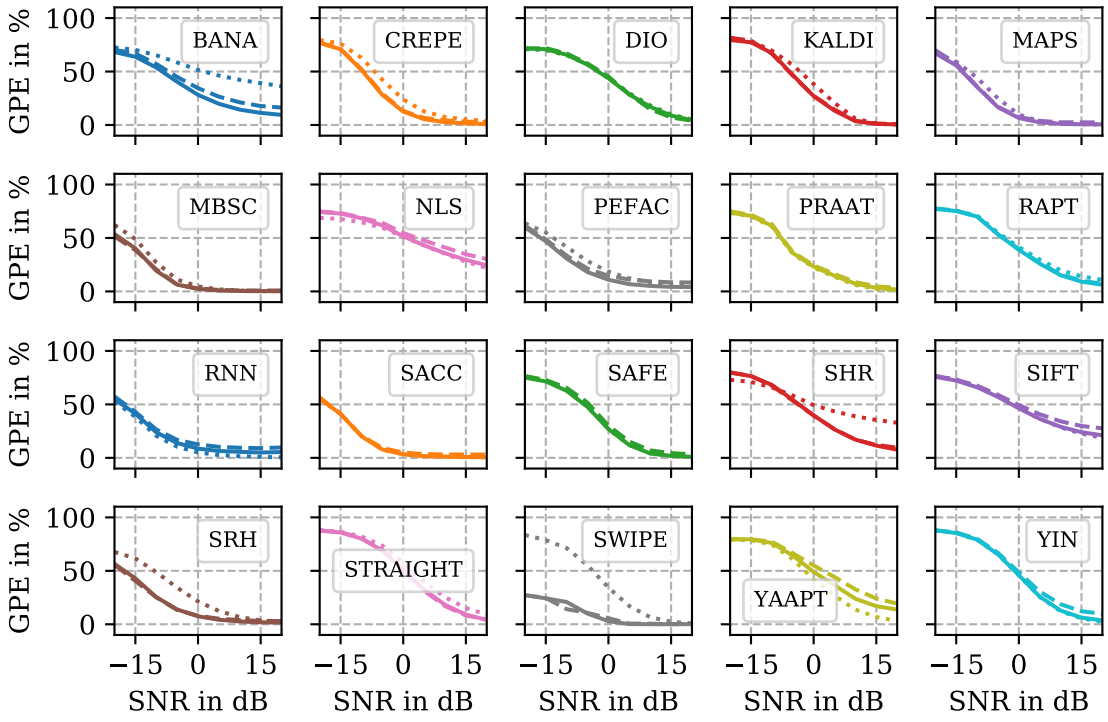


Figure 12.13: Comparison of various fundamental frequency ground truths. Solid lines are the consensus truth used for all previous graphs, dashed lines the speech corpus truth (only available for *FDA*, *KEELE*-Mod, and *PTDB-TUG*), and dotted lines use using each PDA's own estimates of the clean signal as ground truth.

Fundamental Frequency Dependence

Fundamental frequency estimation should perform well for the entire pitch range of human speech. However, depending on the training dataset and intended application, PDAs might acquire unintended biases. Figure 12.14 shows the estimation accuracy of each PDA for pitch ranges corresponding to male voices, intermediate voices, and female voices.

Perhaps surprisingly, every PDA was biased towards either male or female voices. The magnitude of these differences occasionally proved to be very large, up to 50 % GPE in some cases. Like in earlier evaluations, these differences usually only manifested within the transition area. Additionally however, they tended to *widen* towards negative SNRs, indicating that some PDAs are indeed capable of limited fundamental frequency estimation even at strongly negative SNRs for some voices.

Depending on the PDA, either male or female voices yield better accuracies. Interestingly, PDAs favoring higher pitches often seemed to be based on a time-domain signal representation, such as *CREPE*, *SACC*, and *SIFT*, whereas the majority of low-frequency optimized PDAs operate in the frequency domain. This could be attributed to the fact that higher-frequency waveforms offer more repetitions per block, while lower-frequency spectra show a greater number of harmonics.

For example, *DIO*, *KALDI*, *NLS*, *PRAAT*, *SAFE*, and *SHR* worked much better for low-frequency voices, while the aforementioned *CREPE*, *SACC*, and *SIFT* showed a slight preference for high-frequency voices. *BANA*, *CREPE*, *MAPS*, *MBSC*, *PEFAC*, *RNN*, *SACC*, *SIFT*, *SRH*, and *YIN* seemed relatively invariant to voice frequency, and a good choice for mixed-gender applications.

To investigate this in more detail, Figure 12.15 highlights how and when error rates changed

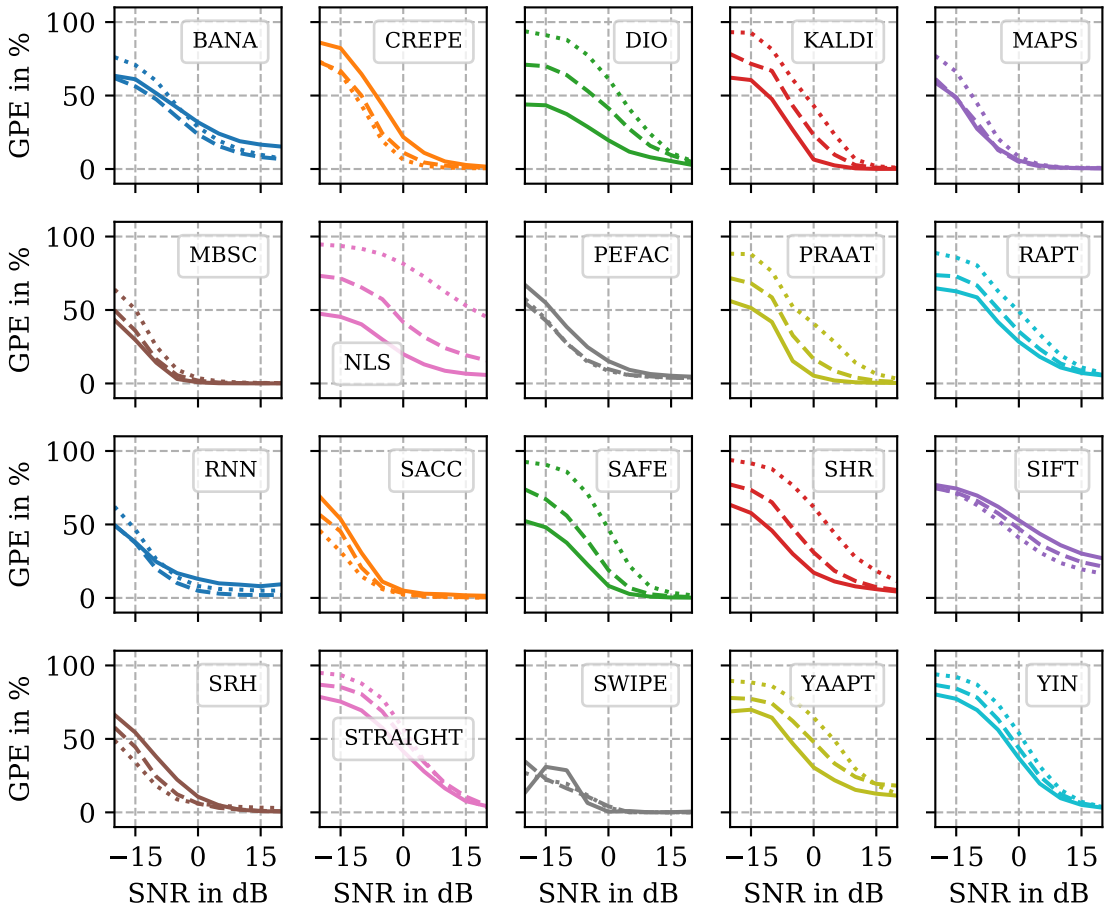


Figure 12.14: GPEs over SNR for various speech pitch ranges, roughly corresponding to male and female voices. Solid lines for pitches below 120 Hz, dashed lines between 120 Hz and 180 Hz, and dotted lines for pitches above 180 Hz. Only for frames that are voiced according to the ground truth and the PDA.

with pitch and SNR. Blue areas denote frequencies where the PDA performs worse than its mean at the given SNR, while yellow areas show better performance. Evidently, the performance of most PDAs differed most strongly at low SNRs, while the high SNR, low-GPE areas of most PDAs showed similar accuracy across all speech pitches. At low SNRs, most PDAs clearly favored higher or lower frequencies, as previously predicted. It should be noted, however, that the impression of a "tipping point" near 150 Hz is merely an artifact of the color map in the visualization, and does not imply this frequency as particularly divisive.

This graph shows that for some PDAs, these differences are strongly local to small frequency bands or SNR regions, while others are more evenly distributed across the frequency and SNR range. For example, *RNN*, *SRH*, *CREPE*, and *SACC* seemed to have low accuracy for very low pitches, but showed no such tendency above ca. 100 Hz. It seems likely that PDAs with deficiencies of very high or very low frequencies were designed for a more limited pitch range than the speech datasets used in this evaluation. Conversely, a strong preference for one frequency might indicate very a narrow training.

These preferences for high- or low-pitched voices might also explain a large part of the speech corpus preference exhibited in some PDAs and discussed earlier. In particular, Figure 12.10 showed

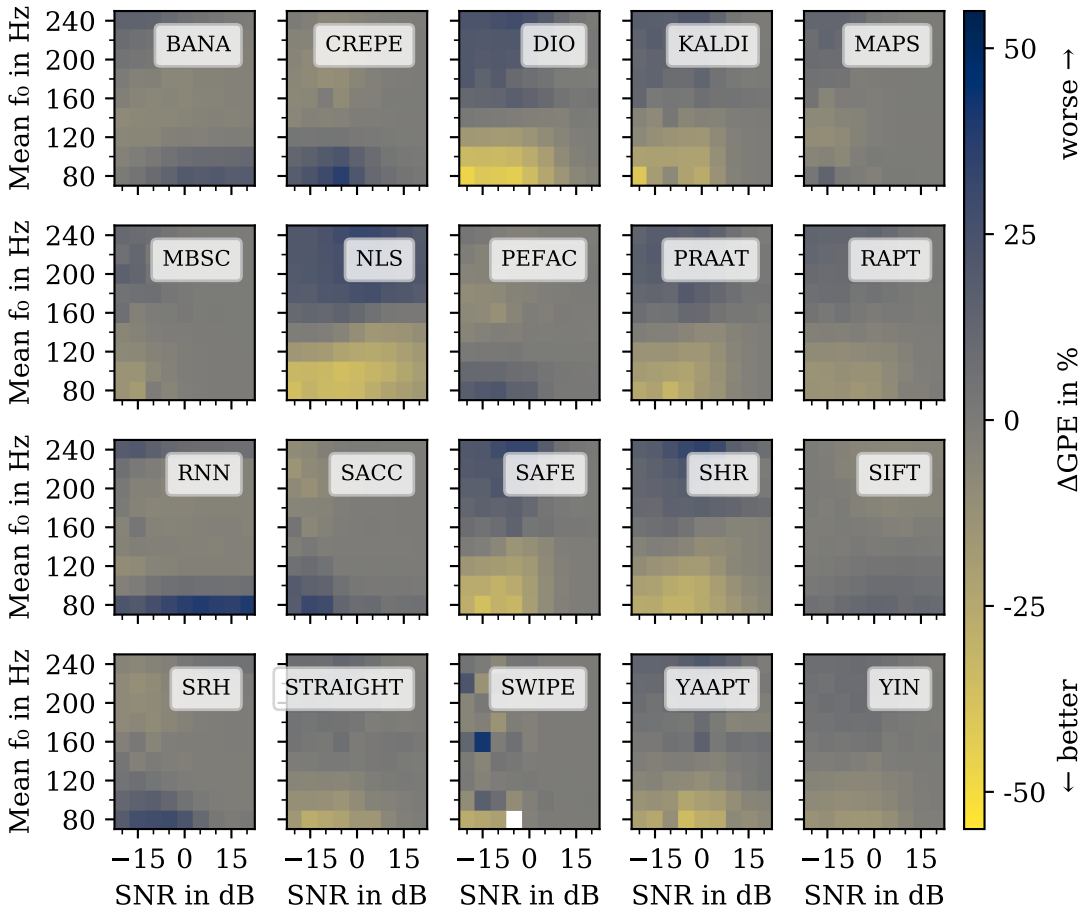


Figure 12.15: Fundamental frequency bias of PDAs. GPE change with respect to the mean GPE of each PDA over SNR and mean speech fundamental frequency. Only for frames that are voiced according to the ground truth and the PDA.

the *TIMIT* corpus to contain many more male voices than female, and the *MOCHA-TIMIT* corpus to be slightly skewed the other way. As for to the present topic, this implies that the former should be strongly preferable to *DIO*, *KALDI*, *NLS*, *PRAAT*, *SAFE*, and *SHR*. However, comparing this list with Figure 12.9 shows no such preference in a significant way. Perhaps this is due to the corpus differences being mostly prevalent in the transition area, while voice pitch biases are most strong at negative SNRs, with only a small overlap between the two.

Nevertheless, the GPE differences across speech pitches are quite large, and might well obscure more subtle differences between PDAs if not carefully controlled. As a preference for high or low pitches is inherently a gender issue, it should be considered with particular care when choosing a PDA for an application. *BANA*, *MAPS*, *MBSC*, *RAPT*, *SIFT*, and *YIN* are relatively unbiased for voice frequencies.

12.3 Conclusions

Publications on fundamental frequency estimation algorithms commonly try to highlight the particular strengths of one PDA over a set of “reference” PDAs. This necessity for superlative claims incentivizes

hiding subtle issues. The present study used a very big data set to establish performance measures and evaluation criteria useful for such a comparison. The analysis of that dataset showed that many seemingly minor parameters such as choice of signal and noise corpora, voice pitch, and the choice of training material can have a commanding influence on the evaluation results. Thus without a complete and identical set of performance metrics and evaluation signals, the results of comparison studies are almost incomparable.

In the face of such obstacles, any simple classification of PDAs into "machine learning vs classic signal processing", or "frequency domain vs time domain", or even "simple vs complex", have been found largely useless, as the performance parameters of each PDA were invariably more complex than any such simple categorization might suggest. As a result, this study must even end without a clear recommendation of a "best" PDA, as no such thing could be found in the data. No PDA was found outstandingly well-suited for all kinds of signal and noise, and no particularly worthwhile trade-off between computation time and estimation performance could be discerned.

Any measure of "good enough" must thus truly be evaluated on a per-application basis, and depends on a wide variety of factors, involving the kinds of speech recordings presented, the expected noise levels and types, the computational capacity available, and the preferred trade-off between recall and precision. Reasonable choices probably favor newer developments such as *MAPS*, *CREPE*, *RNN*, or *SACC*, although all of these come with particular caveats, and many circumstances will favor different PDAs. If noise is of little importance, older PDAs such as *YIN*, *SIFT*, or *PRAAT*, can also work well, and require fewer computational resources.

While the effort was made to find explanations for the differences in estimation accuracy between corpora, there is no reason to assume that our selection of speech and noise signals is in any way complete. It must therefore be assumed that real-world recordings are even more varied than the ones used in this study, and show even more aberrant behavior than the ones we have already seen. A thorough investigation of a PDA's intended application and suitability is therefore essential for getting accurate estimates; and due to the infinite vagaries of the real world, far beyond the scope of any controlled scientific comparison study.

In fact, this study has shown that the influence of these signal characteristics far outweighs any differences between PDAs. This is particularly important for outlier voices such as childrens' or pathologic ones, but also simply non-English speakers. If even the small differences between our speech corpora lead to large differences in accuracy, we have little hope for even more diverse voices. It is our hope that procedures like the consensus truth might open these kinds of voices to the world of pitch estimation, and increase diversity in our databases.

Interestingly, we found that many of the differences between PDAs are most prevalent in a transition area, where gross pitch errors slowly start to rise as SNRs deteriorate, but have not yet risen to a point where the estimates are essentially useless for most applications. It is in this region where frequency dependence, noise dependence, and octave errors are seen most strongly, and it is thus this transition that holds the key for evaluating the estimation accuracy of a PDA. However, it has also been found that this area is exceedingly small, and even a difference of a few decibels in signal-to-noise ratio can change an algorithms' behavior from mostly correct to almost useless.

However, these characteristics were often hidden from the error measures. In particular, most error measures only look at truly voiced frames, and are thus blind to VAD errors. This leads to gross pitch errors looking acceptable at negative SNRs, where in reality VAD false positives would introduce many incorrect estimates, and false negatives would leave nary a correct frame to estimate from. In fact, our results showed that negative SNRs rarely retained much useful data beyond their transition area, despite published claims to the contrary.

Future studies should invest effort into creating new error measures that quantify a PDA's behavior in this transition area. The existing error measures, such as gross pitch errors and voicing decision errors have been found to frequently confound independent variables, and to be easily misread. As

a first step, the appendix to this document includes a summary report on each individual PDA, to give a quick overview over its performance characteristics without any comparison. Additionally, the software repository attached to this dissertation contains source code for calculating the same report for new PDAs.

However, calculating such performance metrics for a large report such as this, is fraught with additional difficulties, such as availability of computer time and the sheer amount of data needing to be processed. It is the hope of the authors that by making the results of our calculations, and indeed the estimated fundamental frequencies and the speech and noise corpora themselves available freely as part of this dissertation[4], future researchers will be able to provide a more nuanced picture of their newly developed PDAs.

[4] see https://bastibe.github.io/Dissertation-Website/

Part VI

Conclusions

Chapter 13

What is the Pitch of Voiced Speech?

In Part I, this question was answered from a human perception point of view: Pitch is not a physical property of a signal, but our perception of it. While a thoroughly unsatisfying answer in terms of signal processing, it is ultimately the only truth there can be.

But if this is the answer, then all further inquiries in algorithmic pitch estimation are futile. If only a human mind can perceive pitch, then pitch estimation algorithms are an error of category, as algorithms may only estimate a numerical value, not what it "feels like" to hear a voice that is high or low.

As such, our definition of pitch had to be altered into a realm measurable and amenable to algorithmic analysis. In doing so we traversed onto the slippery slope between *pitch* on the one side, and *fundamental frequency* on the other. There are technical measures for fundamental frequency that can be derived from signal recordings, but are they pitches?

This is ultimately a philosophical question. An algorithm may estimate a physical property of a sound, such as its rate of repetition or the lowest frequency of a tone complex. And this may correspond to our perception of the sound being high or low. But whether these things are the same is a question that a dissertation on signal processing cannot answer. And yet, it *is* the ultimate question: in order to evaluate the accuracy of pitch determination algorithms, they need to be compared to a truth, and the choice of that truth must fall somewhere on the spectrum between perception and measurement.

Chapter 2 then provided an overview of the human apparatus for speech production and perception, with a particular emphasis on their relationship with pitch. Interestingly, these two perspectives provide two different interpretations of pitch that would become of greater importance later in the text: On the production side, voiced speech is a *periodic* signal produced from repeatedly opening and closing the vocal folds and exciting the vocal tract into resonance. On the perception side, voiced speech is a *harmonic* signal that excites the basilar membrane at regular intervals and with common phases.

To bring pitch estimation into a more technical context, Chapter 3 replaced our definitions of the problem with a simpler one: Pitch is what speech corpora say it is. Thus, the philosophical questions are replaced by concrete questions that can be answered with signal processing. As a corollary, a pitch determination algorithm is now deemed accurate if it produces results similar to those published in existing corpora. Implicit with this simplification comes not considering the influence of room acoustics, multi-channel recordings, multi-speaker recordings, abnormal modes of speaking, abnormal voices, and non-English speakers, because existing corpora do not provide a basis to evaluate these influences.

Part II changes perspective and views speech signals from a digital signal processing perspective. Chapter 5 examined speech within the analytic framework of the short-time Fourier transform, which

disentangles signals along time and frequency, not entirely unlike the filter-bank-like qualities of the human cochlea. Quite different from the organ of hearing, however, the STFT allows a deep introspection of signals frozen in time. Details of signals may be viewed without considering the passage of time, or the progression of frequencies. It is both the power of the STFT to allow this analysis and a great danger to ignore this context.

A particular interest was taken in the choice of window function for the STFT. Much like the shutter speed of a camera, the length of that window compromises between artificially freezing changing signals in time or integrating over their changes. Depending on the window length, the human voice appears either as periodic glottis pulses or as harmonic tone complexes. Additionally, the choice of window shape determines the "sharpening" of the spectra, trading off greater image contrast against lower ringing, the former desirable for STFT magnitudes and the latter for its phases. For STFT phases in particular, the chapter introduced the Hann-Poisson window, which has the unusual property of having no zeros in its magnitude spectrum, and therefore imposes no phase reversals onto STFT phases, making them significantly easier to interpret.

The chapter ended with an investigation into the graphical display of STFTs. The human visual system is just as complex as its auditory system, and thus the same care must be taken to avoid its idiosyncrasies and ambiguities as in audio signal processing. Part of this was dedicated to the design and application of a new color map specific to STFT phases, which avoids visual artifacts due to phase wrapping, and matches visual differences to equal steps in the displayed value.

To extract yet more information from the STFT, Chapter 6 then delved into STFT derivatives as a measure for their developments in time and frequency. Particularly for STFT phases, derivatives reveal structures that are not easily visible in the STFT phases themselves. Finding harmonic and periodic structures not only in STFT magnitudes, but also in its phase derivatives in fact opens a new line of reasoning to this dissertation: STFT phases are often discarded by speech analysis algorithms for being too hard to decipher. This also means that STFT phases are still comparatively unexplored, and thus interesting as an area of research. This was made use of in the following part.

Taken together, the introduction of the phase-focused Hann-Poisson window in Chapter 5, the dedicated color map for phase data in Chapter 5.3, and the STFT phase derivative in Chapter 6, provide a base for analyzing STFT phases in a new level of fidelity.

This was put into use in Part III, which introduced a new fundamental frequency estimation algorithm. Making use of both classical ideas such as a harmonic comb in the STFT magnitude, but also our STFT phase derivatives, the resulting algorithm proved both elegant and accurate. Its major new insights are that phase information and magnitude information can complement one another to correct their respective weaknesses, and that fundamental frequency estimation can be re-interpreted as a per-frequency voice activity determination problem to improve accuracy in the face of ambiguous estimates.

Exploring this solution exposed a problem with pitch estimation's most common error metric, the gross pitch error. Since the GPE is only applied to *voiced* frames, it is blind to some voicing determination errors: The GPE ignores both *unvoiced* frames and frames unlabeled by the ground truth, which essentially hides VAD false positives from evaluations. In real applications, however, such false positives would be visible as estimation errors.

By re-framing the problem of fundamental frequency estimation as a time-frequency voicing activity detection, our algorithm was able to all but avoid such false positives, albeit at the cost of somewhat increased false negatives. However, this conflicts with the standard of today's publications, which define an accurate pitch estimation to have low GPE in reference to a known ground truth. Thus we conclude that a better definition of pitch must include a definition of what pitch is not in order to be useful for real-world applications.

Part IV therefore endeavored to find a new ground truth that is more reliable and applicable than existing corpora. To date, there have been two sources of truth in pitch estimation corpora,

laryngographs and PDAs. The former subscribes to a production theory of speech pitch, where the *true* pitch of voiced speech is determined by the frequency of the vocal cord openings, as recorded by a measurement device called a laryngograph. Ignoring the technical problems of the measuring procedure, the speech production view defines pitches regardless of whether the speaker's mouth is opened or there is sufficient air flow to actually produce an audible sound. The alternative, basing the ground truth on another PDA, is similarly flawed, as it necessarily biases the truth with the signal model of the PDA. Additionally, even laryngograph recordings need to rely on a PDA to ascertain their pitch, and are thus not free from this problem, either.

Chapter 9 analyzed the differences between existing speech corpora and their ground truths, and ascertains how they might affect pitch estimation. Differences indeed proved significant, from varying signal levels, to varying voice diversity and to mere differences in the amount of data included in the corpora. None of the examined databases were found to be perfectly balanced or realistically diverse. They are, however, widely cited, and used for comparisons between publications and algorithms.

On the matter of ground truths, Chapter 10 defined a new ground truth from the consensus of a number of existing PDAs. Since all PDAs are based on reasonable theoretical concepts regarding signal properties that result from a periodicity or harmonicity being present in the signal, their consensus represents a kind of grand average of all these underlying concepts and ideas. While philosophically no "truer" than any other source of truth, this procedure at least ensured that the new *consensus truth* is categorically compatible with a PDA's world view as a perceptive measure, and is in principle available for arbitrary speech recordings without the need for specialized equipment such as a laryngograph.

Furthermore, a significant amount of differences was discovered between the consensus truth and existing, laryngograph-derived ground truths. This included differences in fundamental frequencies as well as differences in voicing activity, which showed the corpora's ground truths diverging from the majority of PDAs. Evaluating PDAs against the consensus truth therefore resulted in fewer errors, and a more truthful assessment of PDA accuracy.

Thus, the search for the nature of pitch has found yet another answer: Pitch is what pitch estimation algorithms estimate. A thoroughly tautological answer, of course, but perhaps philosophically not too dissimilar from the original definition of "what humans perceive as pitch". At least in the context of evaluating the accuracy of new pitch estimation algorithms, this majority vote on pitch is probably a reasonable solution, and allows for the analysis of speech corpora without an existing fundamental frequency ground truth.

Part V applies this new ground truth in the form of a large comparison study of pitch determination algorithms. The first part of this comparison, Chapter 11, assembled 25 PDAs from the entire history of digital fundamental frequency estimation, and ascertained their changing conception of the nature of pitch over the years. In general, earlier, computationally constrained approaches necessarily extracted pitch from relatively simple structures, whereas later PDAs delved into ever-greater depths to describe pitch and the human voice on a variety of levels.

It may not come as a surprise, then, that this diversity of opinions did not arrive at a common consensus. In fact, our comparative literature review revealed dramatic differences in reported accuracy, even if ostensibly identical corpora and algorithms were used in the publications. Pitch determination is not yet an exact science, it seems. These differences, of course, also apply to our own *consensus truth*, indicating that experimental parameters must have been so different between studies in the literature, as to make them virtually useless for comparison.

Consequently, our comparison was explicitly laid open for other researchers to replicate and expand, with complete definitions of all parameters and performance metrics, as well as reproducible open source code and results. Chapter 12 presented the results of this comparison, which was unprecedented in scope and detail. This included comparisons of the PDAs, the speech corpora, noise corpora, and ground truths, among dimensions both traditional and novel.

The results of this comparison study yielded uniquely detailed information about the merits of var-

ious algorithms. In this thorough evaluation, no single algorithm could be identified that performed best across all conditions. Particularly the differences between speech corpora, noise files, voice frequency, and voicing determination were greater than anticipated, which implies that different PDAs are not so much "better" or "worse", but are instead optimized for different applications.

Faced with this diversity, the consensus truth proved invaluable as an impartial basis for comparison between the large range of algorithms and signal conditions. Each PDA was no doubt developed with a particular application in mind, and a specific ground truth as a target. The consensus truth avoided such pre-conceived biases for particular signals and thus formed a more neutral basis for evaluating differential performances between the PDAs than any of the corpus ground truths.

Many of these varied analyses were the result of new performance metrics that expanded significantly upon the common average gross and fine pitch errors. Average gross pitch errors were found particularly problematic, as they do not show voicing determination errors, nor differences between signal conditions. One of the most meaningful aspects of PDA performance was instead the notion of a transition area, a particular SNR where the performance of a PDA starts to deteriorate, and where PDA differences are most pronounced. Both its location and the PDA's behavior at this SNR provided good insight into the overall PDA characteristics.

The evaluation demonstrated that a standardized evaluation protocol and a large dataset can indeed provide the detail necessary to evaluate PDAs for particular applications. However, this can only be a first step towards reproducible research, with error measures in particular still in dire need of revision. Regardless, with this framework in place, it is now possible to select a PDA that is most fit for a particular task, if not "the best" overall.

In summary, there simply cannot be one complete definition of algorithmic pitch. Just as each human's perception of pitch is slightly different, so are there differences between algorithms. In fact, as Chapter 2 illustrated, humans can be internally ambiguous about the pitch of single sounds, perceiving it in a variety of ways depending on context, or simply by choice. This same dynamic seems to be present in pitch determination algorithms designed for different applications. Thus, there can be no *one true pitch*, but only a *variety of pitches* for different applications.

Chapter 14

Epilogue: Whither, Pitch Estimation?

Along the way in this investigation of pitch estimation, many a theory was put forth to make sense of the jungle of definitions. I have consistently resisted the urge to classify approaches into simple categories. The duality of *pitch* versus *fundamental frequency* comes to mind as the most prominent one, which seemed clear-cut to me at the beginning, but truly lost all meaning by the end as the concept of "that which is high or low" simply only exists as a human construct, regardless of whether it is estimated by technology or biology.

Similarly, the differences between *digital* and *analog* algorithms, or *harmonic* and *periodic* interpretations of signals, all too often fail to capture the essence of an algorithm's design. In fact, even comparing the resulting algorithms from starting points such as a statistical model, machine learning, laryngograph ground truths, or simply the intuition of the individual engineer, all seem to converge on a similar set of attributes that are more alike than different.

With the unprecedented depth of analysis in the comparison study at the end of this dissertation, it became abundantly clear that the general problem of *pitch estimation* is in fact a solved one. Minute differences in particular error measures can always be extracted. But does that truly make an algorithm "better"? Only in reference to a ground truth can this question be answered, but this changes the question from one about algorithms to one about truths, and we are none the wiser.

Perhaps it is not surprising that there are few secrets left in an area of research that has garnered over eight hundred publications in the last thirty years.

Thus, if pitch estimation itself is no longer a fruitful topic of study, where does that leave this dissertation? Over the course of my work on this topic, I have found the meta-analysis to be the most interesting topic of all. Parts of this fascination have made it into the actual work in the form of the exploration of STFT phases, and new error measures.

But there are several lines of research I have regrettably never found the time to work on: such as the true meaning of STFT phases. Being that the instantaneous frequency and group delay are derivatives, their zeros must correspond to extrema in the phase manifold. In fact, a simulated hill shading based on the slopes defined by the phase derivatives gives the impression of a continuous surface. It should thus be possible to utilize phase derivatives to completely unwrap the phase manifold, and I would be terribly interested in the emergent shape, so long hidden in plain sight.

Similarly, the derivatives of the STFT magnitude are smooth slopes in the vicinity of dominant transients such as clicks or sinusoids or sweeps (three expressions of the same phenomenon, really). Much like phase derivatives can be used to reassign STFT bins back to their source locations, so could hill climbing be used to reattribute magnitude bins. In fact, the combination of these two methods of reassignment should give a strong indication of whether an STFT bin belongs to a random noise fluctuation or a meaningful and dominant voice partial.

These ideas are particularly interesting, as they view pitch not as a property of a single spectrum,

but as a two-dimensional structure that spans both time and frequency. The possible movements within this structure are limited by the mobility of the human vocal organs in specific, constrained ways. I believe that if we had a more complete map of these states and their possible transitions, we could constrain our speech analysis models to infer speech from much sparser data. In fact, most of our own ability to detect speech in wildly adverse acoustic scenarios is probably based on our intimate knowledge on the constrained possibility space of human voices.

However, these considerations are independent of the estimation of individual pitches. Ultimately, I do not think that there is a holy grail of pitch estimation yet to be found. Pitch is merely a fantasy that seems so close, yet never quite close enough to reach. But trying to reach that goal inevitably leads one down the rabbit hole of the human voice and its infinite intricacies that is far more interesting than pitch itself ever was. Or, in the words of Indiana Jones:

> Henry: Elsa never really believed in the grail. She thought she'd found a prize.
> Indy: And what did you find, Dad?
> Henry: Me? Illumination!
> —Indiana Jones [141]

Acknowledgments

To say that these past six years of working on my PhD have been tumultuous, would be an understatement. Science, it turns out, is a hike through the wilderness; There are dangers along the way, and distractions, but also profound beauty.

Navigating these dangerous waters would have been entirely impossible without the help of my friends and colleagues Jens, Ulrik, Paul, Grace, Matthias, and Menno, all of whom directly contributed to this work with insightful discussions and proofreading.

And I could not have persevered without the love and support of my wife Grace and our daughter Elisa, and my parents Achim and Barbara.

And of course, I am indebted to my supervisors Jörg Bitzer and Steven van de Par, whose guidance was invaluable along the way, and whose wisdom was instrumental in all parts of this work.

Bibliography

[1] G. Aneeja and B. Yegnanarayana. Extraction of Fundamental Frequency From Degraded Speech Using Temporal Envelopes at High SNR Frequencies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):829–838, April 2017.

[2] Elias Azarov, Maxim Vashkevich, and Alexander Petrovsky. Instantaneous pitch estimation algorithm based on multirate sampling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4970–4974, Shanghai, March 2016. IEEE.

[3] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, page 7815—7819. IEEE, 2013.

[4] Paul A Bagshaw. *Automatic prosodic analysis for computer aided pronunciation teaching.* PhD thesis, University of Edinburgh, Edinburgh, UK, 1994.

[5] Paul C Bagshaw, Steven Hiller, and Mervyn A Jack. Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching. In *EUROSPEECH*, 1993.

[6] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors. *Springer handbook of speech processing.* Springer, Berlin ; London, 2008. OCLC: ocn190966783.

[7] W. Bennett. Secret Telephony as a Historical Example of Spread-Spectrum Communication. *IEEE Transactions on Communications*, 31(1):98–104, January 1983.

[8] Boualem Boashash and Saman S. Abeysekera. Two Dimensional Processing Of Speech And Ecg Signals Using The Wigner-Ville Distribution. In Andrew G. Tescher, editor, *Applications of Digital Image Processing IX*, volume 0697, pages 142 – 153. SPIE, 1986. Backup Publisher: International Society for Optics and Photonics.

[9] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, page 97—110. Amsterdam, 1993.

[10] Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic Sciences of the University of Amsterdam, Report*, 132:182, 1996.

[11] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of Artery Visualizations for Heart Disease Diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2479–2488, December 2011.

[12] David Borland and Russell M. Taylor II. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, pages 14–17, 2007.

[13] Ann Bradlow, Cynthia Clopper, Rajka Smiljanic, and Mary Ann Walter. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, 52(11-12):930–942, November 2010.

[14] Mike Brookes. VOICEBOX: speech processing toolbox for MATLAB. online resource. accessed 05 2016.

[15] Henrik Brumm and Sue Anne Zollinger. The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour*, 148(11-13):1173–1198, 2011.

[16] C. Busso, Sungbok Lee, and S. Narayanan. Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596, May 2009.

[17] Denis Byrne, Harvey Dillon, Khanh Tran, Stig Arlinger, Keith Wilbraham, Robyn Cox, Bjorn Hagerman, Raymond Hetu, Joseph Kei, C. Lui, Jurgen Kiessling, M. Nasser Kotby, Nasser H. A. Nasser, Wafaa A. H. El Kholy, Yasuko Nakanishi, Herbert Oyer, Richard Powell, Dafydd Stephens, Rhys Meredith, Tony Sirimanna, George Tavartkiladze, Gregory I. Frolenkov, Soren Westerman, and Carl Ludvigsen. An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America*, 96(4):2108–2120, 1994.

[18] Arturo Camacho. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. PhD thesis, University of Florida, 2007.

[19] D. Chan, A. Fourcen, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Transcoso, C. Veld, and J. Zeiliger. EUROM-A Spoken Language Resource for the EU. In *Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, volume 1, pages 867–870, Madrid, Spain, September 1995.

[20] F. J. Charpentier. Pitch detection using the short-term phase spectrum. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*, volume 11, page 113—116. IEEE, 1986.

[21] Mads Græsbøll Christensen and Andreas Jakobsson. *Multi-pitch estimation*. Morgan & Claypool Publishers, [San Rafael, Calif.], 2009.

[22] Wei Chu and Abeer Alwan. SAFE: a statistical algorithm for F0 estimation for both clean and noisy speech. In *INTERSPEECH*, pages 2590–2593, 2010.

[23] Ryunosuke Daido and Yuji Hisaminato. A Fast and Accurate Fundamental Frequency Estimator Using Recursive Moving Average Filters. In *INTERSPEECH 2016*, pages 2160–2164, September 2016.

[24] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917, 2002.

[25] David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. *Proceedings of Interspeech 2010*, 2010.

[26] David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. *Proceedings of Interspeech 2010*, 2010.

[27] Gilles Degottex and Daniel Erro. A uniform phase representation for the harmonic model in speech synthesis applications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):38, 2014.

[28] Boyuan Deng, Denis Jouvet, Yves Laprie, Ingmar Steiner, and Aghilas Sini. Towards confidence measures on fundamental frequency estimations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5605–5609, New Orleans, LA, March 2017. IEEE.

[29] David Deterding. The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, 36(2):187–196, December 2006.

[30] Jitendra Kumar Dhiman, Nagaraj Adiga, and Chandra Sekhar Seelamantula. A Spectro-Temporal Demodulation Technique for Pitch Estimation. In *Interspeech 2017*, pages 2306–2310. ISCA, August 2017.

[31] Thomas Drugman and Abeer Alwan. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. In *Interspeech*, page 1973—1976, 2011.

[32] Thomas Drugman, Goeric Huybrechts, Viacheslav Klimkov, and Alexis Moinet. Traditional Machine Learning for Pitch Detection. *IEEE Signal Processing Letters*, 25(11):1745–1749, November 2018.

[33] Hendrikus Duifhuis, Lei F. Willems, and R. J. Sluyter. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, 71(6):1568–1580, 1982.

[34] Kelly R. Fitz and Sean A. Fulop. A Unified Theory of Time-Frequency Reassignment. *arXiv:0903.3080 [cs]*, March 2009. arXiv: 0903.3080.

[35] James L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1965.

[36] P. Flandrin, F. Auger, E. Chassande-Mottin, and others. Time-Frequency reassignment from principles to algorithms. *Applications in time-frequency signal processing*, 5:179–203, 2002.

[37] Open Knowledge Foundation. Open Data Commons Open Database License (ODbL).

[38] D. Friedman. Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 1121–1124. IEEE, 1985.

[39] Sean A. Fulop and Kelly Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360–371, January 2006.

[40] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[41] David Gerhard, University of Regina, and Department of Computer Science. *Pitch extraction and fundamental frequency: history and current techniques*. Dept. of Computer Science, University of Regina, Regina, 2003. OCLC: 54005806.

[42] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2494–2498. IEEE, 2014.

[43] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, Inc., Hoboken, NJ, USA, August 2011.

[44] John Goldsmith, Jason Riggle, and Alan C. L. Yu. *The Handbook of Phonological Theory*. Wiley-Blackwell, 2011.

[45] Sira Gonzalez and Mike Brookes. PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):518—530, February 2014.

[46] Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Low-Complexity Pitch Estimation Based on Phase Differences Between Low-Resolution Spectra. In *Interspeech 2017*, pages 2316–2320. ISCA, August 2017.

[47] B Griggs. *The end of rainbow? An exploration of color in scientific visualization*. PhD Thesis, Thesis, University of Oregon, 2014.

[48] Roland Gööck. *Die großen Erfindungen*. Sigloch Edition, 1985.

[49] Habib Hajimolahoseini, Rassoul Amirfattahi, Saeed Gazor, and Hamid Soltanian-Zadeh. Robust Estimation and Tracking of Pitch Period Using an Efficient Bayesian Filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1, 2016.

[50] Kun Han and DeLiang Wang. Neural Network Based Pitch Tracking in Very Noisy Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2158–2168, December 2014.

[51] Philip Harding and Ben Milner. Estimating acoustic speech features in low signal-to-noise ratios using a statistical framework. *Computer Speech & Language*, 42:1–19, March 2017.

[52] M L L Harries, J M Walker, D M Williams, S Hawkins, and I A Hughes. Changes in the male voice at puberty. *Archives of Disease in Childhood*, 77(5):445–447, November 1997.

[53] Cyril M Harris and Mark R Weiss. Pitch Extraction by Computer Processing of High-Resolution Fourier Analysis Data. *The Journal of the Acoustical Society of America*, 35(3):339–343, 1963.

[54] Lars Hausfeld, Alexander Gutschalk, Elia Formisano, and Lars Riecke. Effects of Cross-modal Asynchrony on Informational Masking in Human Cortex. *Journal of Cognitive Neuroscience*, 29(6):980–990, June 2017.

[55] David Hay. First Transatlantic Telephone Cable, 2014.

[56] He Ba, Na Yang, Ilker Demirkol, and Wendi Heinzelman. BaNa: A hybrid approach for noise resilient pitch detection. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 369–372, Ann Arbor, MI, USA, August 2012. IEEE.

[57] Peter Juel Henrichsen and Marcus Uneson. SMALLWorlds – a Multi-lingual Speech Corpus for Cognitive Research. In *The Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, may 2012.

[58] Dik J. Hermes. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

[59] W Hess. *Pitch determination of speech signals.* Springer-Verlag Berlin An, Place of publication not identified, 1983. OCLC: 933705411.

[60] Ian Howard and David Howard. Quantitative comparisons between time domain speech fundamental frequency estimation algorithms. In *Proc. IOA*, volume 8, page 323—330, 1986.

[61] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. Publisher: IEEE COMPUTER SOC.

[62] Anton A. Huurdeman. *The Worldwide History of Telecommunications.* John Wiley & Sons, Inc., Hoboken, NJ, USA, July 2003.

[63] A. Nejat Ince. *Digital speech processing: Speech coding, synthesis, and recognition.* Springer Science+Business Media, 1992.

[64] IPA. *Handbook of the International Phonetic Association.* Cambridge University Press, 1999.

[65] Zhaozhang Jin and DeLiang Wang. A multipitch tracking algorithm for noisy and reverberant speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4218–4221. IEEE, 2010.

[66] Denis Jouvet and Yves Laprie. Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1614–1618, Kos, Greece, August 2017. IEEE.

[67] Sudarsana Reddy Kadiri and Bayya Yegnanarayana. Estimation of Fundamental Frequency from Singing Voice Using Harmonics of Impulse-like Excitation Source. In *Interspeech 2018*, pages 2319–2323. ISCA, September 2018.

[68] Karl-Dirk Kammeyer and Kristian Kroschel. *Digitale Signalverarbeitung.* Vieweg+Teubner, 2009.

[69] Kavita Kasi and Stephen A. Zahorian. Yet Another Algorithm for Pitch Tracking. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–361–I–364, Orlando, FL, USA, May 2002. IEEE.

[70] Hideki Kawahara. Nearly Defect-Free F0 Trajectory Extraction for Expressive Speech Modifications Based on STRAIGHT. *Interspeech 2005*, page 4, 2005.

[71] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207, April 1999.

[72] Hideki Kawahara and Masanori Morise. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana*, 36(5):713–727, October 2011.

[73] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3933–3936. IEEE, 2008.

[74] Hideki Kawahara, Ken-Ichi Sakakibara, Hideki Banno, Masanori Morise, Tomoki Toda, and Toshio Irino. Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 520–529, Hong Kong, December 2015. IEEE.

[75] Hideki Kawahara, Ken-Ichi Sakakibara, Masanori Morise, Hideki Banno, and Tomoki Toda. A Modulation Property of Time-Frequency Derivatives of Filtered Phase and its Application to Aperiodicity and fo Estimation. In *Interspeech 2017*, pages 424–428. ISCA, August 2017.

[76] Hideki Kawahara, Toru Takahashi, Masanori Morise, and Hideki Banno. Development of exploratory research tools based on TANDEM-STRAIGHT. *APSIPA ASC 2009*, page 11, 2009.

[77] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. Dysarthric Speech Database for Universal Access Research. *Interspeech 2008*, page 4, 2008.

[78] Hyun Sik Kim and Alan G. Marshall. Magnitude-mode multiple-derivative spectra for resolution enhancement without loss in signal-to-noise ratio in Fourier transform spectroscopy. *Journal of Mass Spectrometry*, 30(9):1237–1244, September 1995.

[79] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. CREPE: A Convolutional Representation for Pitch Estimation. *arXiv:1802.06182 [cs, eess, stat]*, February 2018. arXiv: 1802.06182.

[80] John Kominek and Alan W Black. CMU ARCTIC database for speech synthesis, 2003.

[81] Peter Kovesi. Good Colour Maps: How to Design Them. *arXiv preprint arXiv:1509.03700*, 2015.

[82] Martin Krawczyk and Timo Gerkmann. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1931–1940, 2014.

[83] Martin Krawczyk-Becker and Timo Gerkmann. Fundamental Frequency Informed Speech Enhancement in a Flexible Statistical Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):940–951, May 2016.

[84] Sandeep Kumar. Performance Evaluation of Novel AMDF-Based Pitch Detection Scheme. *ETRI Journal*, January 2016.

[85] Marie Lebert. Project gutenberg (1971-2008), 2008.

[86] Byung Suk Lee and Daniel PW Ellis. Noise robust pitch tracking by subband autocorrelation classification. In *Interspeech*, pages 707–710, 2012.

[87] H. Levkowitz and G. T. Herman. Color scales for image data. *IEEE Computer Graphics and Applications*, 12(1):72–80, January 1992.

[88] Bin Liu, Jianhua Tao, Dawei Zhang, and Yibin Zheng. A novel pitch extraction based on jointly trained deep BLSTM Recurrent Neural Networks with bottleneck features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340, New Orleans, LA, March 2017. IEEE.

[89] Yuzhou Liu and DeLiang Wang. Robust pitch tracking in noisy speech using speaker-dependent deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5255–5259, Shanghai, March 2016. IEEE.

[90] Philipos C Loizou. *Speech Enhancement*. CRC Press, 2017.

[91] Erfan Loweimi, Jon Barker, and Thomas Hain. On the Usefulness of the Speech Phase Spectrum for Pitch Extraction. In *Interspeech 2018*, pages 696–700. ISCA, September 2018.

[92] Sylvain Marchand and Philippe Depalle. Generalization of the Derivative Analysis Method to Non-Stationary Sinusoidal Modeling. *DAFx 2008*, page 9, 2008.

[93] J. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics*, 20(5):367—377, December 1972.

[94] Philippe Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, volume 7, page 180—183. IEEE, 1982.

[95] MathWorks. Matlab, 1984.

[96] A. McCree. A Scalable Phonetic Vocoder Framework Using Joint Predictive Vector Quantization of Melp Parameters. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–705–I–708, Toulouse, France, 2006. IEEE.

[97] Carol A. McGonegal, Lawrence R. Rabiner, and Aaron E. Rosenberg. A subjective evaluation of pitch detection methods using LPC synthesized speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(3):221—229, 1977.

[98] B. C. J. Moore and H. E. Gockel. Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591):919–931, April 2012.

[99] Brian CJ Moore and Hedwig Gockel. Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, 88(3):320–333, 2002.

[100] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *AES 2009*, page 5, 2009.

[101] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884, 2016.

[102] Na Yang, He Ba, Weiyang Cai, Ilker Demirkol, and Wendi Heinzelman. BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1833–1848, December 2014.

[103] Jesper Kjær Nielsen, Tobias Lindstrøm Jensen, Jesper Rindom Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient. *Signal Processing*, 135:188–197, June 2017.

[104] Jesper Kjø er Nielsen, Tobias Lindstr, Jesper Rindom Jensen, Mads Grø esb, and others. Fast and statistically efficient fundamental frequency estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 86–90. IEEE, 2016.

[105] A. Michael Noll. Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 1967.

[106] A. Michael Noll. Pitch Determination of Human Speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In *Symposium on Computer Processing in Communications*, 1969.

[107] Robert B. Ochsman and Alphonse Chapanis. The effects of 10 communication modes on the behavior of teams during co-operative problem-solving. *International Journal of Man-Machine Studies*, 6(5):579–619, September 1974.

[108] Kyong-Ae Oh and Chong Kwan Un. A performance comparison of pitch extraction algorithms for noisy speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, page 85—88. IEEE, 1984.

[109] A.V. Oppenheim and R.W. Schafer. Dsp history - From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5):95–106, September 2004.

[110] Douglas O'Shaughnessy. *Speech Communication : Human and Machine*. Addison-Wesley, 1987.

[111] Andrew J. Oxenham. How We Hear: The Perception and Neural Coding of Sound. *Annual review of psychology*, 69, 2018.

[112] K. K. Paliwal and A. I. Aarskog. *A Comparative Performance Evaluation of Pitch Estimation Methods for THDS/Subband Coding of Speech*. Elsevier Science, 1984.

[113] Kuldip K. Paliwal and L. Alsteris. Usefulness of phase in speech processing. In *Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan*, pages 1–6, 2003.

[114] Kuldip K. Paliwal and Leigh D. Alsteris. On the usefulness of STFT phase spectrum in human listening tests. *Speech Communication*, 45(2):153–170, February 2005.

[115] Vishala Pannala, G. Aneeja, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Robust Estimation of Fundamental Frequency Using Single Frequency Filtering Approach. In *INTERSPEECH 2016*, pages 2155–2159, September 2016.

[116] Antonio Papandreou-Suppappola. *Applications in Time-Frequency Signal Processing*. CRC Press, 2003.

[117] T. W. Parks and J. D. Wise. Maximum likelihood pitch estimation. In *Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications, 1977 IEEE Conference on*, page 1092—1095, December 1977.

[118] J. R. Pierce. Whither Speech Recognition? *The Journal of the Acoustical Society of America*, 46(4B):1049–1051, October 1969.

[119] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf. A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario. *Interspeech 2011*, page 4, 2011.

[120] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf. A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario. In *INTERSPEECH*, page 1509—1512, 2011.

[121] F. Plante, G. Meyer, and W.A. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Transactions on Speech and Audio Processing*, 6(3):282–287, May 1998.

[122] F Plante, Georg F Meyer, and William A Ainsworth. A Pitch Extraction Reference Database. In *Fourth European Conference on Speech Communication and Technology*, pages 837–840, Madrid, Spain, 1995.

[123] Alexandros Potamianos, Shrikanth Narayanan, and Sungbok Lee. AUTOMATIC SPEECH RECOGNITION FOR CHILDREN. *EUROSPEECH 1997*, page 4, 1997.

[124] K. M. M. Prabhu. *Window Functions and Their Applications in Signal Processing*. CRC Press, 2014.

[125] RaviShankar Prasad and B Yegnanarayana. Robust Pitch Estimation in Noisy Speech Using ZTW and Group Delay Function. *INTERSPEECH 2015*, page 4, 2015.

[126] B. G. Quinn and P. J. Thomson. Estimating the frequency of a periodic function. *Biometrika*, pages 65–74, 1991.

[127] Lawrence Rabiner, Michel J. Cheng, Aaron E. Rosenberg, and Carol A. McGonegal. A comparative performance study of several pitch detection algorithms. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(5):399—418, 1976.

[128] Rajeev Rajan and Hema A. Murthy. Modified Group Delay Based MultiPitch Estimation in Co-Channel Speech. *arXiv:1603.05435 [cs]*, March 2016. arXiv: 1603.05435.

[129] K.R. Rao, D.N. Kim, and J.-J. Hwang. *Fast Fourier Transform - Algorithms and Applications*. Signals and Communication Technology. Springer Netherlands, Dordrecht, 2010.

[130] Myron J. Ross, Harry L. Shaffer, Asaf Cohen, Richard Freudberg, and Harold J. Manley. Average magnitude difference function pitch extractor. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 22(5):353—362, 1974.

[131] Thomas D Rossing, F Richard Moore, and Paul A Wheeler. *The science of sound*, volume 3. Addison Wesley San Francisco, 2002.

[132] Martin Rothenberg and James J. Mahshie. Monitoring Vocal Fold Abduction through Vocal Fold Contact Area. *Journal of Speech, Language, and Hearing Research*, 31(3):338–351, September 1988.

[133] David Rowetel. Codec2, 2011.

[134] Florian Schiel, Christian Heinrich, and Sabine Barfüsser. Alcohol language corpus: the first public corpus of alcoholized German speech. *Language Resources and Evaluation*, 46(3):503–521, September 2012.

[135] M. R. Schroeder. Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

[136] H. Schulzrinne and S. Casner. RTP Profile for Audio and Video Conferences with Minimal Control, 2003.

[137] Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, April 2011.

[138] M. P. Simunovic. Colour vision deficiency. *Eye*, 24(5):747–755, 2010.

[139] Julius O. Smith. Spectral audio signal processing. online book, 2011 edition. accessed 08 2016.

[140] Man Mohan Sondhi. New methods of pitch extraction. *Audio and Electroacoustics, IEEE Transactions on*, 16(2):262—266, 1968.

[141] Steven Spielberg. Indiana Jones and the last Crusade, 1989.

[142] Miroslav Stanek and Tomas Smatana. Comparison of fundamental frequency detection methods and introducing simple self-repairing algorithm for musical applications. In *2015 25th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 217–221, Pardubice, Czech Republic, April 2015. IEEE.

[143] Anthony P. Stark and Kuldip K. Paliwal. Speech analysis using instantaneous frequency deviation. In *INTERSPEECH*, pages 2602–2605. Citeseer, 2008.

[144] H J M Steeneken and F W M Geurtsen. DESCRIPTION OF THE RSG-10 NOISE DATA-BASE. *Report IZF*, page 12, 1988.

[145] Simon Stone, Peter Steiner, and Peter Birkholz. A Time-Warping Pitch Tracking Algorithm Considering Fast f0 Changes. In *Interspeech 2017*, pages 419–423. ISCA, August 2017.

[146] Sofia Strömbergsson. Today's Most Frequently Used F0 Estimation Methods, and Their Accuracy in Estimating Male and Female Pitch in Clean Speech. In *INTERSPEECH 2016*, pages 525–529, September 2016.

[147] Sofia Strömbergsson. Today's Most Frequently Used F0 Estimation Methods, and Their Accuracy in Estimating Male and Female Pitch in Clean Speech. In *INTERSPEECH 2016*, pages 525–529, September 2016.

[148] Lyudmila Sukhostat and Yadigar Imamverdiyev. A Comparative Analysis of Pitch Detection Methods Under the Influence of Different Noise Conditions. *Journal of Voice*, 29(4):410—417, July 2015.

[149] Xuejing Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, page I—333. IEEE, 2002.

[150] David Talkin. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495:518, 1995.

[151] Lee Ngee Tan and Abeer Alwan. Multi-band summary correlogram-based pitch detection for noisy speech. *Speech Communication*, 55(7-8):841–856, September 2013.

[152] Abhay Upadhyay and Ram Bilas Pachori. A new method for determination of instantaneous pitch frequency from speech signals. In *2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, pages 325–330, Salt Lake City, UT, USA, August 2015. IEEE.

[153] Luc M. Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with an auditory model. *The Journal of the Acoustical Society of America*, 91(6):3511–3526, 1992.

[154] B.D. Van Veen and K.M. Buckley. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, April 1988.

[155] Andrew Varga and Herman J.M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, July 1993.

[156] Peter Vary, Rudolf Hofmann, Karl Hellwig, and Robert J. Sluyter. A regular-pulse excited linear predictive codec. *Speech Communication*, 7(2):209–215, July 1988.

[157] E.F. Velez and R.G. Absher. Transient analysis of speech signals using the Wigner time-frequency representation. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2242–2245, Glasgow, UK, 1989. IEEE.

[158] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, editors. *Audio source separation and speech enhancement*. John Wiley & Sons, Hoboken, NJ, 2018.

[159] Kevin Walker and Stephanie Strassel. The RATS Radio Traffic Collection System. *Odyssey 2012*, page 7, 2012.

[160] DeLiang Wang and Guy J. Brown, editors. *Computational auditory scene analysis: principles, algorithms, and applications*. IEEE Press ; Wiley Interscience, Piscataway, N.J. : Hoboken, N.J, 2006. OCLC: ocm67870956.

[161] Dongmei Wang and John H. L. Hansen. F0 estimation for noisy speech by exploring temporal harmonic structures in local time frequency spectrum segment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6510–6514, Shanghai, March 2016. IEEE.

[162] Dongmei Wang, John H. L. Hansen, and Emily Tobey. F0 estimation for noisy speech based on exploring local time-frequency segment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, New Paltz, NY, USA, October 2015. IEEE.

[163] Dongmei Wang, Chengzhu Yu, and John H. L. Hansen. Robust Harmonic Features for Classification-Based Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):952–964, May 2017.

[164] Deso A. Weiss. The Pubertal Change of the Human Voice (Mutation). *Folia Phoniatrica et Logopaedica*, 2(3):126–159, 1950.

[165] J.G. Wilpon and C.N. Jacobsen. A study of speech recognition for children and the elderly. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 349–352, Atlanta, GA, USA, 1996. IEEE.

[166] Michael Wohlmayr and Franz Pernkopf. Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model. *INTERSPEECH 2009*, page 4, 2009.

[167] Bang Wong. Points of view: Color blindness. *nature methods*, 8(6):441–441, 2011.

[168] Alan Wrench. MOCHA MultiCHannel Articulatory database: English, November 1999.

[169] M. Wu, D. Wang, and G. J. Brown. A multi-pitch tracking algorithm for noisy speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–369–I–372, May 2002.

[170] Jichen Yang and Rohan Kumar Das. Low frequency frame-wise normalization over constant-Q transform for playback speech detection. *Digital Signal Processing*, 89:30–39, June 2019.

[171] Grace H. Yeni-Komshian, James F. Kavanagh, Charles Albert Ferguson, and National Institute of Child Health and Human Development, editors. *Child phonology. Vol. 1: Production*. Number 7 in Communicating by language. Acad. Press, New York, 1980. OCLC: 247391516.

[172] Stephen A Zahorian, Princy Dikshit, and Hongbing Hu. A Spectral-Temporal Method for Pitch Tracking. *INTERSPEECH 2006*, page 4, 2006.

[173] Stephen A. Zahorian and Hongbing Hu. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559, 2008.

[174] Xueliang Zhang, Hui Zhang, Shuai Nie, Guanglai Gao, and Wenju Liu. A Pairwise Algorithm Using the Deep Stacking Network for Speech Separation and Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1066–1078, June 2016.

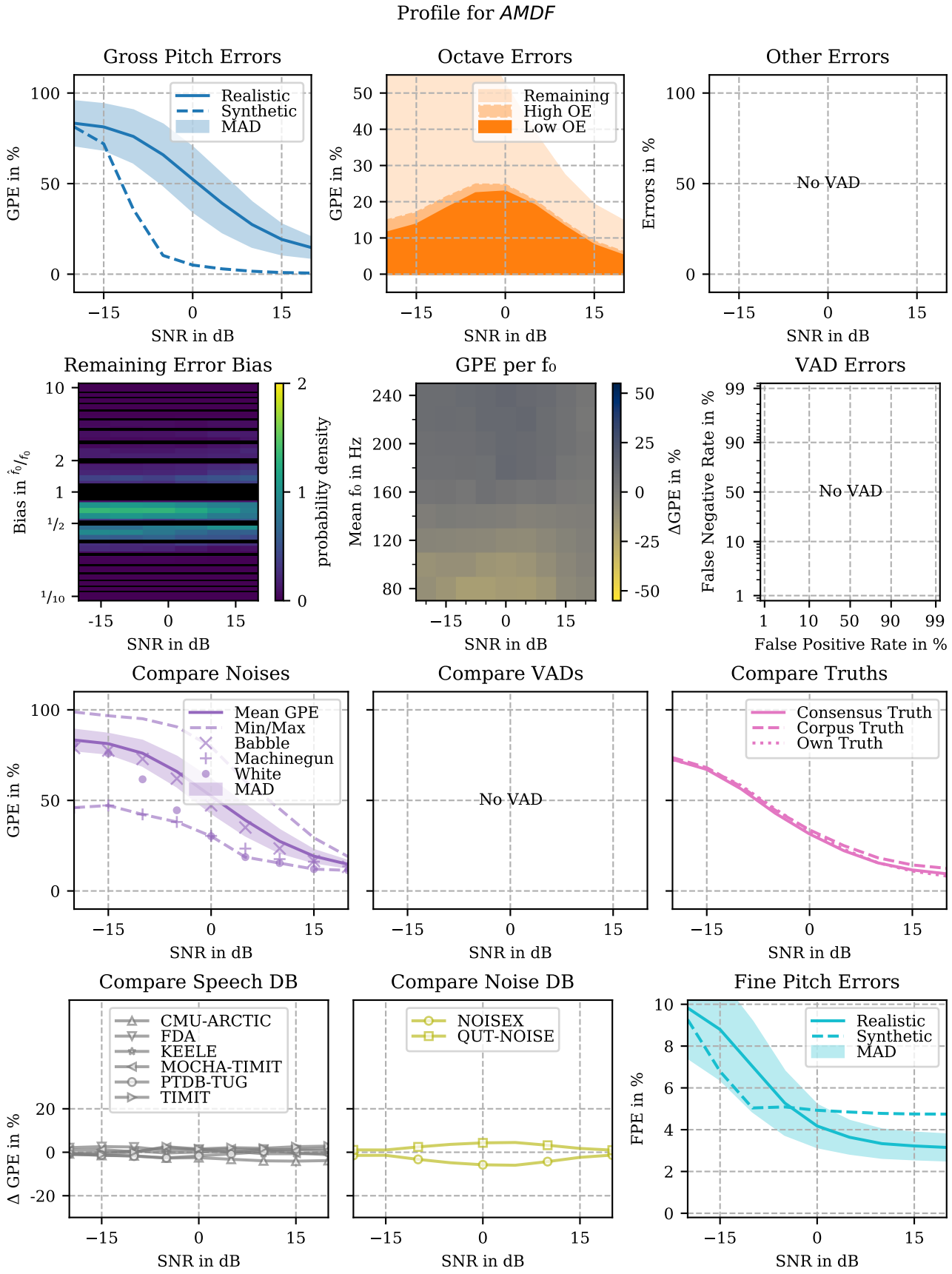
[175] V. W. Zue. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, 73(11):1602–1615, 1985.

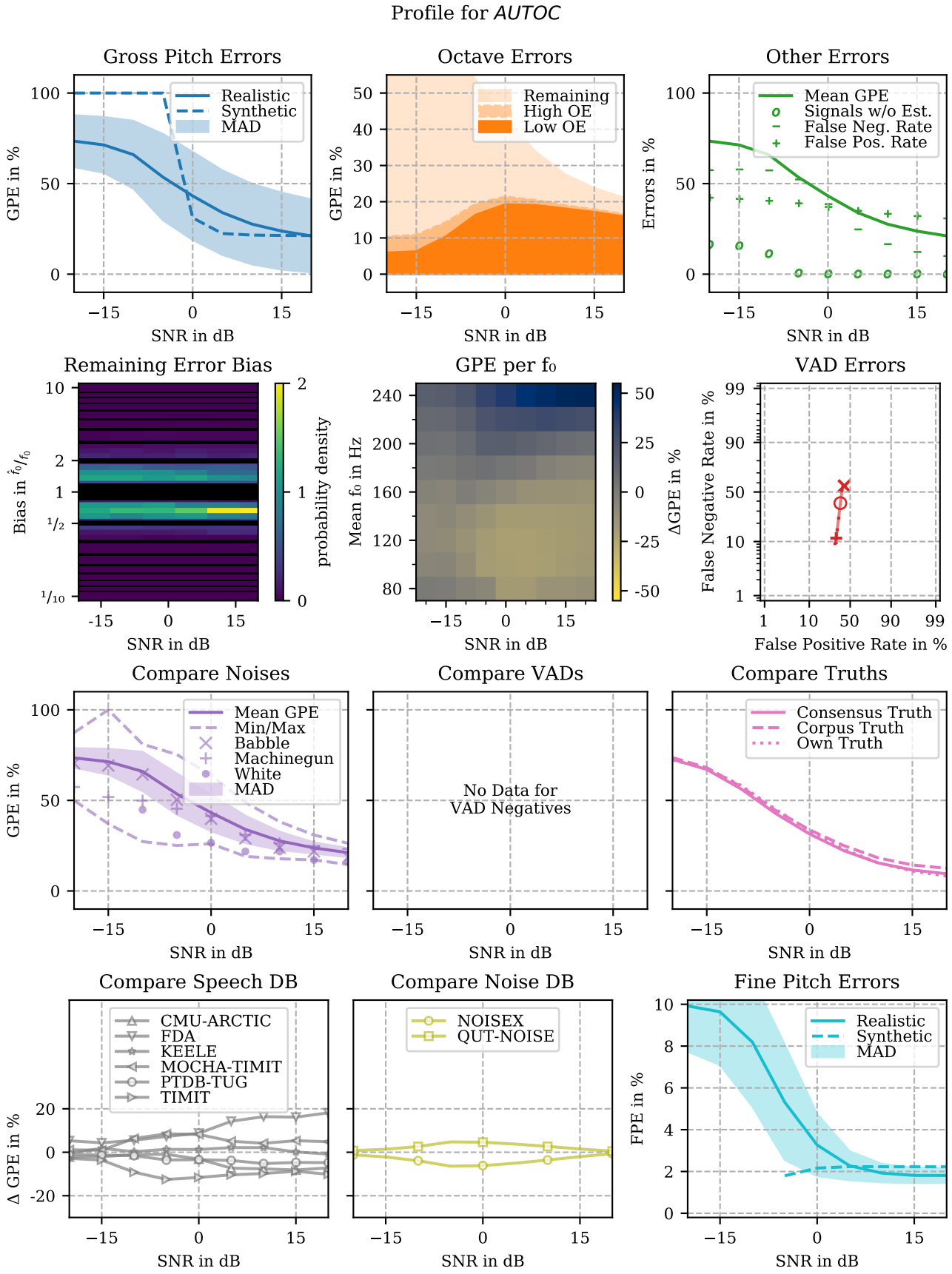
Part VII

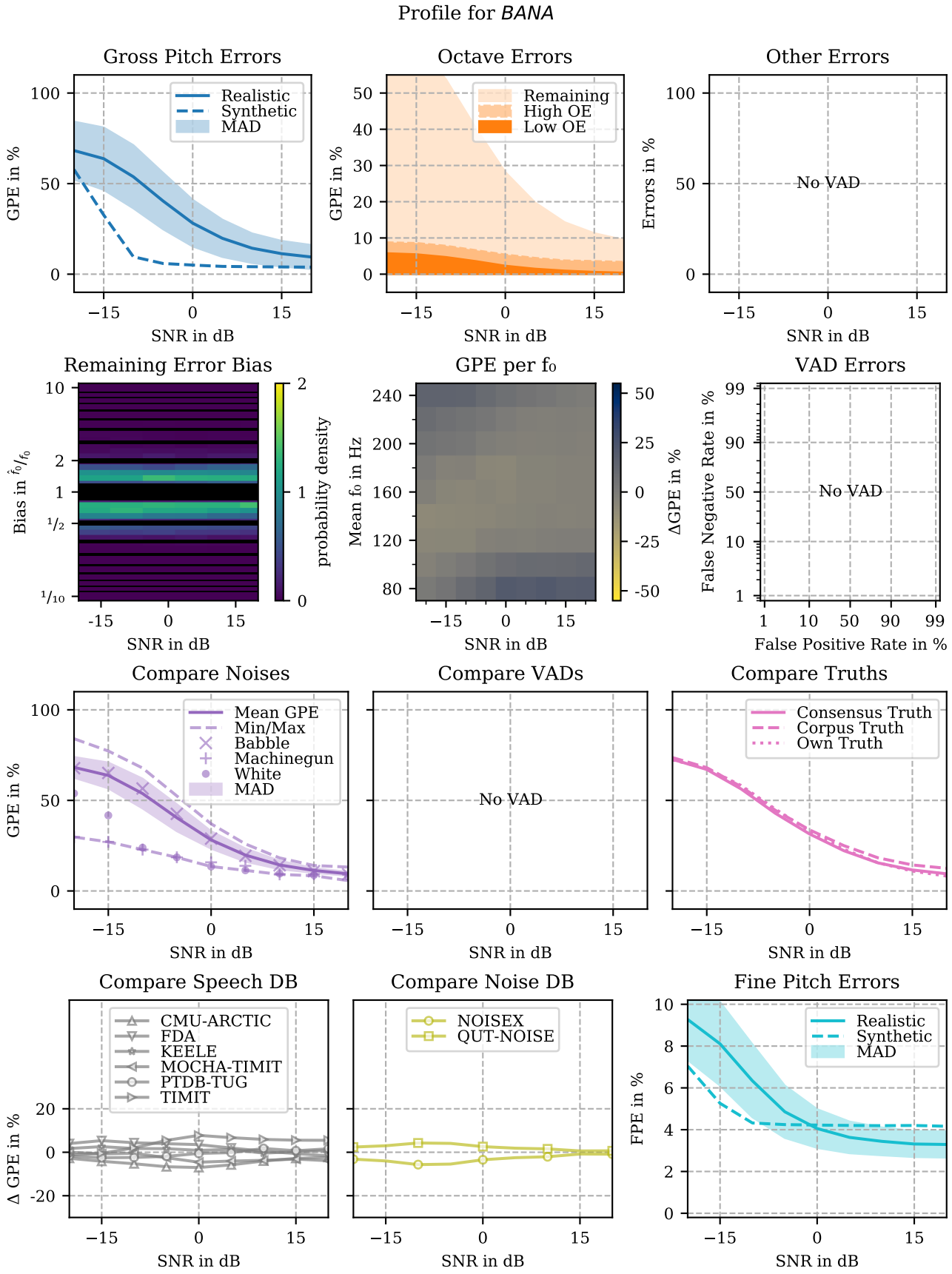
Appendix

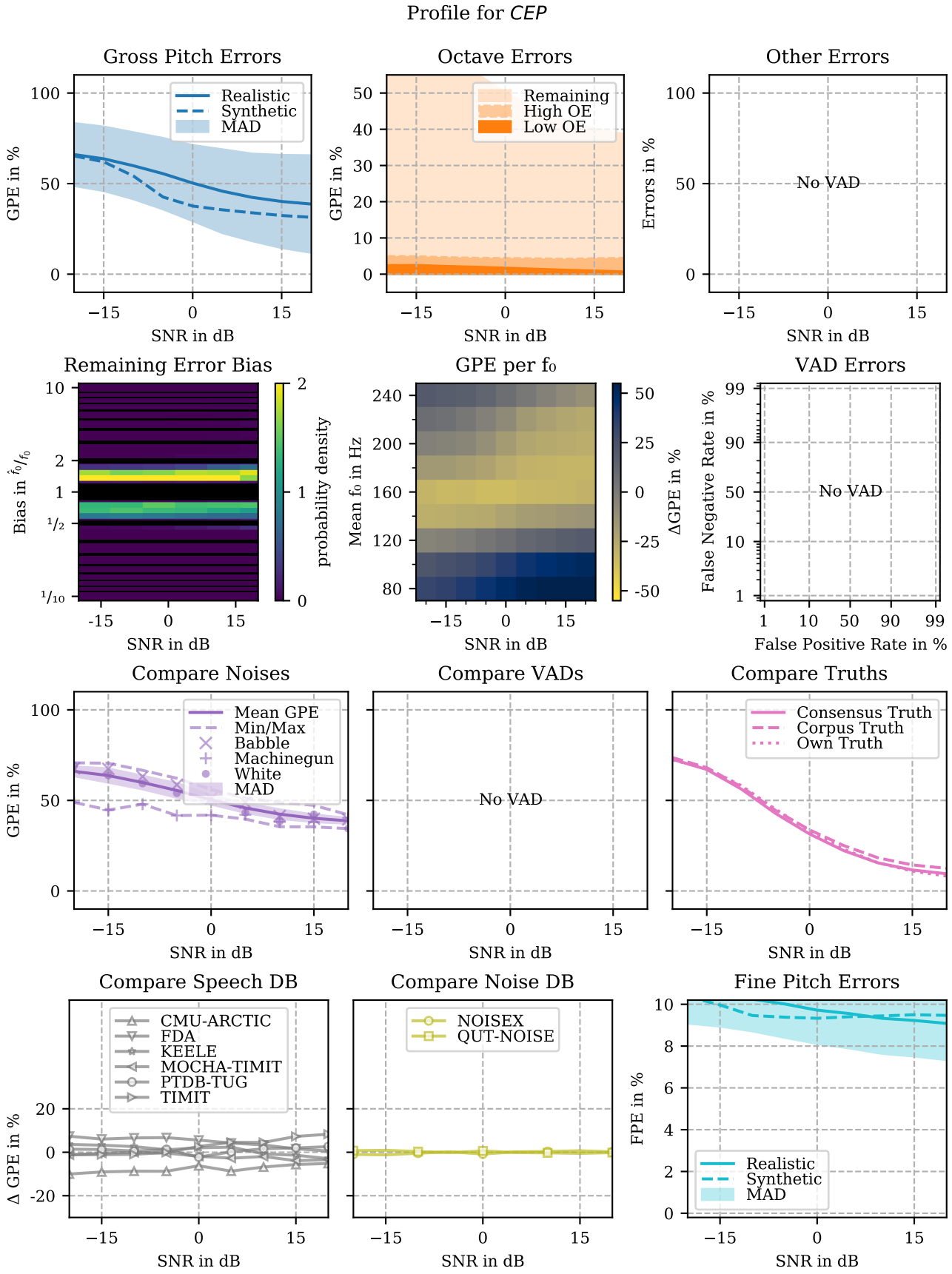
PDA Profiles

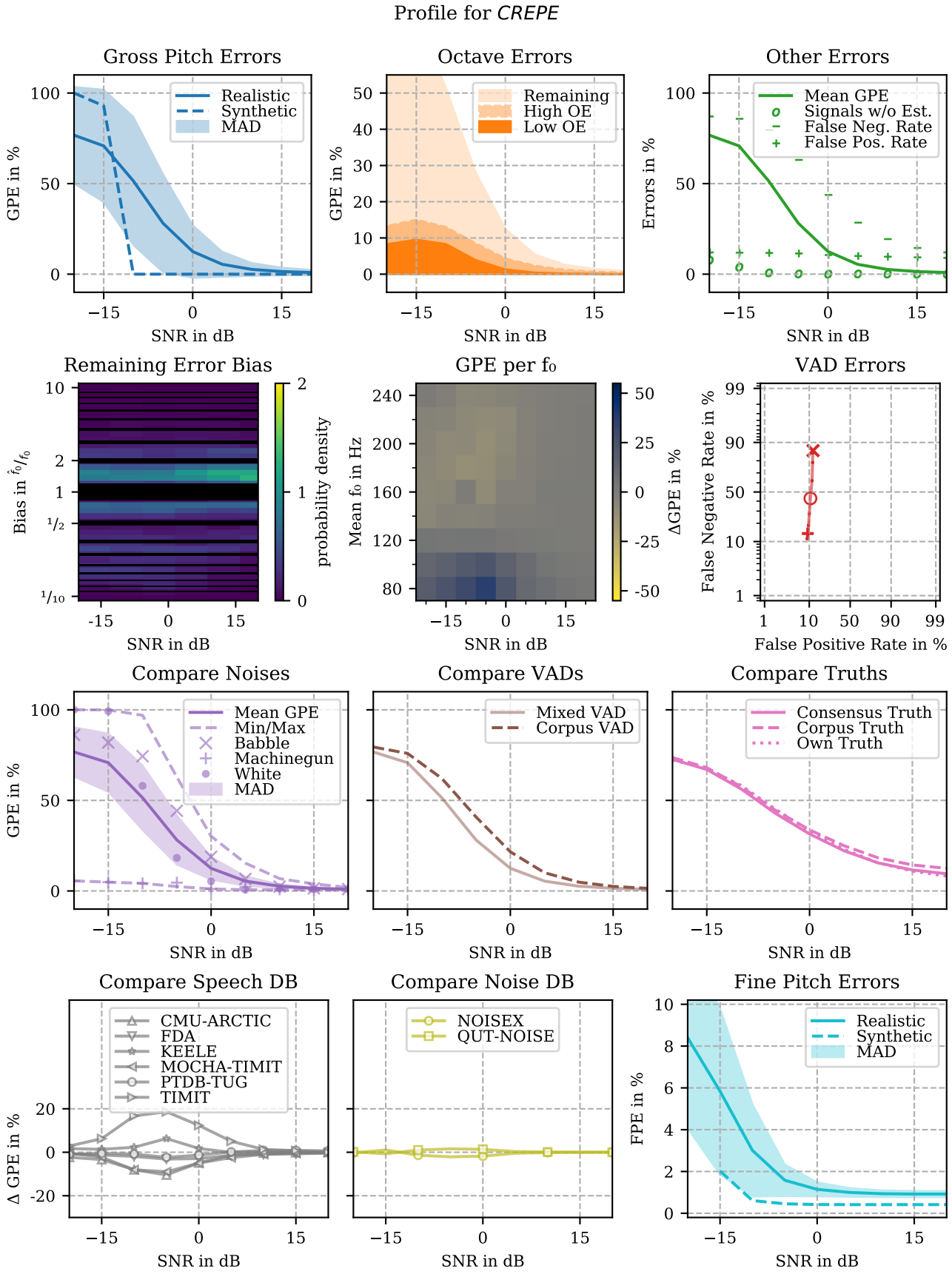
This Appendix includes a summary page for each PDA, with all the most important graphs on one page:

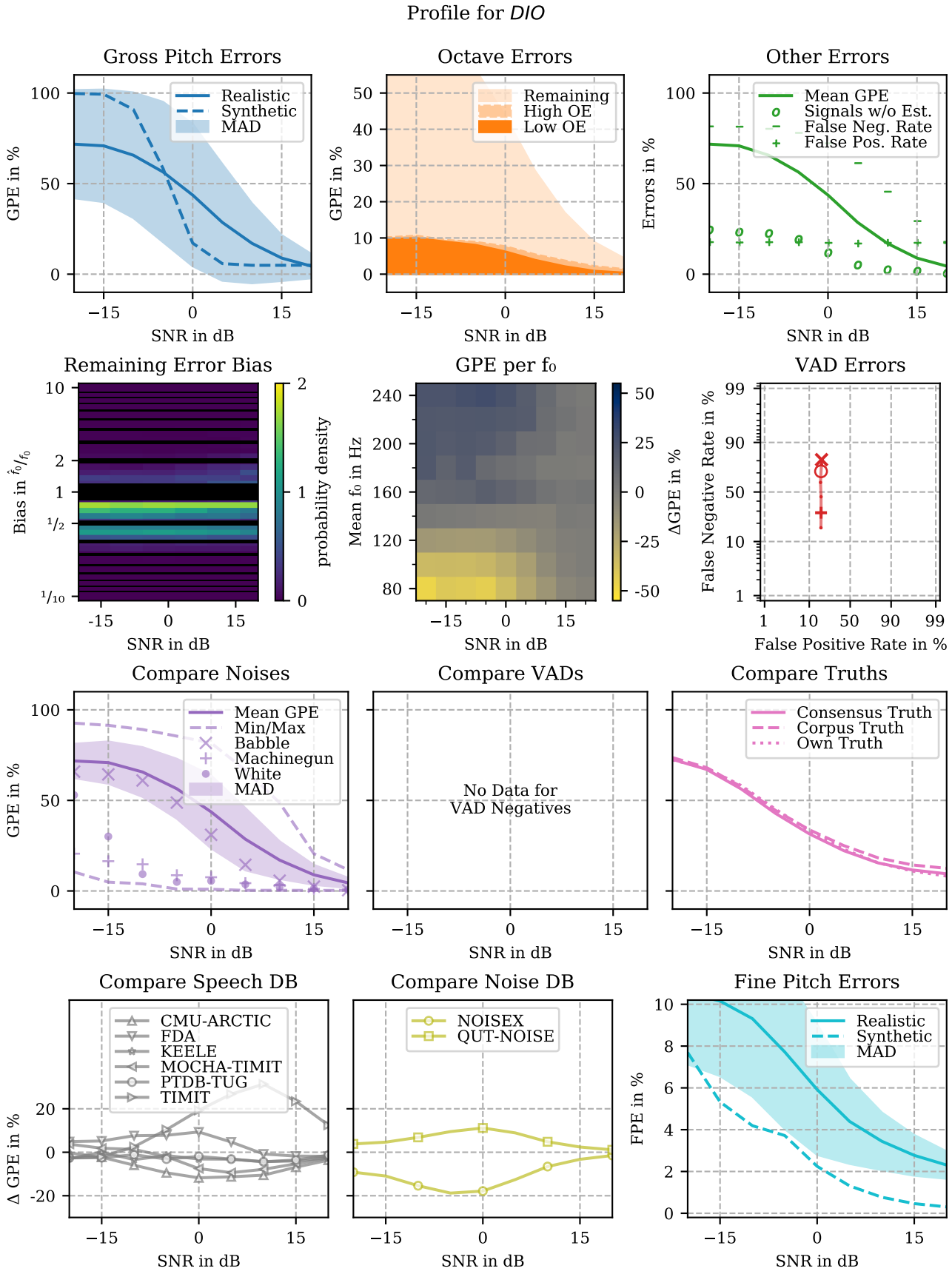

Profile for AMDF
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

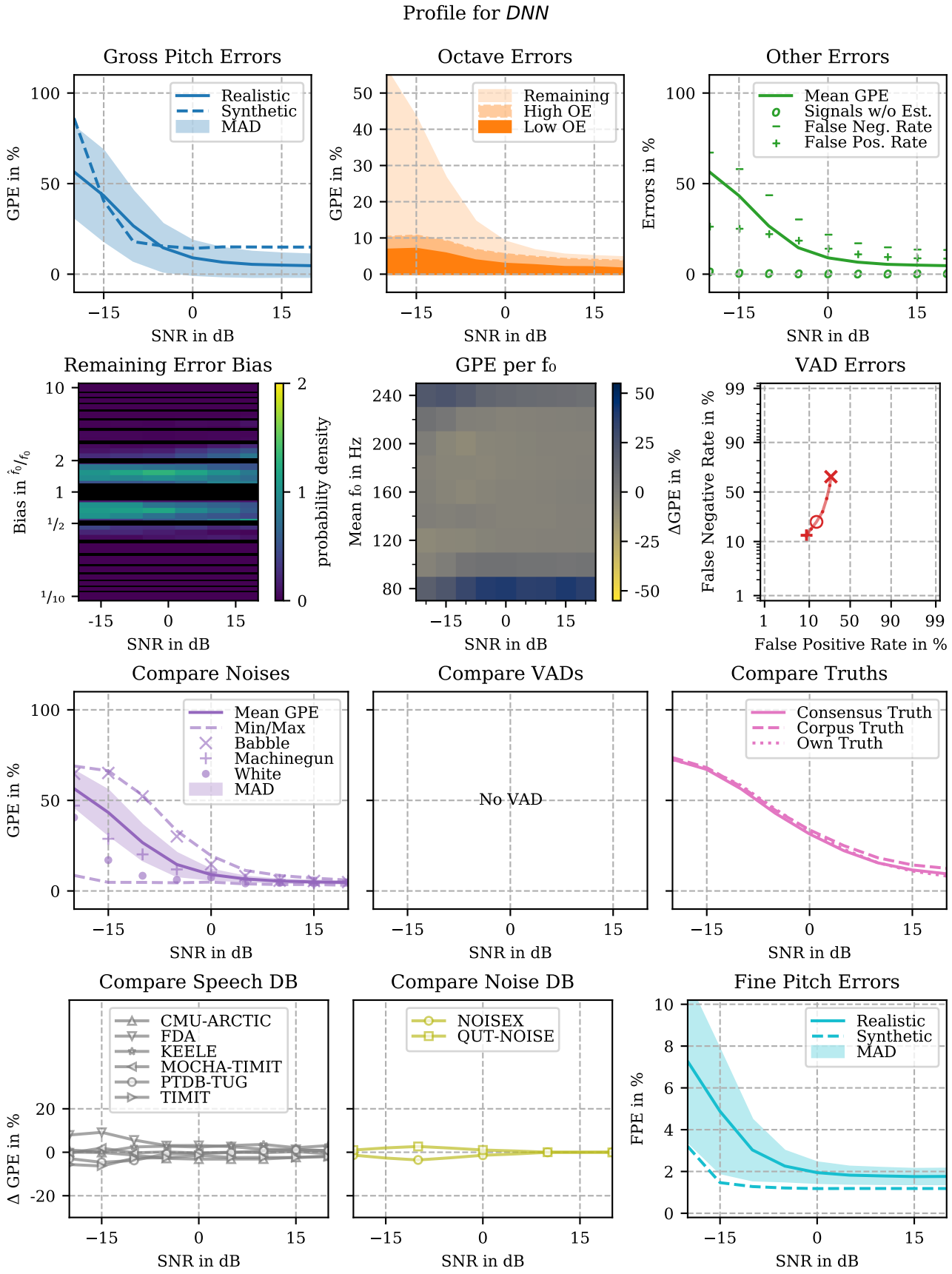

Profile for *AUTOC*
Gross Pitch Errors
GPE in %
SNR in dB
Realistic
Synthetic
MAD
Octave Errors
Remaining
High OE
Low OE
Other Errors
Errors in %
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
Compare VADs
No Data for
VAD Negatives
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
Compare Noise DB
NOISEX
QUT-NOISE
Fine Pitch Errors
FPE in %
Realistic
Synthetic
MAD


Profile for BANA
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

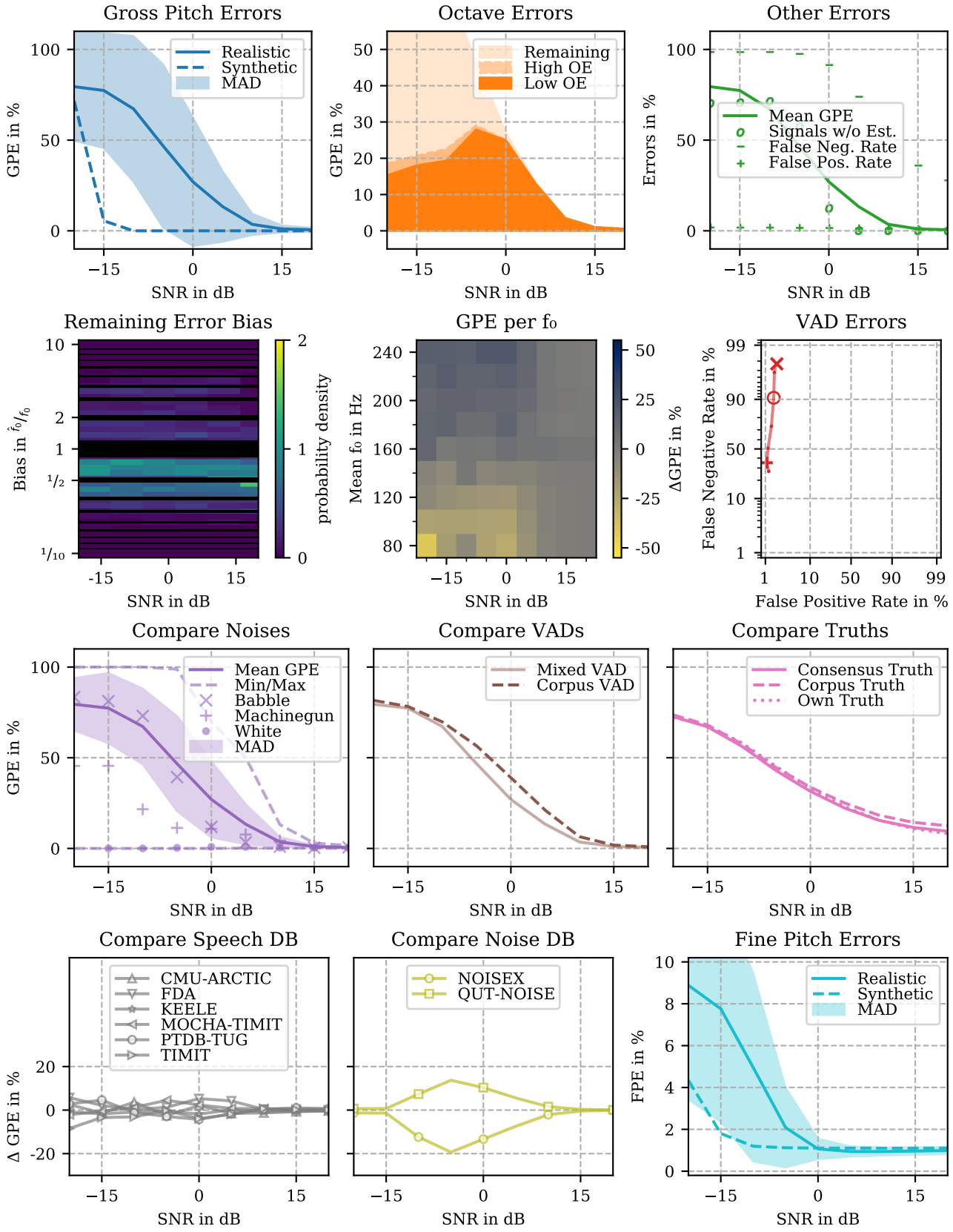

Profile for *CEP*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

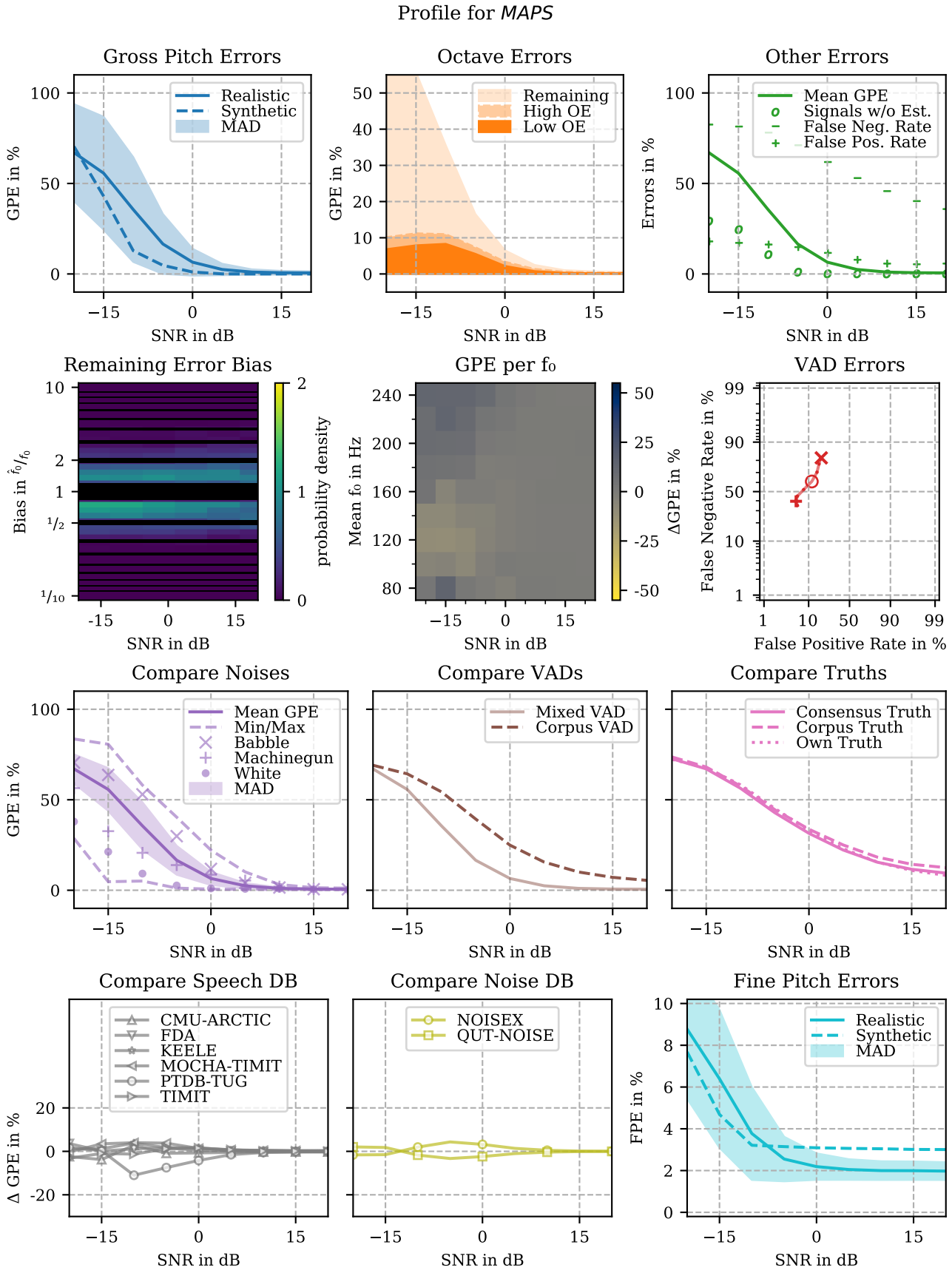

Profile for CREPE
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

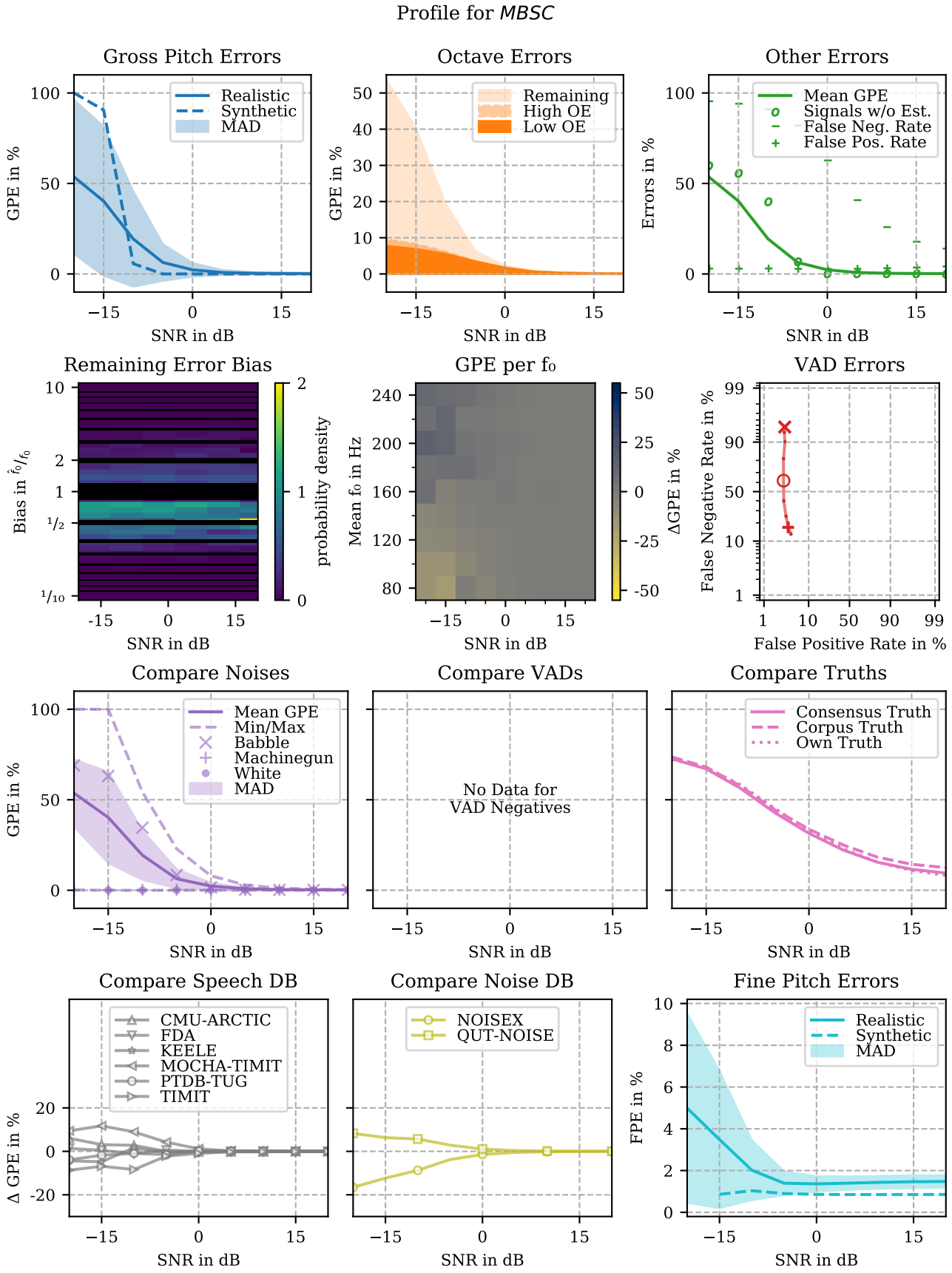

Profile for *DIO*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No Data for
VAD Negatives
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

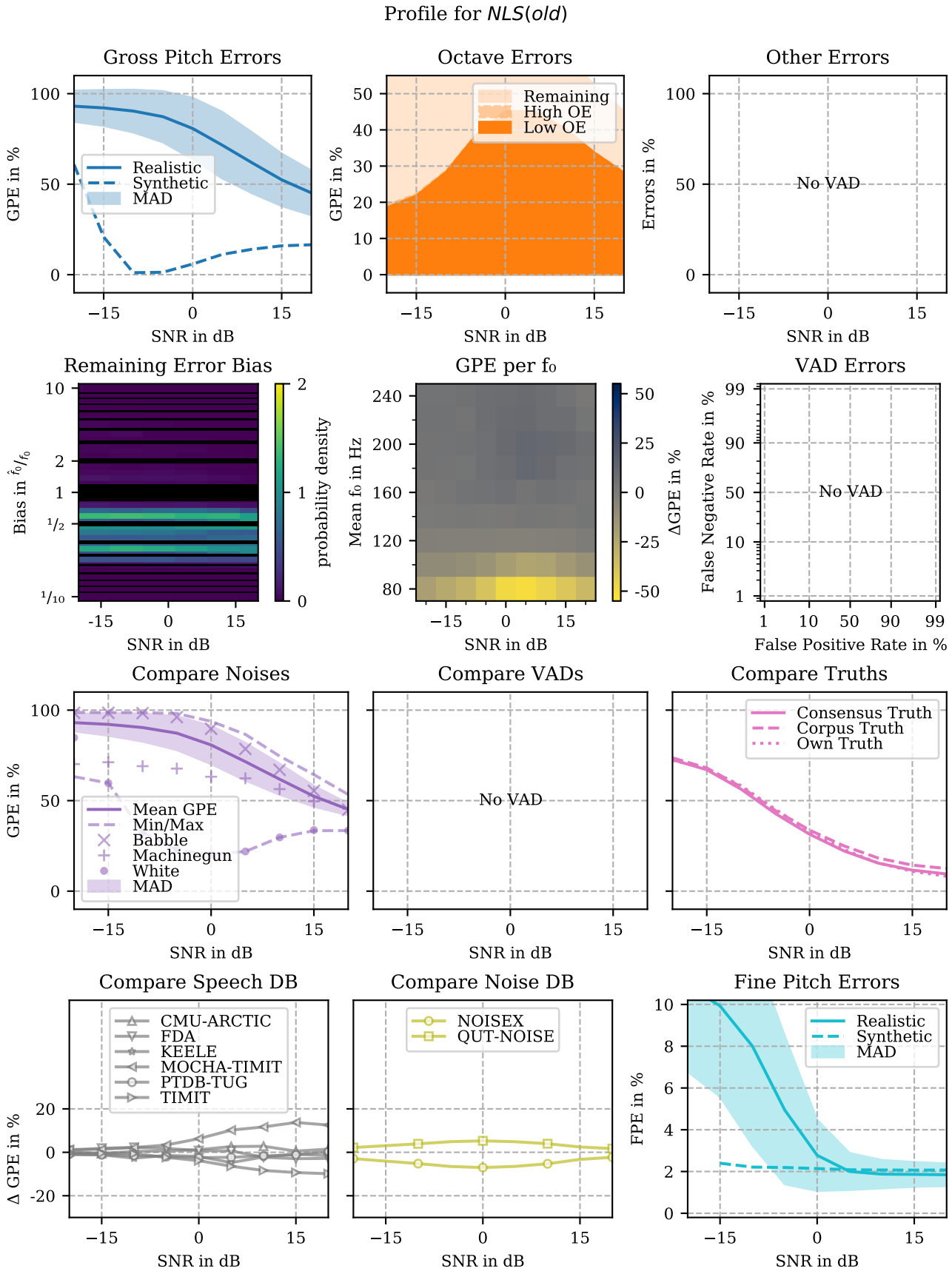

Profile for DNN
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

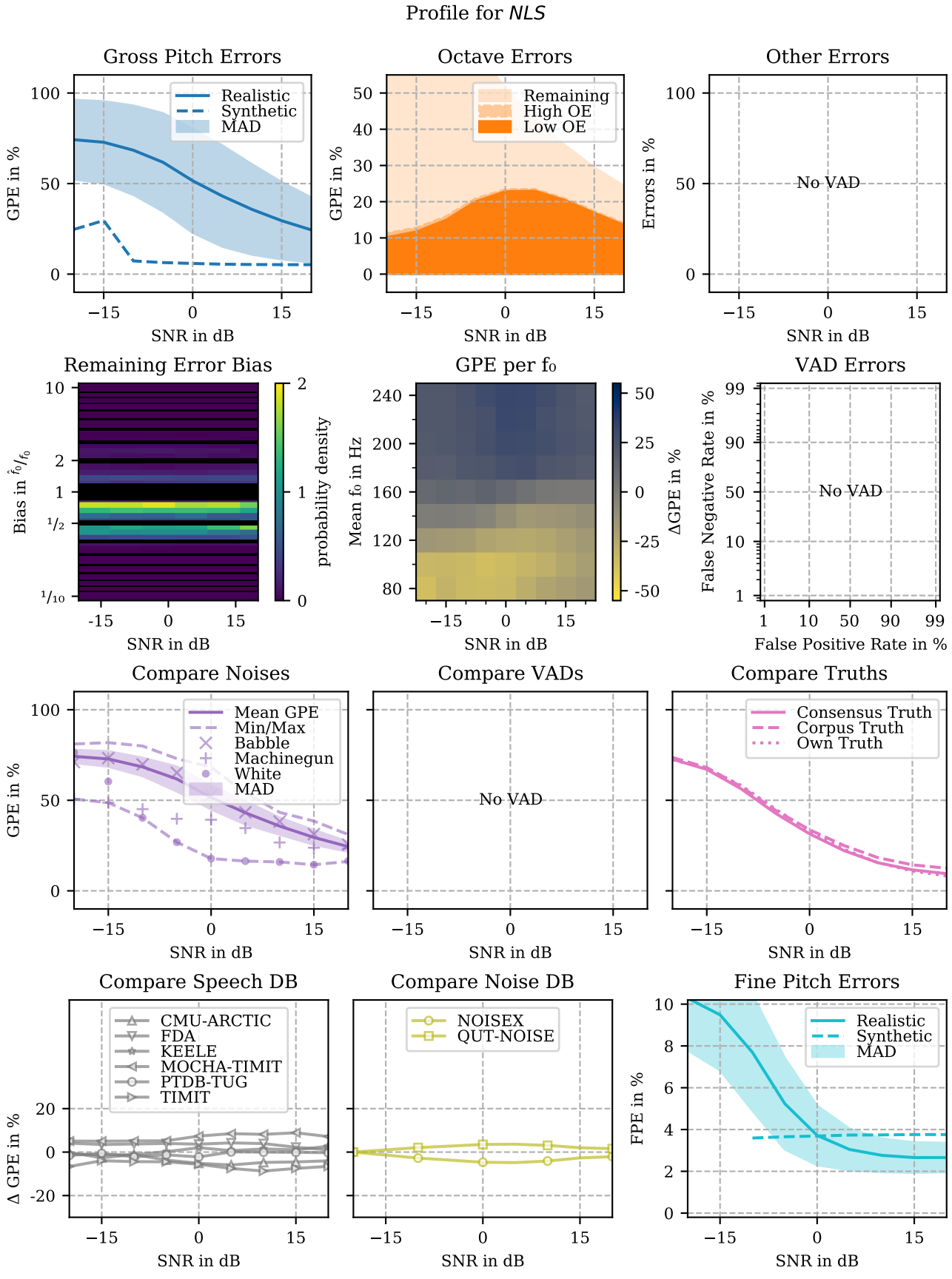
Profile for *KALDI*

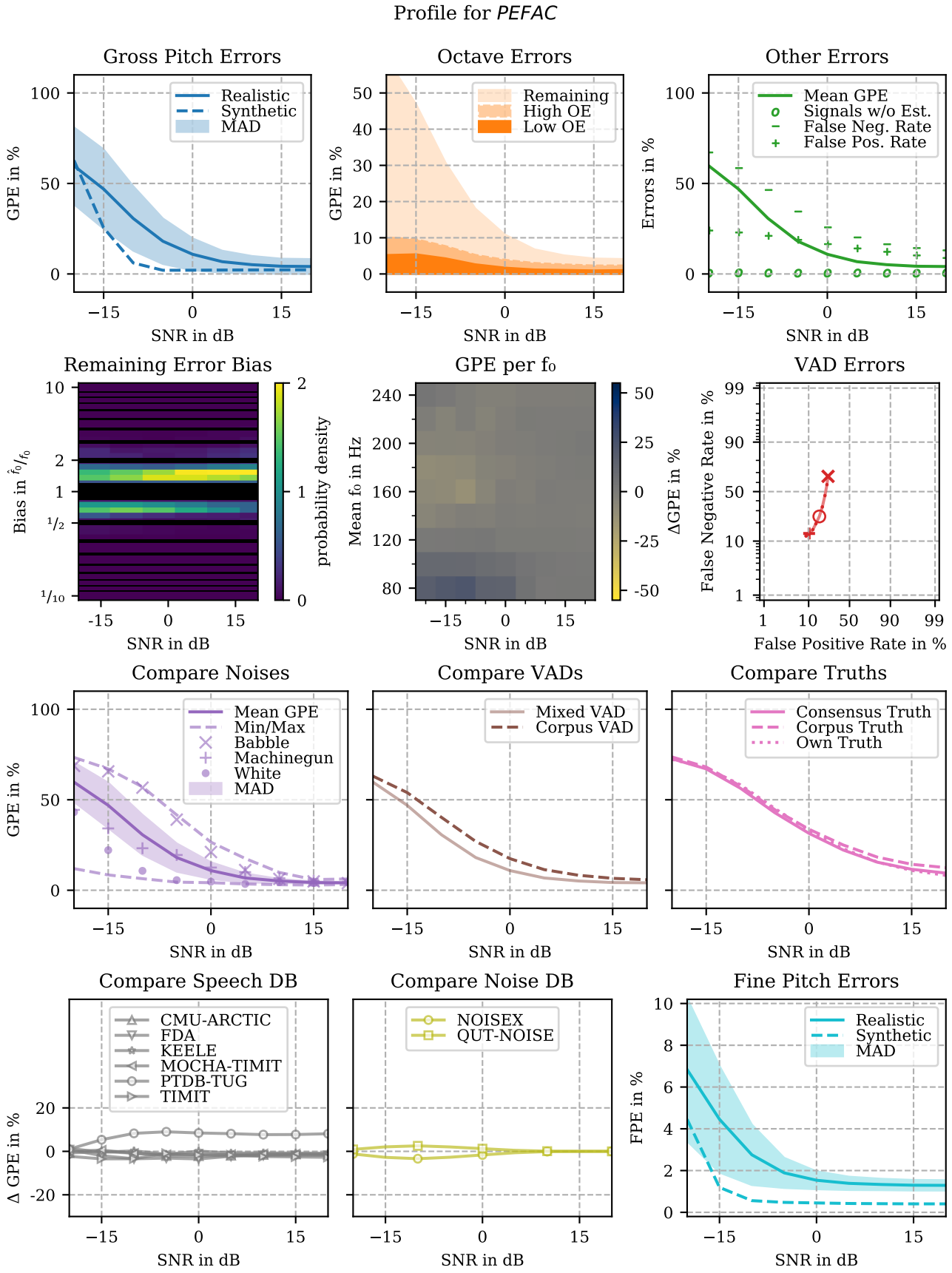



Profile for MAPS
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f0
Mean f0 in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

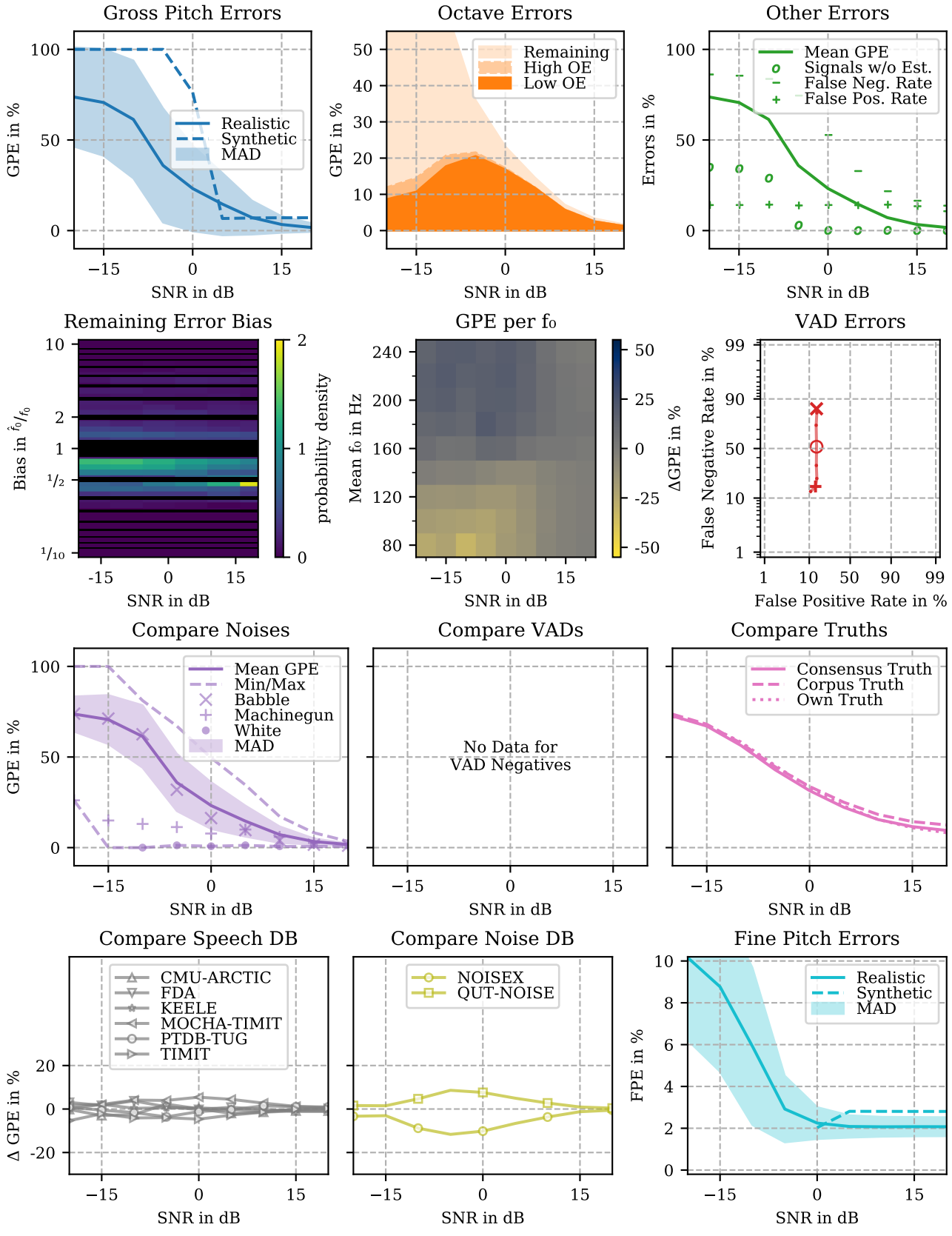

Profile for *MBSC*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No Data for
VAD Negatives
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

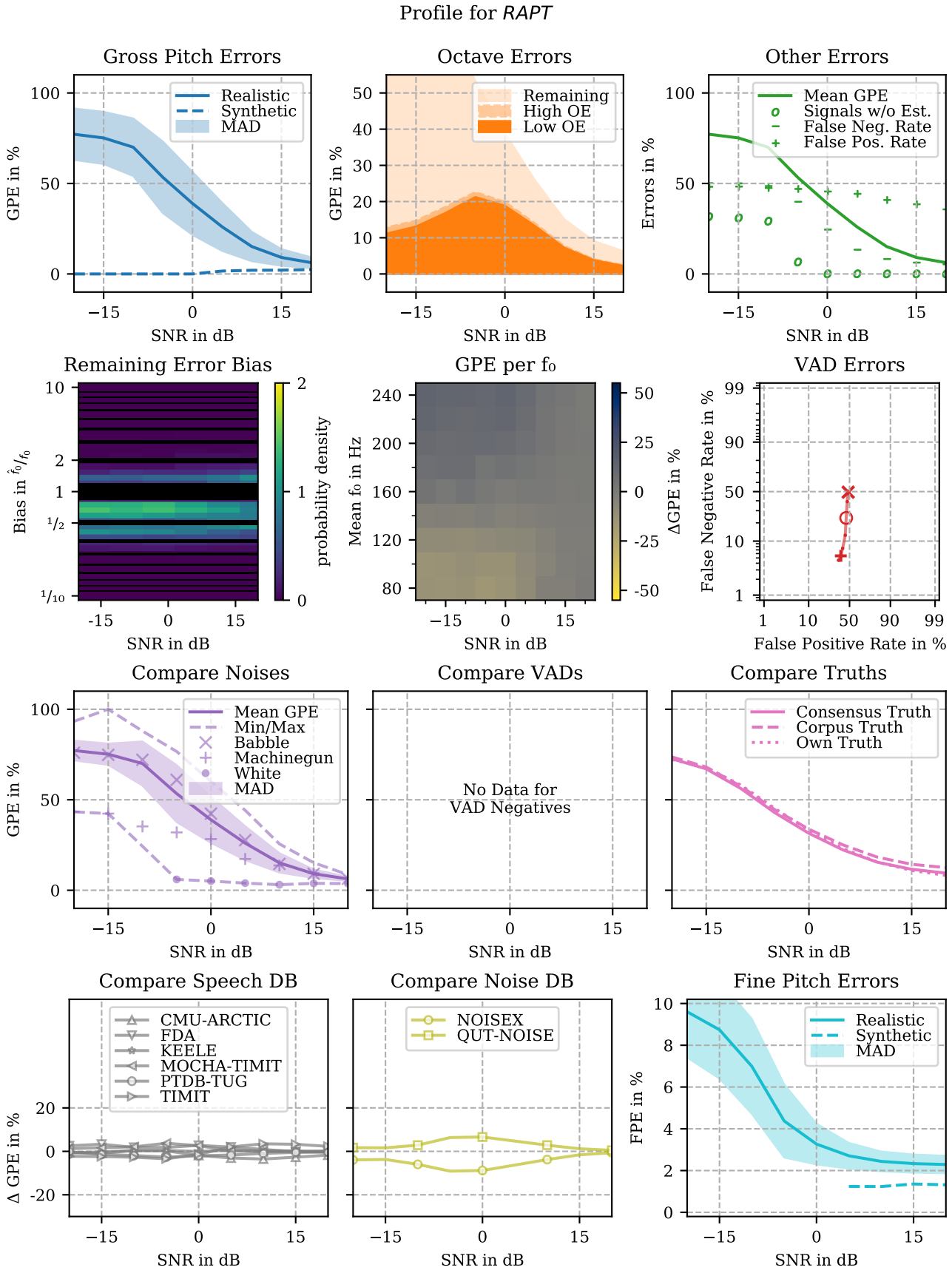

Profile for NLS(old)
Gross Pitch Errors
GPE in %
SNR in dB
Realistic
Synthetic
MAD
Octave Errors
Remaining
High OE
Low OE
Other Errors
Errors in %
No VAD
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
Compare VADs
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
Compare Noise DB
NOISEX
QUT-NOISE
Fine Pitch Errors
FPE in %

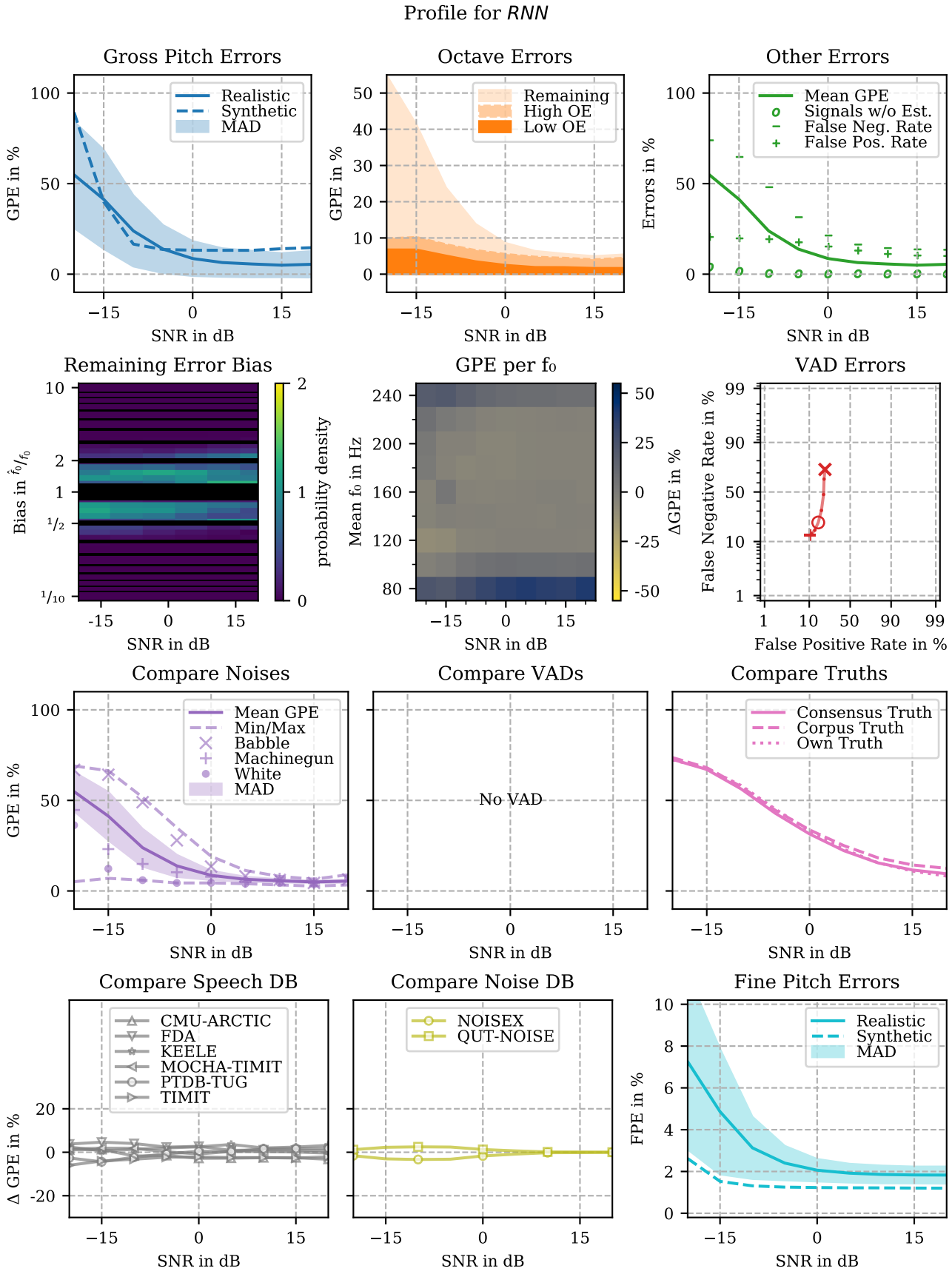

Profile for *NLS*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

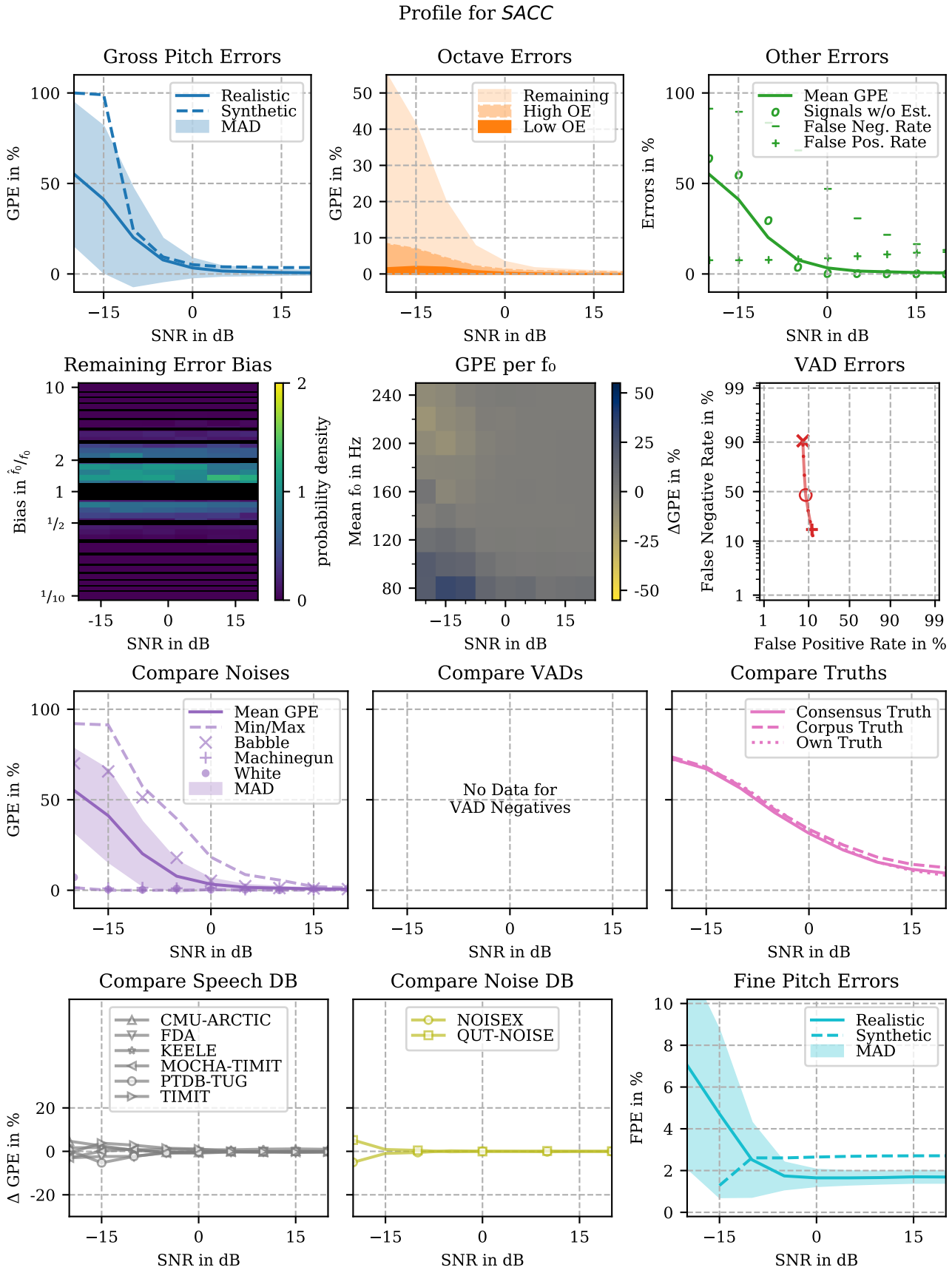

Profile for PEFAC
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

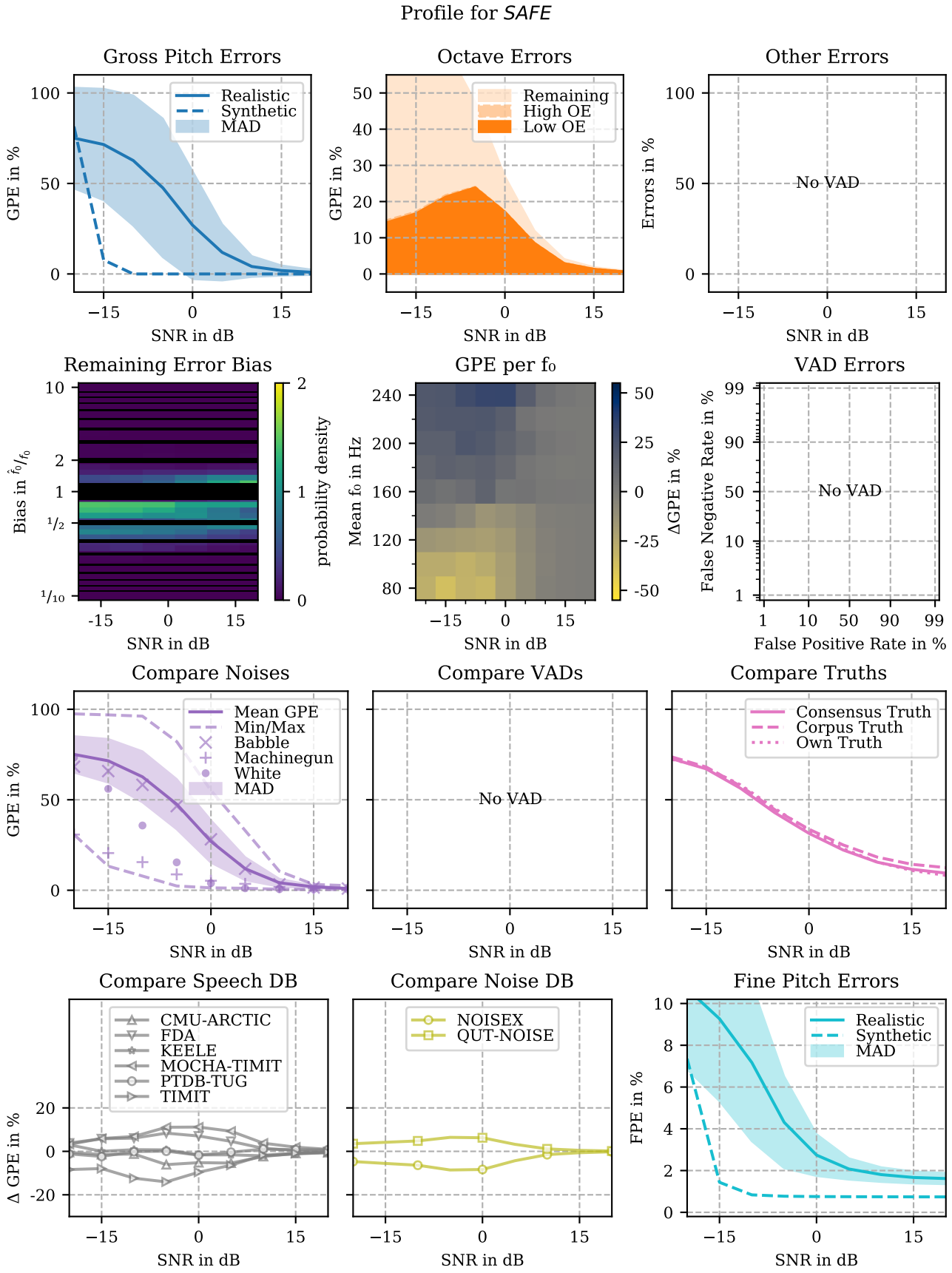
Profile for *PRAAT*

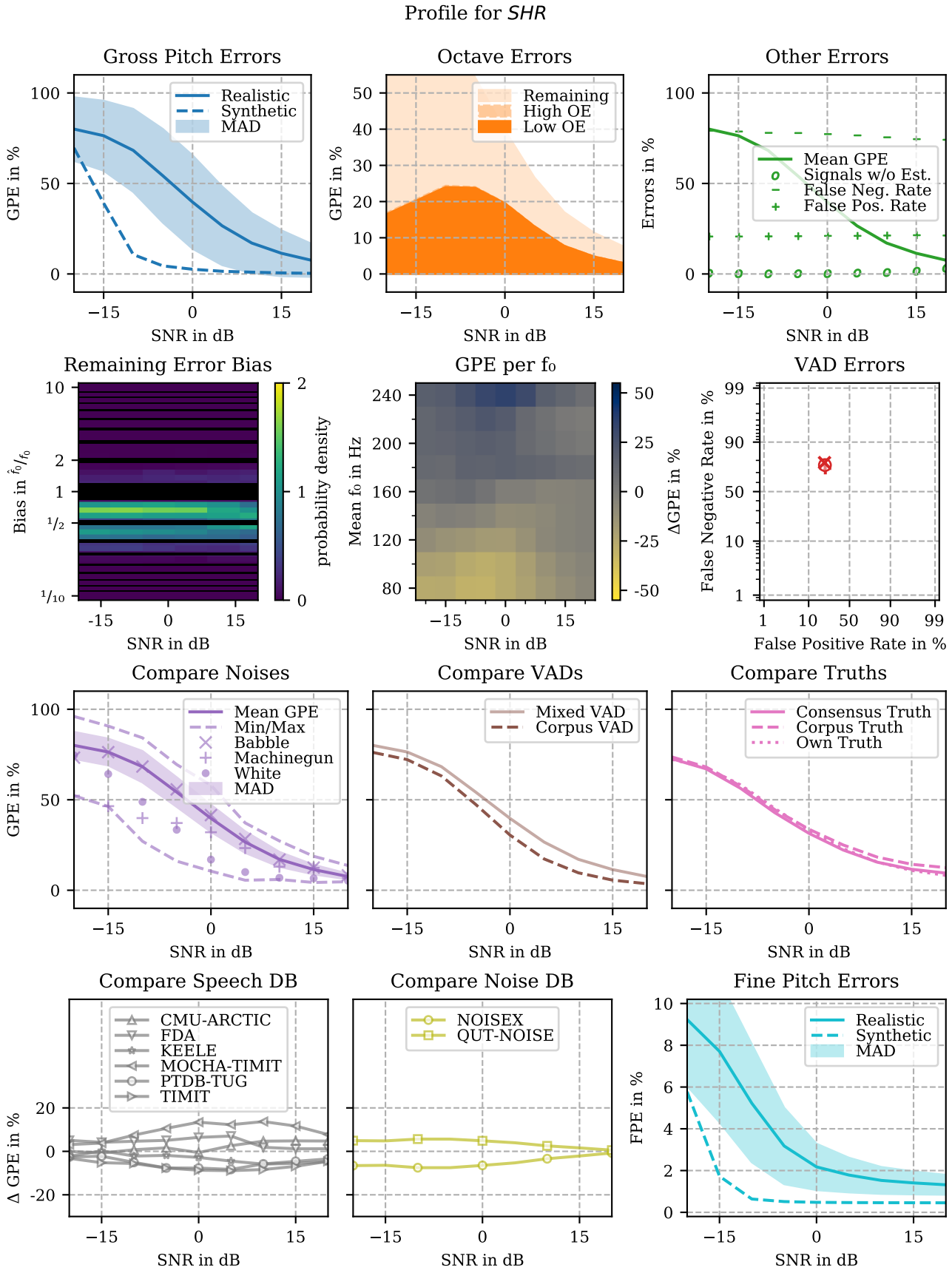


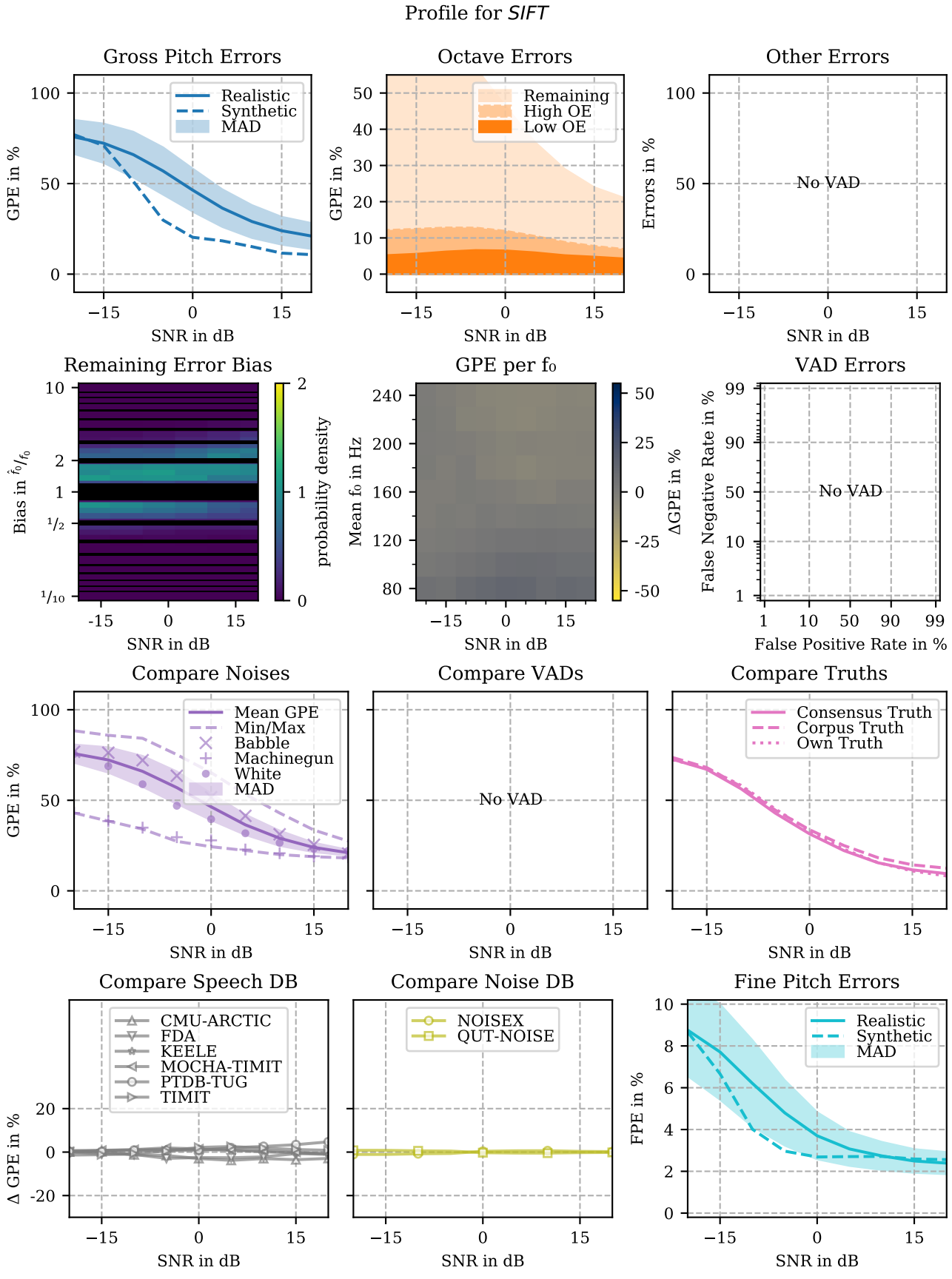

Profile for RAPT
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No Data for
VAD Negatives
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

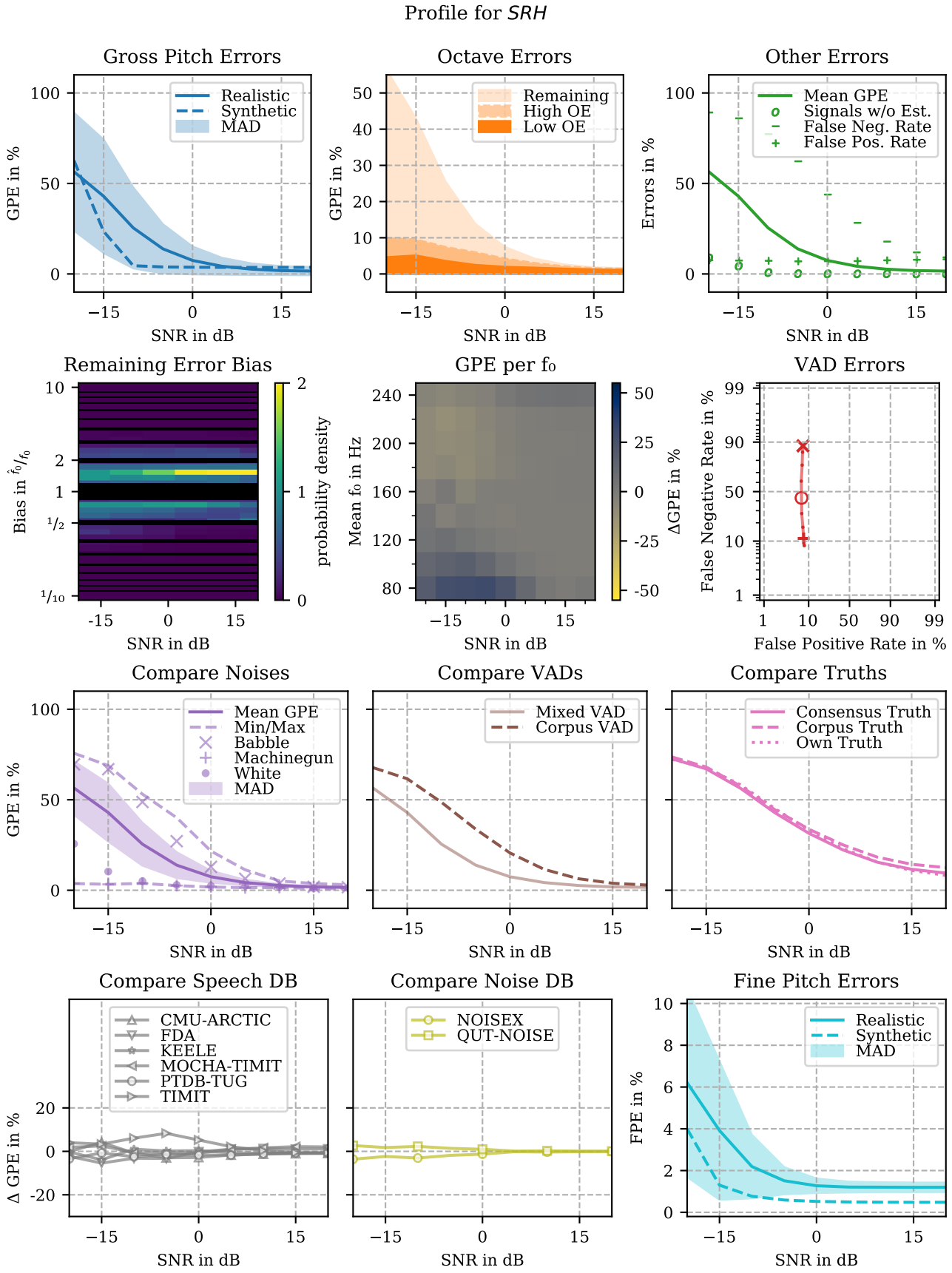

Profile for *RNN*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

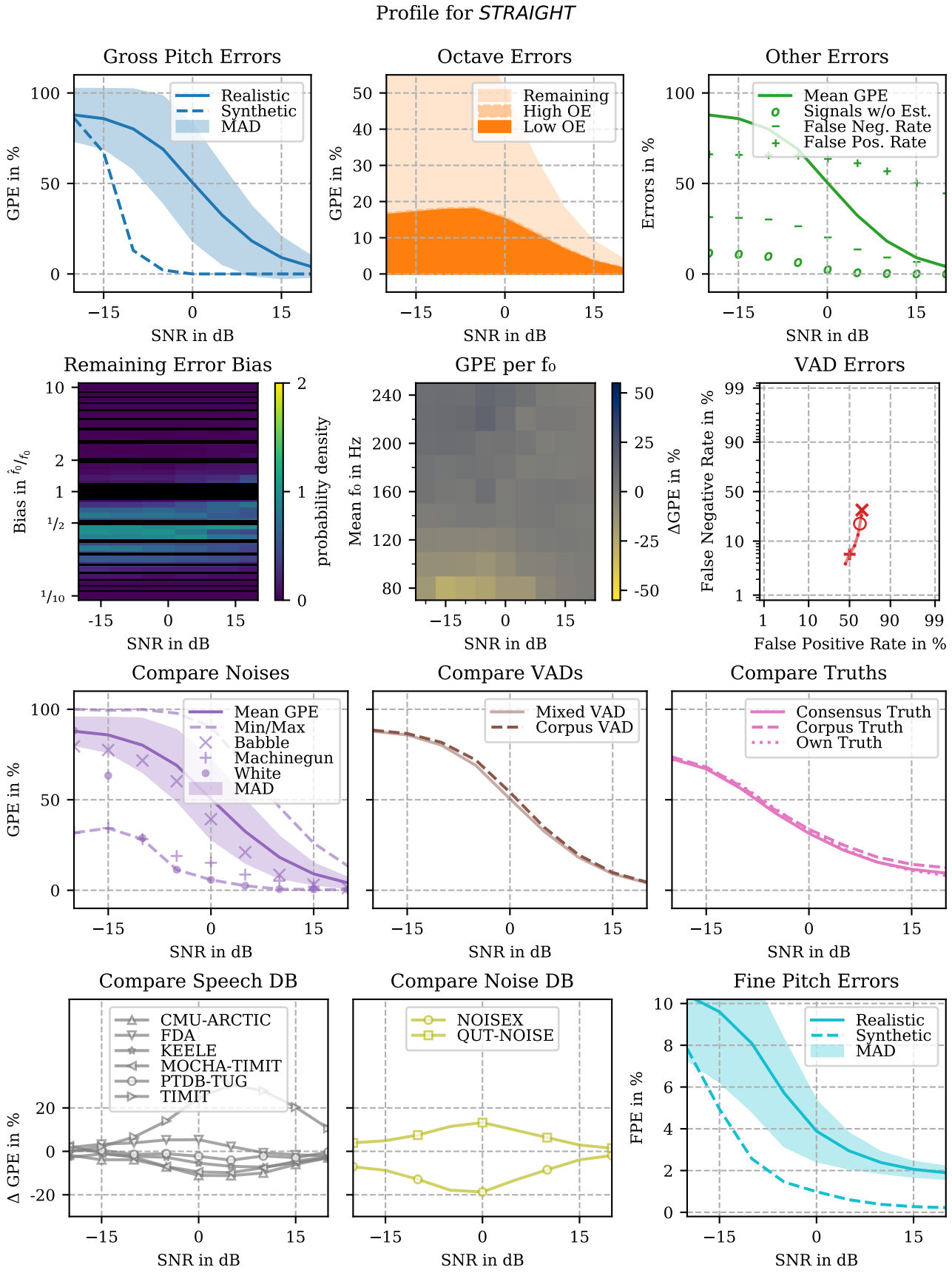

Profile for SACC
Gross Pitch Errors
GPE in %
SNR in dB
Realistic
Synthetic
MAD
Octave Errors
Remaining
High OE
Low OE
Other Errors
Errors in %
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
GPE per f0
Mean f0 in Hz
ΔGPE in %
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
Compare VADs
No Data for
VAD Negatives
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
Compare Noise DB
NOISEX
QUT-NOISE
Fine Pitch Errors
FPE in %
Realistic
Synthetic
MAD

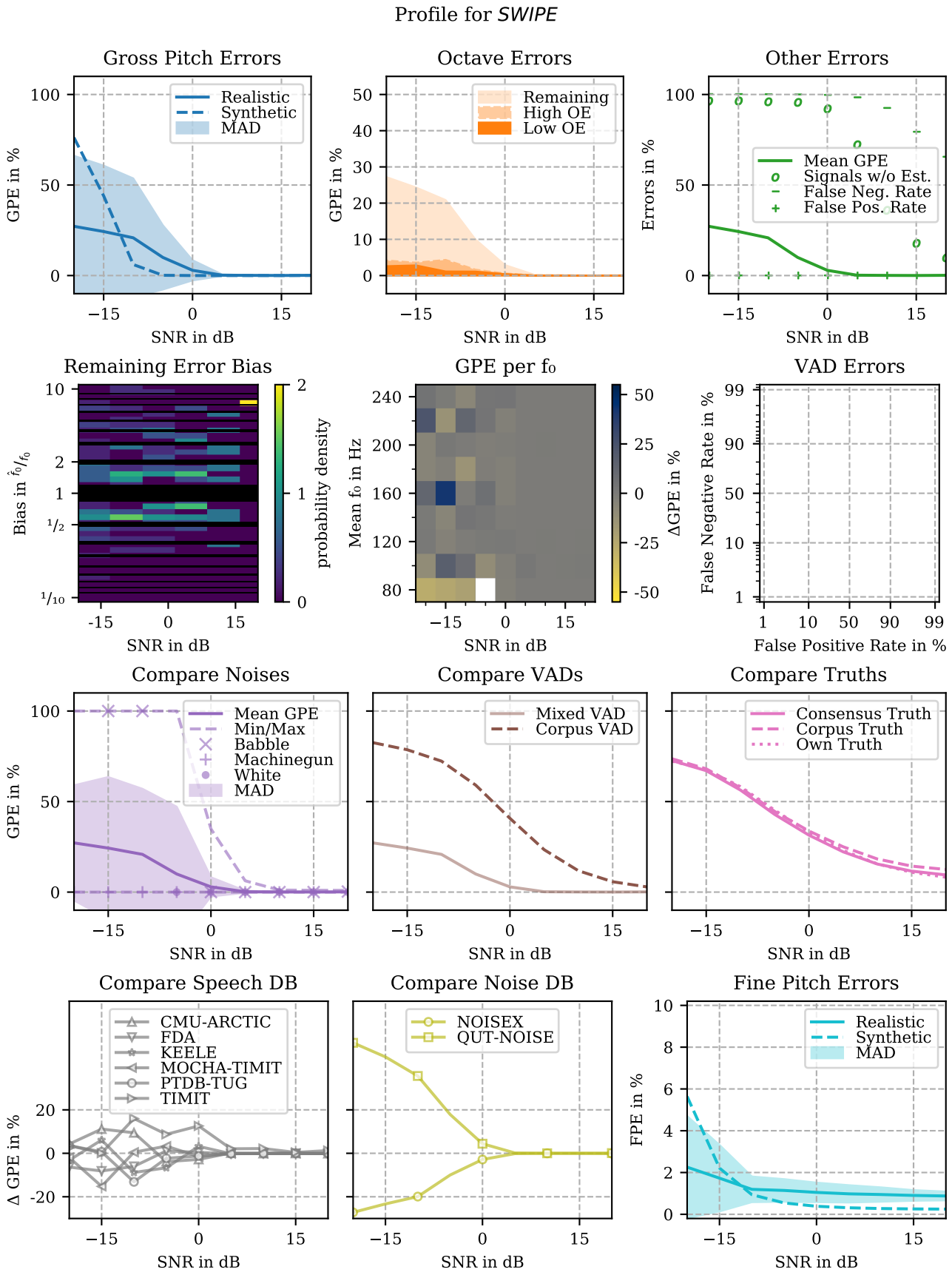

Profile for *SAFE*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f_0
Mean f_0 in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

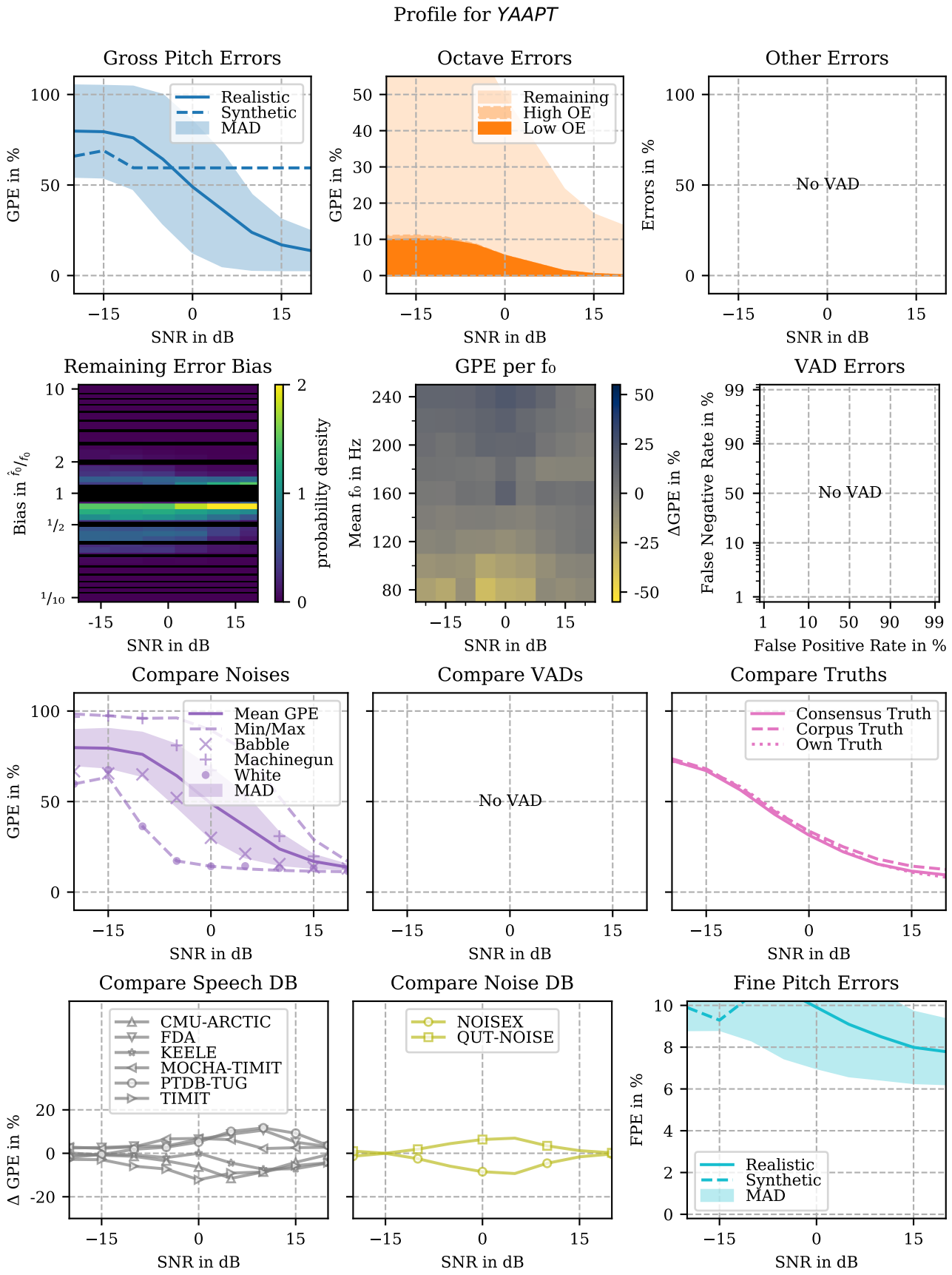

Profile for SHR
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

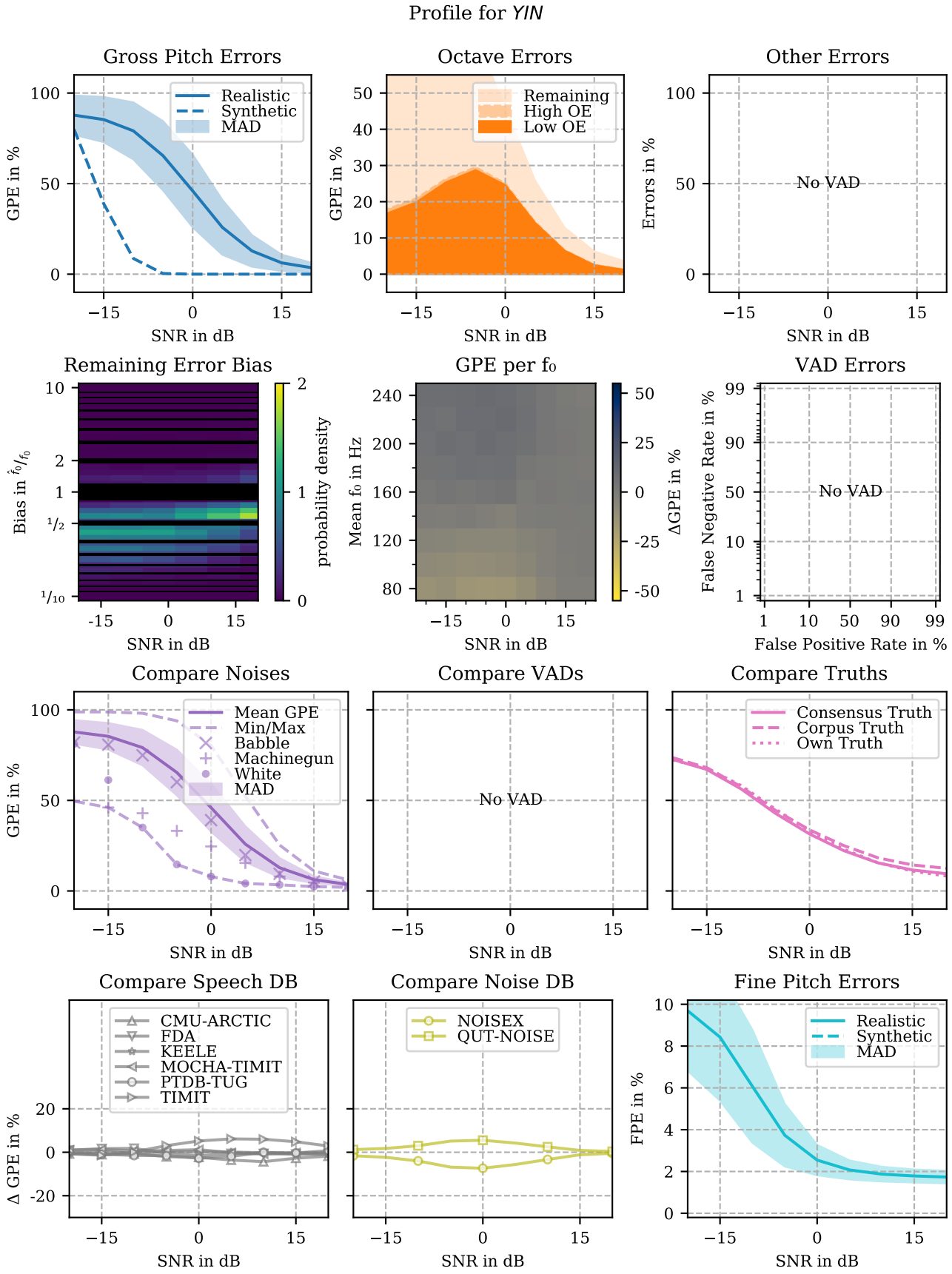

Profile for *SIFT*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB


Profile for SRH
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f0
Mean f0 in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB


Profile for *STRAIGHT*
Gross Pitch Errors
GPE in %
Realistic
Synthetic
MAD
SNR in dB
Octave Errors
GPE in %
Remaining
High OE
Low OE
SNR in dB
Other Errors
Errors in %
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
GPE in %
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
Δ GPE in %
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
FPE in %
Realistic
Synthetic
MAD
SNR in dB


Profile for SWIPE
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
Mean GPE
Signals w/o Est.
False Neg. Rate
False Pos. Rate
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
Mixed VAD
Corpus VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB


Profile for *YAAPT*
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB


Profile for YIN
Gross Pitch Errors
Realistic
Synthetic
MAD
GPE in %
SNR in dB
Octave Errors
Remaining
High OE
Low OE
GPE in %
SNR in dB
Other Errors
No VAD
Errors in %
SNR in dB
Remaining Error Bias
Bias in $\hat{f}_0/f_0$
probability density
SNR in dB
GPE per f₀
Mean f₀ in Hz
ΔGPE in %
SNR in dB
VAD Errors
No VAD
False Negative Rate in %
False Positive Rate in %
Compare Noises
Mean GPE
Min/Max
Babble
Machinegun
White
MAD
GPE in %
SNR in dB
Compare VADs
No VAD
SNR in dB
Compare Truths
Consensus Truth
Corpus Truth
Own Truth
SNR in dB
Compare Speech DB
CMU-ARCTIC
FDA
KEELE
MOCHA-TIMIT
PTDB-TUG
TIMIT
Δ GPE in %
SNR in dB
Compare Noise DB
NOISEX
QUT-NOISE
SNR in dB
Fine Pitch Errors
Realistic
Synthetic
MAD
FPE in %
SNR in dB

Source Code

MAPS

```
import scipy.signal
import numpy
from base64 import b64decode
from scipy.interpolate import RectBivariateSpline

_basefrequencies=numpy.logspace(numpy.log10(50), numpy.log10(450), 200)

def estimate_f0(blocks, basefrequencies=_basefrequencies):
    """A base frequency estimation in the magnitude and phase spectrum.

    MaPS-f0 uses a signal's magnitude STFT and it's IF deviation to
    estimate the maximum-likelihood base frequency of a tone complex
    for a given series of base frequencies.

    blocks: A SignalBlocks instance. Must use a samplerate of 48 kHz.
    basefrequencies: an ordered vector of base frequencies in Hz.

    Returns a vector of times, a vector of likely base frequencies,
    and a vector of their likelihood.

    """
    probability = pitch_probability(blocks, basefrequencies)
    return max_track_viterbi(probability, blocks.hopsize/blocks.samplerate, basefrequencies)


def pitch_probability(blocks, basefrequencies=_basefrequencies):
    """A probability-of-pitch estimation in the magnitude and phase spectrum.

    MaPS-f0 uses a signal's magnitude STFT and it's IF deviation to
    estimate the likelihood that a given range of base frequencies are
    base frequencies of a tone complex.

    blocks: A SignalBlocks instance.
    basefrequencies: an ordered vector of base frequencies in Hz.

    Returns a len(blocks) x len(basefrequencies) matrix of likelihood values.

    """

    mc = magnitude_correlation(blocks, basefrequencies)
    ifdd = ifd_difference(blocks, basefrequencies)
    return value2posterior(mc, ifdd)


def max_track(probabilities, delta_t, basefrequencies=_basefrequencies):
```

```
    """Extract frequency track from probabilities.

    The frequency track is located at the maximum likelihood per time
    frame.

    probabilities: a time x frequency matrix of pitch probabilities.
    delta_t: the time distance between two likelihood frames.
    basefrequencies: The second axis of probabilities in Hz.

    Returns the time of each frame, it's most likely frequency, and
    the frequency's likelihood.

    """
    time = numpy.arange(len(probabilities))*delta_t
    freq = basefrequencies[numpy.argmax(probabilities, axis=1)]
    prob = numpy.max(probabilities, axis=1)
    return time, freq, prob


def max_track_viterbi(probabilities, delta_t, basefrequencies):
    """Extract frequency track from probabilities.

    The frequency track is the track of maximum probability with a
    minimum of frequency steps. Frequency steps are penalized
    propertionally to the multiplicative step size.

    probabilities: a time x frequency matrix of pitch probabilities.
    frequency: The second axis of probabilities in Hz.

    Returns each frame's most likely frequency, and the frequency's
    probability.

    """

    time = numpy.arange(len(probabilities))*delta_t

    # transition probability between two frequencies is the quotient
    # between those frequencies, normalized to < 1.
    transition = basefrequencies[:, None] / basefrequencies[None, :]
    transition[transition>1] = 1/transition[transition>1]

    # accumulate probabilities for each time step and select the
    # highest cumulative probability path per basefrequencies and time:
    cum_probability = probabilities.copy()
    # save step that lead to this time/basefrequencies:
    idx_probability = numpy.empty(probabilities.shape, dtype=int)
    for idx in range(1, len(probabilities)):
        step_probs = cum_probability[idx-1][:,None] * probabilities[idx][None,:] * transition
        max_prob_idx = step_probs.argmax(axis=0)
        idx_probability[idx] = max_prob_idx
        cum_probability[idx] = step_probs[[max_prob_idx, numpy.arange(len(basefrequencies))]]
        # normalize, so large products of small numbers don't end up zero
        cum_probability[idx] /= cum_probability[idx].mean()

    # walk backwards and select the path that lead to the maximum
    # cumulative probability. For each step in the path, extract the
    # basefrequencies and the local (non-cumulative) probability:
    freq = numpy.empty(len(probabilities))
    prob = numpy.empty(len(probabilities))
    f_idx = numpy.argmax(cum_probability[-1])
```

```
    for t_idx in reversed(range(len(probabilities))):
        freq[t_idx] = basefrequencies[f_idx]
        prob[t_idx] = probabilities[t_idx, f_idx]
        f_idx = idx_probability[t_idx, f_idx]

    return time, freq, prob


def magnitude_correlation(blocks, basefrequencies=_basefrequencies):
    """Correlate synthetic tone complex spectra with a true spectrum.

    Generate spectra at a number of given base frequencies, then
    correlate each of these spectra with the magnitude signal
    STFT-spectra.

    Before correlation, each frequency bin is log-weighted to make the
    correlation perceptually accurate.

    blocks: A SignalBlocks instance.
    basefrequencies: An ordered list of base frequencies in Hz.

    Returns a len(blocks) x len(basefrequencies) matrix of correlation values.

    """
    specsize = blocks.blocksize//2+1

    # weigh differences according to perception:
    f = numpy.linspace(0, blocks.samplerate/2, specsize)
    log_f_weight = 1 / (blocks.samplerate/2)**(f / (blocks.samplerate/2))

    correlation = numpy.zeros([len(blocks), len(basefrequencies)])
    synthetic_magnitudes = synthetic_magnitude(blocks.samplerate, specsize, basefrequencies)

    for idx, spectrum in enumerate(stft(blocks)):
        # the correlation for real signals does not require the conj():
        correlation[idx] = numpy.sum(numpy.abs(spectrum) *
                                     synthetic_magnitudes *
                                     log_f_weight, axis=1)

    return correlation


def stft(blocks, *, nfft=None, windowfunc=scipy.signal.hann):
    """Short-time Fourier Transform of a signal.

    The signal is cut into short overlapping blocks, and each block is
    transformed into the frequency domain using the FFT. Before
    transformation, each block is windowed by windowfunc.

    blocks: A SignalBlocks instance.
    nfft: None for blocksize, or a number.
    windowfunc: A function that returns a window.

    Returns a complex spectrum.

    """
    window = windowfunc(blocks.blocksize)
    nfft = nfft or blocks.blocksize
    specsize = nfft//2+1
    for idx, block in enumerate(blocks):
```

```
        yield numpy.fft.rfft(block*window, nfft)


def synthetic_magnitude(samplerate, specsize, basefrequencies=_basefrequencies):
    """Synthetic magnitude spectra of a range of tone complexes.

    samplerate: The sampling rate of the tone complexes.
    specsize: The length of each spectrum.
    basefrequencies: An ordered vector of tone complex base
        frequencies in Hz.

    Returns a len(basefrequencies) x specsize matrix of tone complex spectra.

    """
    freqs = numpy.linspace(0, samplerate/2, specsize)
    synthetic_spectra = numpy.empty((len(basefrequencies), specsize), numpy.float64)
    for idx, basefrequency in enumerate(basefrequencies):
        synthetic_spectra[idx, :] = hannwin_comb(samplerate, basefrequency, specsize)
    return synthetic_spectra


def hannwin_comb(samplerate, basefreq, specsize):
    """Approximate a speech-like correlation spectrum of a tone complex.

    This is an approximation of time_domain_comb that runs much
    faster.

    Instead of calculating the FFT of a series of hann-windowed
    sinuses, this models the spectrum of a tone-complex as a series of
    hann-window-spectrums.

    For a perfect reconstruction, this would need to calculate the sum
    of many hann-window-spectra. Since hann window spectra are very
    narrow, this assumes that each window spectrum extends from
    n*basefreq-basefreq/2 to n*basefreq+basefreq/2 and that
    neighboring spectra do not influence each other.

    This assumtion holds as long as basefreq >> 1/specsize.

    Amplitudes are normalized by specsize.

    To make the spectrum more speech-like, frequencies above 1000 Hz
    are attenuated by 24 dB/oct.

    To make the correlation of this spectrum and some other spectrum
    have a normalized gain, the spectrum is shifted to be zero-mean.

    samplerate: The sampling rate in Hz of the signal.
    basefreq: The base frequency in Hz of the tone complex.
    specsize: The length of the resulting spectrum in bins
              (typically 2**N+1 for type(N) == int).

    Returns a real magnitude spectrum.

    """

    freqs = numpy.linspace(0, samplerate/2, specsize)
    # create a local frequency vector around each harmonic, going from
    # -basefreq/2 to basefreq/2 within the area around the nth
    # harmonic n*basefreq-basefreq/2 to n*basefreq+basefreq/2:
```

```
    closest_harmonic = (freqs + basefreq/2) // basefreq
    # ignore first half-wave:
    closest_harmonic[closest_harmonic==0] = 1
    local_frequency = closest_harmonic*basefreq - freqs
    # convert from absolute frequency to angular frequency:
    local_angular_freq = local_frequency/(samplerate/2)*2*numpy.pi
    # evaluate hannwin_spectrum at the local frequency vector:
    comb_spectrum = numpy.abs(hannwin_spectrum(local_angular_freq, specsize))
    # normalize to zero mean:
    comb_spectrum -= numpy.mean(comb_spectrum)
    # attenuate high frequencies:
    comb_spectrum[freqs>1000] /= 10**(numpy.log2(freqs[freqs>1000]/1000)*24/20)
    return comb_spectrum


def hannwin_spectrum(angular_freq, specsize):
    """Spectrum of a hann window

    The hann window is a linear combination of modulated rectangular
    windows r(n) = 1 for n=[0, N-1]:

    w(n) = 1/2*(1 - cos((2*pi*n)/(N-1)))
         = 1/2*r(n) - 1/4*exp(i*2*pi * n/(N-1))*r(n) - 1/4*exp(-i*2*pi * n/(N-1))*r(n)

    It's spectrum is then

    W(omega) = 1/2*R(omega) - 1/4*R(omega + (2*pi)/(N-1)) - 1/4*R(omega - (2*pi/(N-1)))

    with the spectrum of the rectangular window

    R(omega) = exp(-i*omega * (N-1)/2) * sin(N*omega/2) / sin(omega/2)

    (Source: https://en.wikipedia.org/wiki/Hann_function)

    angular_freq: Angular Frequency omega (0...2*pi), may be a vector.
    specsize: Length N of the resulting spectrum

    Returns the spectral magnitude for angular_freq.

    """

    def rectwin_spectrum(angular_freq):
        # In case of angular_freq == 0, this will calculate NaN. This
        # will be corrected later.
        spectrum = ( numpy.exp(-1j*angular_freq*(specsize-1)/2) *
                     numpy.sin(specsize*angular_freq/2) /
                     numpy.sin(angular_freq/2) )
        # since sin(x) == x for small x, the above expression
        # evaluates to specsize for angular_freq == 0.
        spectrum[angular_freq == 0.0] = specsize
        return spectrum

    angular_freq = numpy.asarray(angular_freq, dtype='float64')
    delta_f = 2*numpy.pi / (specsize-1)
    # don't warn about division by zero, NaNs will be corrected.
    with numpy.errstate(invalid='ignore'):
        return (1/2 * rectwin_spectrum(angular_freq) -
                1/4 * rectwin_spectrum(angular_freq + delta_f) -
                1/4 * rectwin_spectrum(angular_freq - delta_f)) / specsize
```

```
def ifd_difference(blocks, basefrequencies=_basefrequencies):
    """Compare generated IF deviations with true IF deviation.

    Generate IF deviations at given base frequencies, then subtract
    these from the true IF deviation of each block. The minimum
    difference marks the base frequency of the block.

    Each difference is log-weighted along the frequency to account for
    human perception, and bias-corrected along the base frequency to
    compensate for higher variances at higher base frequencies.

    blocks: A SignalBlocks instance.
    basefrequencies: an ordered vector of base frequencies in Hz.

    Returns a len(blocks) x len(basefrequencies) matrix of difference values.

    """
    specsize = blocks.blocksize//2+1

    synthetic_ifds = synthetic_ifd(blocks.samplerate, specsize, basefrequencies)

    # weigh differences according to perception:
    f = numpy.linspace(0, blocks.samplerate/2, specsize)
    log_f_weight = 1 / (blocks.samplerate/2)**(f / (blocks.samplerate/2))
    speech_weight = numpy.ones(f.shape)

    max_f0 = basefrequencies[-1]
    ifds = ifd(blocks, max_f0=max_f0)

    # larger base frequencies lead to larger IFDs:
    difference_bias = numpy.nanmean(numpy.abs(synthetic_ifds), axis=1)
    # larger signal variability leads to larger IFDs:
    signal_bias = numpy.nanmean(numpy.abs(ifds), axis=1)

    difference = numpy.zeros([len(blocks), len(basefrequencies)])
    for idx, this_ifd in enumerate(ifds):
        # the difference between the synthetic baseband instantaneous
        # frequency and the actual baseband instantaneous frequency
        # is minimal at the probable f0.
        difference_matrix = synthetic_ifds - this_ifd
        bias = numpy.sqrt(difference_bias**2 + signal_bias[idx]**2)
        # scale frequencies logarithmically, and correct for bias:
        difference[idx] = numpy.nanmean(numpy.abs(difference_matrix) *
                                        log_f_weight *
                                        speech_weight, axis=1) / bias

    return difference


def ifd(blocks, *, max_f0=450):
    """Instantaneous Frequency Deviation of a signal.

    Each time-frequency bin in the IFD has as value the difference
    between the bin's frequency and the most prominent frequency track
    in the bin's vicinity. As an example, if there is a prominent
    frequency track 100 Hz above a bin, it's value will be 100. If a
    bin is situated right on top of a frequency track, it's value will
    be 0. If it is above a frequency track, it's value is negative.
```

```
    This is equivalent to the frequency differentiation of a
    baseband-transformed STFT spectrum; aka BPD in [1].

    [1]: Krawczyk, M.; Gerkmann, T., "STFT Phase Reconstruction in
         Voiced Speech for an Improved Single-Channel Speech
         Enhancement," in Audio, Speech, and Language Processing,
         IEEE/ACM Transactions on , vol.22, no.12, pp.1931-1940, Dec.
         2014 doi: 10.1109/TASLP.2014.2354236

    The signal is cut into overlapping blocks. Each block is
    differentiated by splitting it in two strongly-overlapping
    sub-blocks, each sub-block is Fourier-transformed, the phase
    spectra are calculated, and the difference between the
    sub-blocks-spectra is calculated, and converted to frequencies.

    The frequencies thus obtained fall into a narrow range that is
    limited by the sub-block overlap. The maximum visible difference
    is given by max_f0, and governs the sub-block overlap. Lower
    max_f0 decrease the sub-block overlap, and wrap IFD frequencies at
    max_f0/2.

    blocks: A SignalBlocks instance.
    max_f0: The max frequency distance visible in the IFD.

    Returns a len(blocks) x blocks.samplerate//2+1 matrix of IFD values.

    """
    specsize = blocks.blocksize//2+1

    # number of samples that each sub-block-pair overlaps:
    dt = int(blocks.samplerate//max_f0)
    # a time window for each sub-block that maximizes phase accuracy:
    window = hannpoisson(blocks.blocksize, sym=False)

    longer_blocks = SignalBlocks(blocks.data, blocks.samplerate,
                                 blocks.blocksize+dt, blocks.hopsize)
    instfreq = numpy.zeros([len(longer_blocks), specsize], dtype='complex')
    for idx, block in enumerate(longer_blocks):
        # differentiate the spectrum phase in the time direction:
        # split the block in two strongly-overlapping blocks, then calculate the
        # angle difference (aka instantaneous frequency).
        spectrum1 = numpy.fft.rfft(window*block[:-dt], n=blocks.blocksize)
        spectrum2 = numpy.fft.rfft(window*block[dt:], n=blocks.blocksize)
        instfreq[idx] = spectrum2 * spectrum1.conj()

    # pure sinusoids change phase by this much in the given overlap time:
    baseband_phase_change = numpy.exp(1j * numpy.linspace(0, numpy.pi, specsize) * dt)
    # calculate the baseband phase difference / instantaneous frequency deviation:
    instfreq_deviation = numpy.angle(instfreq * baseband_phase_change.conj())
    instfreq_deviation *= max_f0/2 / numpy.pi # display as frequencies

    return instfreq_deviation


def hannpoisson(length, *, alpha=2, sym=True):
    """A window function with no side lobes.

    The Hann-Poisson window is a Hann Window times a Poisson window.
    It has the unusual feature of having "no side lobes" in the sense
    that, for alpha >= 2, the window-transform magnitude has negative
```

```
    slope for all positive frequencies [1][2].

    [1]:
    http://www.dsprelated.com/freebooks/sasp/Hann_Poisson_Window.html

    [2]: eq5.24 on p154 in Window Functions and Their Applications in
    Signal Processing by K. M. M. Prabhu, CRC Press)

    length: The length of the window.
    alpha: A slope constant (smooth slope for alpha >= 2)

    Returns the window array.

    """
    if not sym:
        length += 1
    normalization = (length-1) / 2
    n = numpy.arange(-normalization, normalization+1)
    poisson = numpy.exp(-alpha*numpy.abs(n) / normalization)
    hann = 1/2 * (1+numpy.cos(numpy.pi*n / normalization))
    if sym:
        return poisson * hann
    else:
        return (poisson * hann)[:-1]


def synthetic_ifd(samplerate, specsize, basefrequencies=_basefrequencies):
    """Synthetic instfreq deviations of a range of tone complexes.

    samplerate: The sampling rate of the tone complexes.
    specsize: The length of each IFD.
    basefrequencies: An ordered vector of tone complex base
        frequencies in Hz.

    Returns a len(basefrequencies) x specsize matrix of IFDs.

    """
    max_f0 = basefrequencies[-1]
    synthetic_ifds = numpy.zeros((len(basefrequencies), specsize), numpy.float64)
    f = numpy.linspace(0, samplerate/2, specsize)
    for idx, basefrequency in enumerate(basefrequencies):
        # the baseband_instreq shows the frequency difference to the closest
        # dominant harmonic:
        closest_harmonic = (f + basefrequency/2) // basefrequency
        instfreq = closest_harmonic*basefrequency - f
        # since it is derived from the phase, it wraps:
        instfreq = wrap_angles(instfreq, max_f0/2, inplace=True)
        instfreq[f < basefrequency/2] = 0
        synthetic_ifds[idx] = instfreq

    return synthetic_ifds


def wrap_angles(angles, limit, *, inplace=False):
    """something like modulo, but wraps n ≥ limit to n ± 2*limit.

    Useful for confining angular values to a wrapping data range.

    angles: Some real values.
    limit: The maximum/minimum valid value.
```

```
    inplace: Whether to overwrite values in angles (faster).

    Returns wrapped angles.

    """
    angles = numpy.array(angles, copy=not inplace)
    too_big = angles > limit
    angles[too_big] -= numpy.ceil((angles[too_big]-limit)/(2*limit))*2*limit
    too_small = angles < -limit
    angles[too_small] -= numpy.floor((angles[too_small]+limit)/(2*limit))*2*limit
    return angles


class SignalBlocks:
    """A generator for short, overlapping signal blocks.

    Each SignalBlocks instance contains the signal `data` and its
    `samplerate`. It generates short, possibly overlapping signal
    blocks of length `blocksize`. Each block starts `hopsize` after
    the previous block.

    The SignalBlocks' `len` is its number of blocks, and its
    `duration` is the signal length in seconds.

    """

    def __init__(self, data, samplerate, blocksize=2048, hopsize=1024):
        self.data = data
        self.samplerate = int(samplerate)
        self.blocksize = int(blocksize)
        self.hopsize = int(hopsize)

    def __iter__(self):
        idx = 0
        while idx+self.blocksize < len(self.data):
            yield(self.data[idx:idx+self.blocksize])
            idx += self.hopsize

    def __len__(self):
        return int(numpy.ceil( (len(self.data)-self.blocksize) / self.hopsize ))

    @property
    def duration(self):
        return len(self.data)/self.samplerate

_magnitude_correlation = numpy.frombuffer(b64decode(
    b'mGrQJey9QMCe/0La69g3wBtUytH+ayzA8lEd3ktMEsBQBFrnZT8UQEqtaNaLZS1ANyySXLJVOEDk'
    b'APhmT/xAQKzrpp/FzUVAdNZV2DufSkA9wQQRsnBPQAPW2SQUIVJAZ0sxQc+JVEDMwIhdivJWQDA2'
    b'4HlFW1lAlKs3lgDEW0D4II+yuyxeQC5Lc2e7SmBA4AWf9Rh/YUCSwMqDdrNiQA=='
    ), dtype='double')

_ifd_difference = numpy.frombuffer(b64decode(
    b'QM+ExWYxsj9WRR487eOyP2y7t7JzlrM/gTFRKfpItD+Wp+qfgPu0P6wdhBYHrrU/wZMdjY1gtj/W'
    b'CbcDFBO3P+x/UHqaxbc/Avbp8CB4uD8XbINnpyq5PyziHN4t3bk/Qli2VLSPuj9Xzk/LOkK7P2xE'
    b'6UHB9Ls/grqCuEenvD+XMBwvzlm9P62mtaVUDL4/whxPHNu+vj/XkuiSYXG/Pw=='
    ), dtype='double')

_posterior = numpy.frombuffer(b64decode(
    b'FAvfatwgsj+6t4fUv1awPzCfCvN3Mas/KLtCfM76pD9zS4ObSPmbP5IFkiZpzo4/BomuaYtqfD86'
    b'0dREyE9mP8tCzWxIulA/zjotfjR2QT89OM4nQVg+P6JaRjc4O0A/Oq/XmvaJQT8b+sytyclCP/z3'
```

```
    b'FqV6+EM/eqPgbJ5hRD8OTYuTNSlEP/LKwhzy90M/TnYNKhRyQz+Pcne44TlCP+yWorIkD7A/cGBM'
    b'rDFcrT8QjswRE92oPxCxdo0eMaM/KARH6Dh9mT8uDmFlDjmMP32GGH63TXo/pj2yJW/OZD+q/PT8'
    b'kitPP/iBYIoNOkA/O2hCRdghPD+EGiKIaSg+P5cMtd+tTEA/TFG0lN1HQT+QF4PSOxZCP3bT1HBg'
    b'dkI/YlvTsad6Qj+cx1tSY1lCP6UrchjeqEE/pL6/hZluQD/MxcByV52tP94d9TS6sqo/W7fTTjlU'
    b'pj+eVBTCCCShP+ZOfOKV7pY/Ok2O/B/IiT9iKXgBgVx4P/jpwDYAlGM/3KHHceL3TT+DkEXkXbU/'
    b'P7hCRDQ5Wzs/tmiLVlwBPT9wIYgcPBM/P/jDlc6oPkA/ukLRGpy/QD+dXIP9cAlBP4BPSITsGEE/'
    b'a8q7o5ftQD/jektu+UFAPzL7pKLQnT4/UMlq2evCuT9qbMhjaQK3P4qHixXe4LI/UgpxGruorD9i'
    b'mvHRO0KjP5M0f3mu75U/o6uK6kP7hD9F5CyBdh9xP3ia9H+LFVs//COlP3JrTT90jx86NyRJP1CW'
    b'twsuCEo/JL7i2GZdSz9oQDeDiRhMPwpnqnaobkw/xRv4l0d7TD+QroCy7EJMP2ryWaP5rEs/fP/K'
    b'wSt/Sj/a97CXYjxJP971ghvX39M/qsNy4wnP0T8sMXxhQzjNPyJhtqNvC8Y/E9ZxtDR3vT9wO6Vi'
    b'D8iwP95ppFmUC6A/uHGoaEcYij8bQzvzBqt0P1D9qcpTfGY/uqRUZqINYz9WRqZs/IVjP0SdemZw'
    b'a2Q/OgTNp3r0ZD843ytYES9lP0owKXorMGU/BM/eTyX+ZD/K8uJnloVkP6L0veWBn2M/3hlsRfOr'
    b'Yj+iYgdleh7qPwJA+O76P+c/tA4Z9Ev+4j8Yp4P6NWDcP5tfidjLxNI/trzXPshRxT+y4KsEI1m0'
    b'P9Q9MgNyfqA/UvNmR54Iij/yt3oVPCV8PwT9PcqvvHc/CkVZezBCeD/dG7YN1WR5P26cY8ilP3o/'
    b'IN/qDuTKej/UrtROGhR7P/ah3GIU63o/+jCKZ1w4ej9a35Pd1A15Pywc3Qws5Hc/V6rT2A6P+T+I'
    b'8E/N0Y/2P9ueKjokcPI/JlxeATKQ6z/I9H8RHSLiP4ZTP7m0itQ/jHRiBMiZwz+sh23mM7yvPx3f'
    b'Q1rd85g/cEExpe9ij9GxwE2xoOGP7ahBsXsCYc/b7cSG3IniD/Zujo/BCOJPwfZrvOi5ok/vNFP'
    b'84Juij+jSjpUmmOKP1FJaXjFrYk/Kv86FmKNiD/WwDNWyH+HP8nmHMCeUARAdJ0dqSWtAUAcK8qO'
    b'MfT8P5BvmE9eDPY/+v7bA2cn7T/tWMDA12TgP+YPRh7eMM8/VBHFq2I9uT9aNmtK+sKjP6gPVKVK'
    b'BpU/wNIzBsu8kT8czwUuuzuSP0CCkM61MpM/SC5UjZMVlD8JP84QjMCUPwEmOmjmMJU/hlq0FmQ6'
    b'lT+g2AmmQL2UP7Kx1L+95ZM/2zJf9IUukz9SRon7gqgNQEfapATtDQlAqZVcC7tHBEAUU4/g7lz/'
    b'PyLexlZ20fQ/bB3nK70n5z9VkdzsEObVP8QNnJYsvME/YYqoz5mxqz9AD7ZNcE6dPybnhmIuypg/'
    b'2Gv5Z7KymT+LKRIeSUabP1qpBicNqJw/2jGetjSgnT8IlAgIPhqePwivbuxrFJ4/XBazXNlsnT+f'
    b'1ek73EWcPygYaNpvWJs/LKSfrNKeE0BKOZw9QXkQQJnl6aCEqgpAzo8l+yPCBEDsE5OHjJr7PzJH'
    b'79IEkO4/mgH3wJzi3D8BgspRp3PHP3aRKgdLLbI/Sn/hDG4Aoz/xR3pVRQugP0KEiiYHx6A/5aHm'
    b'MF79oT9fxQ6wdwijP/bh/NOlvaM/AtQ+J3/zoz+oWistMcujPzNayU7lY6M/izTahXivoj/+orur'
    b'bu6hP6RLib4uDRdAMDOn6QlkFEB+6LFjYgIRQDnd4X7SeApAxAf5NQm+AUCWfFlYyMbzP6Q6nNVX'
    b'puI/DqcWqpIrzj8YK8hdlwG3PwBhXtuZk6c/XuYYZeHHoz/MSewtfMWkP1+BJWlkiKY/wDwqezAa'
    b'qD/+T49t0hipP0xZgYgUVak/9gtDKUQSqT9uR8k78qyoP1bY2E1g9qc/x9dbPyjmpj8Y4PJpnBsa'
    b'QPmBx+QH7hhA1q5zT6WSFUAeFBI1X5YQQGpjITFgdAZA/uu6sOdc+T8QtHTpKXvnP/hdap16hdI/'
    b'vr6lxNXWuz8YD2iLc0ysP2npu+5Fpqc/4titT7DiqD+iXefg3yqrP0Eiz1QEMq0/ugPQQNeFrj9g'
    b'a7QisASvP3x7eyow4q4/LCLR29GQrj/+k7NRRvatPyL4vQ3786w/nvlc4IiwG0Bv/nX6Z1kcQCT4'
    b'OC4O7BlAlfyS8GtKFEBmqOtXLOELQLEC1qe0q/8/pYqQRBSD7D9yccUiyaDVP1Rm0xEHEcA/Tqn3'
    b'OHeFsD992VDk18mrP57dz2nqT60/7EoCrwHqrz+RRrEMCw2xPx5VfX3N17E/zmvVSdxLsj9QgToX'
    b'uGOyP/I33nfdS7I/UCVTfxoUsj8ULkl1GcSxPxCogC4xsxxAmjbKIr1cHkA8YTkqNkgdQOWobM2W'
    b'ChhAuiyIGJ7JEEAEsLl4TfkCQFB/3Kjk2vA/8NU9qxzy2D9cmuz1wSPCP23t1g9EsrI/SP/uG9e+'
    b'rz8EnEuIf8+wP4AnvKzwPbI/hd/tnAlesz+Whewxn0O0PwQguLUH/rQ/sHlXCe1QtT8NsxZALzu1'
    b'P7YomyaD6bQ/W19cdryWtD8arxwpwlQfQGZsnVaE1CBAZxLm2+66IEC0xtBweWccQGA1ksFqrBNA'
    b'qjfBGN+6BUDtH2jsBlbzP8r9Pqqno9w/on5LXtdlxD9zHDz6PqS0P7zo01aJlLE/QuuT9ruysj+t'
    b'rVbQA0W0P+TYwcRHd7U/osk6yhuGtj9wZsmA25O3PxYDxHsTRbg/xAG10jF3uD+IDP3iOP63P7LM'
    b'72CLNrc/8bGoqwSmIEBzKmwME+giQJJGBgxjPiNALrpTYnCDIEC8ou1k0FsWQLtX2BIovwdAOMQK'
    b'Tjj79D9xUwBa+WLfPzQiE2L1YsY/8lqN89uItj94iqg4iDGzP654A3HXdLQ/xXpfd3wztj9ldleI'
    b'w323P8Q7ykyTvbg/bQtjVgH5uT9GBtZijQq7P340VzKrDrw/juyd9d/2uz/+rqayrOW6PxoqIMh4'
    b'0iJAiMhHxw45JUAwubZgchMlQG18idYAzSFA7qOP3AEMGEBneUjtfGgJQLjb7zb+IPY/BEw6oOdc'
    b'4D/YiM77w6PHPwXSUPgSarg/LoDjuW7StD8ErHYgYwy2P406PPr0D7g/kMqCjT2duT9DXk/j8vu6'
    b'P2B+Pm0kHrw/KrXLmWKXvT/uN8o6+NC/P2B2tpSwb8A/E6hHWe5TwD+ks881EFQmQIwtoexmDihA'
    b'ekpW8XumJkACH9RGvioiQF5PGPqF1BhAeEPzCOeQC0D0kHNfxiz4P1Qu+z1GXuE/4oztxQzTyD9u'
    b'D+v7UxC6P+X6hqgyc7Y/rjKiF66qtz/nXAPRRt+5PxzyUekLtLs/tvRW+xsavT+XXjpCqDy+P1Ss'
    b'RbmAN8A/pPum0f2jwT+Km8SbbkTCP6x1g356gsI//XZXbSPFKkDkfm2dqrkrQByPRPenPihAx/qc'
    b'op2YIUCox6GHY+8XQMAJzoOCkgxA5i0KkJEj+j/yc03/tNviP/ZpefEJk8o/Kv06pvCauz+yF1ov'
    b'jge4PyaphgqBuLk/wG3KqGMWvD+8vRKDo6i9P3hgJ0x+xL4/lt/8LNcqwD8WqAUWxdzBP7LYiHxs'
    b'JcM/VKMAp83Nwj+68dVWpnDCP5RhejaQqDBApFWXOEWHMECsmG48xj4rQHDDR7P45yFABlmUG9aQ'
    b'F0CpzOYzHl8NQA6RgStXgPs/eMUCr+0B5D8qLgyflzfMP8iIzNv9/7w/OHNobvskuT/yeh/nnYK7'
    b'P2E310u3N74/HPP3O80pvz/evsf1CJi/P4ml5gnb1cA/oxwI+UtFwz+5hNDy43rEP77MvtLKAsM/'
    b'RJpAgj24wT8='
), dtype='double').reshape((20, 20))

value2posterior = RectBivariateSpline(_magnitude_correlation, _ifd_difference, _posterior).ev
```

AMDF

```
import numpy

# Ross et al., 1974
def amdf(signal, samplerate):
    delays = range(samplerate//450, samplerate//50)
    blocklen = int(0.036*samplerate)
    hoplen = int(0.01*samplerate)
    result = []
    for blockidx in range(delays[-1], len(signal)-blocklen, hoplen):
        diff = numpy.zeros(len(delays))
        for delayidx, delay in enumerate(delays):
            diff[delayidx] = 1/blocklen * numpy.sum(
                numpy.abs(signal[blockidx      :blockidx+blocklen] -
                          signal[blockidx-delay:blockidx+blocklen-delay]))
        result.append(samplerate/delays[numpy.argmin(diff)])
    return (numpy.arange(len(result))*hoplen/samplerate, # time
            numpy.array(result),                         # frequency
            numpy.ones(len(result)))                     # probability
```

CEP

```
import numpy
import scipy.signal

# Noll, 1967
def cep(signal, samplerate):
    # low-pass filter to 4 kHz (filter type is unspecified)
    coeffs = scipy.signal.butter(4, 4000/(samplerate/2))
    signal = scipy.signal.lfilter(*coeffs, signal)
    blocklen = int(0.040*samplerate) # 40 ms
    hoplen = int(0.010*samplerate) # 10 ms
    window = scipy.signal.hamming(blocklen)
    # search range between 1 ms ... 15 ms
    one_ms = int(blocklen/2 * 0.001/0.040) # in quefrency bins
    fifteen_ms = int(blocklen/2 * 0.015/0.040) # in quefrency bins
    # weighed with 1 to 5 from 1 to 15 ms
    weight = numpy.linspace(1, 5, fifteen_ms-one_ms)
    result = []
    for blockidx in range(0, len(signal)-blocklen, hoplen):
        block = signal[blockidx:blockidx+blocklen]
        cepstrum = numpy.fft.rfft(numpy.log(abs(numpy.fft.fft(window*block))**2))
        region_of_interest = numpy.real(cepstrum[one_ms:fifteen_ms]) # this is already approximately real
        region_of_interest *= weight
        result.append(numpy.argmax(region_of_interest) /
                      blocklen * 0.040 + 0.001) # convert to s
    return (numpy.arange(len(result))*hoplen/samplerate, # time
            1/numpy.array(result),                       # frequency
            numpy.ones(len(result)))                     # probability
```

AUTOC

```
import numpy
import scipy.signal

# Sondhi, 1968
def autoc(signal, samplerate):
```

```
    signal = signal.copy()

    # center clipping:
    blocklen = int(0.005*samplerate)
    for blockidx in range(0, len(signal)-blocklen, blocklen):
        block = signal[blockidx:blockidx+blocklen]
        threshold = 0.3*numpy.max(numpy.abs(block))
        # these modify signal:
        block[numpy.abs(block) < threshold] = 0
        block[block != 0] -= threshold * numpy.sign(block[block != 0])

    # autocorrelation up to 15 ms:
    blocklen = int(0.030*samplerate)
    hoplen = int(0.010*samplerate)
    window = scipy.signal.hamming(blocklen)
    fifteen_ms = int(blocklen * 0.015/0.030)
    weight = numpy.linspace(1, 5, fifteen_ms) # unspecified in the paper
    result = []
    for blockidx in range(0, len(signal)-blocklen, hoplen):
        block = signal[blockidx:blockidx+blocklen]*window
        autocorrelation = numpy.correlate(block, block, 'full')[-blocklen:]
        region_of_interest = numpy.abs(autocorrelation[:fifteen_ms])
        argmax = numpy.argmax(region_of_interest*weight)
        result.append(samplerate/argmax if argmax != 0 else 0)
    return (numpy.arange(len(result))*hoplen/samplerate, # time
            numpy.array(result),                         # frequency
            numpy.ones(len(result)))                     # probability
```

SIFT

```
import numpy
import scipy.signal

# Markel, 1972
def sift(signal, samplerate):
    order = 16
    blocklen = int(0.032*samplerate)
    two_ms = int(blocklen * 0.002/0.032)
    result = []
    for blockidx in range(0, len(signal)-blocklen, blocklen):
        block = signal[blockidx:blockidx+blocklen]
        autocorrelation = numpy.correlate(block, block, 'full')
        autocorrelation = autocorrelation[len(autocorrelation)//2:]
        toeplitz = scipy.linalg.toeplitz(autocorrelation[:order-2])
        coeffs = numpy.linalg.inv(toeplitz) @ -autocorrelation[1:order-1]
        filtered = scipy.signal.lfilter([1, *coeffs], [1], block)
        autocorrelation = numpy.correlate(filtered, filtered, 'full')
        autocorrelation = autocorrelation[len(autocorrelation)//2:]
        argmax = numpy.argmax(autocorrelation[two_ms:]) + two_ms
        result.append(samplerate/argmax)

    return (numpy.arange(len(result))*blocklen/samplerate, # time
            numpy.array(result),                           # frequency
            numpy.ones(len(result)))                       # probability
```

Literature Review Database

[1] Muhammad Navid Anjum Aadit, Sharadindu Gopal Kirtania, and Mehnaz Tabassum Mahin. Pitch and formant estimation of bangla speech signal using autocorrelation, cepstrum and LPC algorithm. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 371–376, Dhaka, Bangladesh, December 2016. IEEE.

[2] M. Abe and S. Ando. Nonlinear time-frequency domain operators for decomposing sounds into loudness, pitch and timbre. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1368–1371, Detroit, MI, USA, 1995. IEEE.

[3] M. Abe and S. Ando. Application of loudness/pitch/timbre decomposition operators to auditory scene analysis. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 5, pages 2646–2649, Atlanta, GA, USA, 1996. IEEE.

[4] T. Abe, T. Kobayashi, and S. Imai. Harmonics tracking and pitch extraction based on instantaneous frequency. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 756–759, Detroit, MI, USA, 1995. IEEE.

[5] T. Abe, T. Kobayashi, and S. Imai. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 1277–1280, Philadelphia, PA, USA, 1996. IEEE.

[6] Jakob Abeser and Meinard Muller. Fundamental Frequency Contour Classification: A Comparison between Hand-crafted and CNN-based Features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 486–490, Brighton, United Kingdom, May 2019. IEEE.

[7] M N Abhijith, Prasanta K Ghosh, and K Rajgopal. Multi-pitch tracking using Gaussian mixture model with time varying parameters and Grating Compression Transform. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1473–1477, Florence, Italy, May 2014. IEEE.

[8] Iman Haji Abolhassani, Douglas O'Shaughnessy, and Sid-Ahmed Selouani. A method utilizing window function frequency characteristics for noise-robust spectral pitch estimation. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[9] A.R. Abu-El-Quran and R.A. Goubran. Adaptive Pitch-Based Speech Detection for Hands-Free Applications. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 3, pages 305–308, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[10] N. Abu-Shikhah and M. Deriche. A novel pitch estimation technique using the Teager energy function. In *ISSPA '99. Proceedings of the Fifth International Symposium on Signal Processing and its Applications (IEEE Cat. No.99EX359)*, volume 1, pages 135–138, Brisbane, Qld., Australia, 1999. Queensland Univ. Technol.

[11] S. I. Adalbjornsson, A. Jakobsson, and M. G. Christensen. Estimating multiple pitches using block sparsity. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6220–6224, Vancouver, BC, Canada, May 2013. IEEE.

[12] Stefan I. Adalbjörnsson, Andreas Jakobsson, and Mads G. Christensen. Multi-pitch estimation exploiting block sparsity. *Signal Processing*, 109:236–247, April 2015.

[13] S. Ahmadi and A.S. Spanias. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, May 1999.

[14] Kazi Jamir Uddin Ahmed and Md. Rezwan khan. Estimation of Pitch of Noisy Speech using AR Model Based Inverse Filtering. In *2006 International Conference on Electrical and Computer Engineering*, pages 447–450, Dhaka, Bangladesh, December 2006. IEEE.

[15] R. Ahn and W. H. Holmes. An improved harmonic-plus-noise decomposition method and its application in pitch determination. In *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, pages 41–42, September 1997.

[16] Philipp Aichinger, Martin Hagmuller, Berit Schneider-Stickler, Jean Schoentgen, and Franz Pernkopf. Tracking of Multiple Fundamental Frequencies in Diplophonic Voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):330–341, February 2018.

[17] Aimilios Chalamandaris, Pirros Tsiakoulis, Sotiris Karabetsos, and Spyros Raptis. An efficient and robust pitch marking algorithm on the speech waveform for TD-PSOLA. In *2009 IEEE International Conference on Signal and Image Processing Applications*, pages 397–401, Kuala Lumpur, Malaysia, 2009. IEEE.

[18] William A Ainsworth, Charles R Day, and Georg F Meyer. Improving Pitch Estimation with Short Duration Speech Samples. *5th International Conference on Spoken Language Processing (ICSLP 98)*, page 4, 1998.

[19] Francesc Alias, Carlos Monzo, and Joan Claudi Socoro. A Pitch Marks Filtering Algorithm Based on Restricted Dynamic Programming. *INTERSPEECH 2006*, page 4, 2006.

[20] Francesc Alias and Natàlia Munne. Reliable Pitch Marking of Affective Speech at Peaks or Valleys Using Restricted Dynamic Programming. *IEEE Transactions on Multimedia*, 12(6):481–489, October 2010.

[21] A. K. Alimuradov. An Algorithm for Measurement of the Pitch Frequency of Speech Signals Based on Complementary Ensemble Decomposition Into Empirical Modes. *Measurement Techniques*, 59(12):1316–1323, March 2017.

[22] Alan Alimuradov. Research of Frequency-Selective Properties of Empirical Mode Decomposition Methods for Speech Signals' Pitch Frequency Estimation. In *2015 International Conference on Engineering and Telecommunication (EnT)*, pages 77–79, Moscow, Russia, November 2015. IEEE.

[23] A. Alkulaibi, J.J. Soraghan, and T.S. Durrani. Fast HOS based simultaneous voiced/unvoiced detection and pitch estimation using 3-level binary speech signals. In *Proceedings of 8th Workshop on Statistical Signal and Array Processing*, pages 194–197, Corfu, Greece, 1996. IEEE Comput. Soc. Press.

[24] G. A. Alzamendi, G. Schlotthauer, and M. E. Torres. A New Method for Structural Analysis of Perturbed Pitch Period Series. In Ariel Braidot and Alejandro Hadad, editors, *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*, volume 49, pages 492–495. Springer International Publishing, Cham, 2015.

[25] B. Anantharaman, K.R. Ramakrishnan, and S.H. Srinivasan. Wavelet based pitch extraction in the mpeg compressed domain. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 124–127, Tokyo, Japan, 2001. IEEE.

[26] Kristian Timm Andersen and Marc Moonen. An adaptive time-frequency analysis scheme for improved real-time speech enhancement. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6265–6269, Florence, Italy, May 2014. IEEE.

[27] M. S. Andrews. Eigenstructure based pitch estimation in speech processing applications. In *[1992] Conference Record of the Twenty-Sixth Asilomar Conference on Signals, Systems Computers*, pages 1111–1115 vol.2, October 1992.

[28] M.S. Andrews, J. Picone, and R.D. Degroat. Robust pitch determination via SVD based cepstral methods. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 253–256, Albuquerque, NM, USA, 1990. IEEE.

[29] G. Aneeja and B. Yegnanarayana. Extraction of Fundamental Frequency From Degraded Speech Using Temporal Envelopes at High SNR Frequencies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):829–838, April 2017.

[30] Manjare Chandraprabha Anil and S. D. Shirbahadurkar. Expressive speech synthesis using prosodic modification for Marathi language. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 126–130, Noida, Delhi-NCR, India, February 2015. IEEE.

[31] Y. Arai, R. Mochizuki, H. Nishimura, and T. Honda. An excitation synchronous pitch waveform extraction method and its application to the VCV-concatenation synthesis of Japanese spoken words. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1437–1440 vol.3, October 1996.

[32] Luc Ardaillon and Axel Roebel. Fully-Convolutional Network for Pitch Estimation of Speech Signals. In *Interspeech 2019*, pages 2005–2009. ISCA, September 2019.

[33] Andrei Sebastian Ardeleanu and Marinel Temneanu. Fundamental frequency estimation based on mean values. In *2013 8TH INTERNATIONAL SYMPOSIUM ON ADVANCED TOPICS IN ELECTRICAL ENGINEERING (ATEE)*, pages 1–4, Bucharest, Romania, May 2013. IEEE.

[34] L. Arevalo. Linear predictive, eigenvalue oriented pitch-contour measurement for forensic voice identification. In *Fifth ASSP Workshop on Spectrum Estimation and Modeling*, pages 299–303, October 1990.

[35] Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma. Energy and F0 Contour Modeling with Functional Data Analysis for Emotional Speech Detection. *INTERSPEECH 2013*, page 5, 2013.

[36] I. O. Arkhipov and V. B. Gitlin. Pitch extraction using a generated decision function. *Acoustical Physics*, 46(5):511–517, September 2000.

[37] Meysam Asgari and Izhak Shafran. Improving the Accuracy and the Robustness of Harmonic Model for Pitch Estimation. *INTERSPEECH 2013*, page 5, 2013.

[38] K. Ashouri and M. H. Savoji. Automatic and accurate pitch marking of speech signal using an expert system based on logical combinations of different algorithms outputs. In *2004 12th European Signal Processing Conference*, pages 995–998, September 2004.

[39] I.A. Atkinson, B.G. Evans, and A.M. Kondoz. Pitch detection of speech signals using segmented autocorrelation. *Electronics Letters*, 31(7):533–535, March 1995.

[40] E Azarov, M Vashkevich, D Likhachov, and A Petrovsky. A low-delay algorithm for instantaneous pitch estimation. page 9, 2015.

[41] Elias Azarov, Maxim Vashkevich, and Alexander Petrovsky. Instantaneous pitch estimation based on RAPT framework. *20th European Signal Processing Conference (EUSIPCO 2012)*, page 5, 2012.

[42] Elias Azarov, Maxim Vashkevich, and Alexander Petrovsky. Instantaneous pitch estimation algorithm based on multirate sampling. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4970–4974, Shanghai, March 2016. IEEE.

[43] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7815–7819. IEEE, 2013.

[44] M.M. Babu. Efficient and accurate pitch estimation using FFT. In *Proceedings. IEEE International Joint Symposia on Intelligence and Systems (Cat. No.98EX174)*, pages 354–358, Rockville, MD, USA, 1998. IEEE Comput. Soc.

[45] F.R. Bach and M.I. Jordan. Discriminative Training of Hidden Markov Models for Multiple Pitch Tracking. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages 489–492, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[46] Tom Backstrom, Stefan Bayer, and Sascha Disch. Pitch Variation Estimation. *INTERSPEECH 2009*, page 4, 2009.

[47] Roland Badeau, Valentin Emiya, and Bertrand David. Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3073–3076, Taipei, Taiwan, April 2009. IEEE.

[48] MyungJin Bae, SangMok Shon, and HahYoung Yo. Soongsil University, (*)ETFU-VLSI Lab Seoul 156-743, Korea mjbae@saintsoongsil.ac.kr. *International Symposium on Signal Processing and its Applications (ISSPA)*, page 4, 1996.

[49] Fadoua Bahja, Joseph Di Martino, and El Hassane Ibn Elhaj. Real-time pitch tracking using the eCATE algorithm. In *2010 5th International Symposium On I/V Communications and Mobile Network*, pages 1–4, Rabat, Morocco, September 2010. IEEE.

[50] Fadoua Bahja, Joseph Di Martino, Elhassan Ibn Elhaj, and Driss Aboutajdine. A corroborative study on improving pitch determination by time–frequency cepstrum decomposition using wavelets. *SpringerPlus*, 5(1):564, December 2016.

[51] Fadoua Bahja, El Hassan Ibn Elhaj, and Joseph Di Martino. On the use of wavelets and cepstrum excitation for Pitch Determination in real-time. In *2012 International Conference on Multimedia Computing and Systems*, pages 150–153, Tangiers, Morocco, May 2012. IEEE.

[52] Tomasz Bandurski, Lukasz Hamerski, Michal Papaj, Agnieszka Paruzel, and Krzysztof Swider. Pitch estimation of narrowband-filtered speech signal using Instantaneous Complex Frequency. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2007*, pages 157–162, Poznan, Poland, September 2007. IEEE.

[53] Andras Banhalmi, Andras Kocsor, Kornel Kovacs, and Laszlo Toth. Fundamental Frequency Estimation by Combinations of Various Methods. In *Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006*, pages 330–333, Reykjavik, Iceland, June 2006. IEEE.

[54] Andras Banhalmi, Kornel Kovacs, Andras Kocsor, and Laszlo Toth. Fundamental Frequency Estimation by Least-Squares Harmonic Model Fitting. *INTERSPEECH 2005*, page 4, 2005.

[55] Hynek BaNI and Petr Pollak. Direct time domain fundamental frequency estimation of speech in noisy conditions. *2004 12th European Signal Processing Conference*, page 4, 2004.

[56] Sahil Bansal, Anindita Ghosh, Chandra Sekhar Seelamantula, Gurunath Gurrala, and Prasanta Kumar Ghosh. Adaptive frequency estimation using iterative DESA with RDFT-based filter. In *2017 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 1–6, Bangalore, November 2017. IEEE.

[57] N Barbot, Olivier Boeffard, and D Lolive. F_0 Stylisation with a Free-Knot B-Spline Model and Simulated-Annealing Optimization. *INTERSPEECH 2005*, page 4, 2005.

[58] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva. Pitch detection with a neural-net classifier. *IEEE Transactions on Signal Processing*, 39(2):298–307, February 1991.

[59] K.E. Barner. Nonlinear estimation of DEGG signals with applications to speech pitch detection. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 4, pages 2243–2246, Philadelphia, PA, USA, 1996. IEEE.

[60] K.E. Barner. Colored L-l filters and their application in speech pitch detection. *IEEE Transactions on Signal Processing*, 48(9):2601–2606, September 2000.

[61] K.E. Barner and J.A. Gallant. Nonlinear estimation of EGG signals with applications to speech pitch detection. In *Proceedings of 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1340–1341, Baltimore, MD, USA, 1994. IEEE.

[62] Jan Bartošek and Václav Hanžl. Exploring abilities of merged normalized forward-backward correlation for speech pitch analysis. *2011 International Conference on Applied Electronics*, page 4, 2011.

[63] Mohamed Anouar Ben Messaoud, Aicha Bouzid, and Noureddine Ellouze. A Robust pitch estimation approach for clean speech. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 756–760, Sousse, Tunisia, March 2012. IEEE.

[64] Mohamed Anouar Ben Messaoud and Aïcha Bouzid. Pitch estimation of speech and music sound based on multi-scale product with auditory feature extraction. *International Journal of Speech Technology*, 19(1):65–73, March 2016.

[65] Mohamed Anouar Ben Messaoud, Aïcha Bouzid, and Noureddine Ellouze. Autocorrelation of the Speech Multi-Scale Product for Voicing Decision and Pitch Estimation. *Cognitive Computation*, 2(3):151–159, September 2010.

[66] S.L. Bernadin and S.Y. Foo. Wavelet Processing for Pitch Period Estimation. In *2006 Proceeding of the Thrity-Eighth Southeastern Symposium on System Theory*, pages 123–126, Cookeville, TN, USA, 2006. IEEE.

[67] Sanjivani S. Bhabad, G. K. Kharate, and Smita C. Stunde. Pitch detection in time, frequency and cepstral domain for articulatory handicapped people. In *2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE)*, pages 80–84, Jaipur, India, December 2013. IEEE.

[68] V. Bharathi, Asaph Abraham A., and R. Ramya. Vocal pitch detection for musical transcription. In *2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies*, pages 724–726, Thuckafay, July 2011. IEEE.

[69] Peter Birkholz, Patrick Schmaser, and Yi Xu. Estimation of Pitch Targets from Speech Signals by Joint Regularized Optimization. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2075–2079, Rome, September 2018. IEEE.

[70] Rachel M. Bittner, Avery Wang, and Juan P. Bello. Pitch contour tracking in music using Harmonic Locked Loops. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 191–195, New Orleans, LA, March 2017. IEEE.

[71] Bo Li, Ying-ying Li, Cheng-you Wang, Chao-jing Tang, and Er-yang Zhang. A new efficient pitch-tracking algorithm. In *IEEE International Conference on Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003*, volume 2, pages 1102–1107, Changsha, Hunan, China, 2003. IEEE.

[72] N.M. Botros and R.S. Adamjee. Speech-pitch detection using maximum likelihood algorithm. In *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. No.99CH37015)*, volume 2, page 882, Atlanta, GA, USA, 1999. IEEE.

[73] B. Boyanov, T. Ivanov, S. Hadjitodorov, and G. Chollet. Robust hybrid pitch detector. *Electronics Letters*, 29(22):1924, 1993.

[74] D. R. Brown, R. Mudumbai, and S. Dasgupta. Fundamental limits on phase and frequency tracking and estimation in drifting oscillators. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5225–5228, Kyoto, Japan, March 2012. IEEE.

[75] Karen Bryden, Andre Brind'Amour, and Hisham Hassanein. A Robust Pitch and Voicing Detector for Harmonic Coders. *International Symposium on Signal Processing and its Applications (ISSPA)*, page 4, August 1996.

[76] Luis Buera, Jasha Droppo, and Alex Acero. Speech enhancement using a pitch predictive model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4885–4888, Las Vegas, NV, USA, March 2008. IEEE.

[77] I S Burnett and U B Gambino. Pitch Detection Based on Prototype Waveform. *International Symposium on Signal Processing and its Applications (ISSPA)*, page 4, August 1996.

[78] D. Burshtein. Joint maximum likelihood estimation of pitch and AR parameters using the EM algorithm. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 797–800, Albuquerque, NM, USA, 1990. IEEE.

[79] N. R. Butt, S. I. Adalbjornsson, S. D. Somasundaram, and A. Jakobsson. Robust fundamental frequency estimation in the presence of inharmonicities. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5499–5503, Vancouver, BC, Canada, May 2013. IEEE.

[80] Dmitry Bykhovsky and Ofer Hadar. Evaluation of a GLRT threshold for voiced-unvoiced decision and pitch tracking in noisy speech. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 000680–000683, Eilat, Israel, November 2010. IEEE.

[81] Tom Bäckström. Fundamental Frequency. In *Speech Coding*, pages 91–96. Springer International Publishing, Cham, 2017.

[82] Benedikt T Bönninghoff, Robert M Nickel, Steffen Zeiler, and Dorothea Kolossa. Unsupervised Classification of Voiced Speech and Pitch Tracking Using Forward-Backward Kalman Filtering. *Speech Communication*, page 5, 2016.

[83] Lutfiye Cagan and Umut Arioz. Turkish pitch frequency detection: AutoCorreletaion and cepstral method. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 248–251, Malatya, Turkey, May 2015. IEEE.

[84] Runshen Cai, Yaoting Zhu, and Shaoqiang Shi. A Modified Pitch Detection Method Based on Wavelet Transform. In *2010 Second International Conference on Multimedia and Information Technology*, pages 246–249, Kaifeng, China, 2010. IEEE.

[85] Arturo Camacho. On the use of auditory models' elements to enhance a sawtooth waveform inspired pitch estimator on telephone-quality signals. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1080–1085, Montreal, QC, Canada, July 2012. IEEE.

[86] Arturo Camacho and John G. Harris. A Pitch Estimation Algorithm Based on the Smooth Harmonic Average Peak-to-Valley Envelope. In *2007 IEEE International Symposium on Circuits and Systems*, pages 3940–3943, New Orleans, LA, USA, May 2007. IEEE.

[87] Roudra Chakraborty, Debapriya Sengupta, and Sagnik Sinha. Pitch tracking of acoustic signals based on average squared mean difference function. *Signal, Image and Video Processing*, 3(4):319–327, December 2009.

[88] C.-F. Chan and E.W.M. Yu. Improving pitch estimation for efficient multiband excitation coding of speech. *Electronics Letters*, 32(10):870, 1996.

[89] Liang Chang, Jingde Xu, Kun Tang, and Huijuan Cui. A new robust pitch determination algorithm for telephone speech. *ISITA*, page 3, 2012.

[90] Bao Changchun, Dai Yisong, and Fan Changxin. Two kings of pitch predictors in speech compressing coding. *Journal of Electronics (China)*, 14(3):200–208, July 1997.

[91] Changle Zhou and Lanlan Lv. The fundamental frequency estimation of Guqin timbre based on wavelet transform. In *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, pages 89–92, Shanghai, China, November 2009. IEEE.

[92] Chao Wang and S. Seneff. Robust pitch tracking for prosodic modeling in telephone speech. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1343–1346, Istanbul, Turkey, 2000. IEEE.

[93] D. Charalampidis and V.B. Kura. Novel wavelet-based pitch estimation and segmentation of non-stationary speech. In *2005 7th International Conference on Information Fusion*, page 5 pp., Philadelphia, PA, USA, 2005. IEEE.

[94] Indira Chatterjee, Priya Gupta, Parthasarathi Bera, and Joy Sen. Pitch Tracking and Pitch Smoothing Methods-Based Statistical Approach to Explore Singers' Melody of Voice on a Set of Songs of Tagore. In Rabindranath Bera, Subir Kumar Sarkar, and Swastika Chakraborty, editors, *Advances in Communication, Devices and Networking*, volume 462, pages 509–515. Springer Singapore, Singapore, 2018.

[95] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski. Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1299–1302, Istanbul, Turkey, 2000. IEEE.

[96] D. Chazan, Y. Stettiner, and D. Malah. Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 728–731 vol.2, Minneapolis, MN, USA, 1993. IEEE.

[97] H. Chen, W.C. Wong, and C.C. Ko. A comparison of pitch prediction algorithms in forward and backward adaptive CELP systems. In *[Proceedings] Singapore ICCS/ISITA '92*, pages 821–825, Singapore, 1990. IEEE.

[98] H. Chen, W.C. Wong, and C.C. Ko. Comparison of pitch prediction and adaptation algorithms in forward and backward adaptive CELP systems. *IEE Proceedings I Communications, Speech and Vision*, 140(4):240, 1993.

[99] P. Chen and S. Ando. Pitch from zeros of bank-filtered signals. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 530–533 vol.2, Minneapolis, MN, USA, 1993. IEEE.

[100] S.-H. Chen and J.-F. Wang. Noise-robust pitch detection method using wavelet transform with aliasing compensation. *IEE Proceedings - Vision, Image, and Signal Processing*, 149(6):327, 2002.

[101] Xiao Chen and Bo Xu. An improved pitch extraction algorithm for speech processing. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 399–399, Singapore, Singapore, September 2014. IEEE.

[102] Xuemei Chen and Ruolun Liu. Multiple Pitch Estimation Based on Modified Harmonic Product Spectrum. In Wei Lu, Guoqiang Cai, Weibin Liu, and Weiwei Xing, editors, *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, volume 211, pages 271–279. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[103] A. Cherif. Pitch and formants extraction algorithm for speech processing. In *ICECS 2000. 7th IEEE International Conference on Electronics, Circuits and Systems (Cat. No.00EX445)*, volume 1, pages 595–598, Jounieh, Lebanon, 2000. IEEE.

[104] Soumeya Cherouat and Farid Marir. Pitch detection and voicing/unvoicing decision of Arabic speech signal by HOS-polycesptre. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 768–771, Sousse, Tunisia, March 2012. IEEE.

[105] Chen-Yu Chiang. On Smoothing and Enhancing Dynamics of Pitch Contours Represented by Discrete Orthogonal Polynomials for Prosody Generation. In *INTERSPEECH 2016*, pages 2303–2307, September 2016.

[106] Y. Chisaki, T. Usagawa, and M. Ebata. Improvement of pitch estimation using harmonic wavelet transform. In *Proceedings of IEEE. IEEE Region 10 Conference. TENCON 99. 'Multimedia Technology for Asia-Pacific Information Infrastructure' (Cat. No.99CH37030)*, volume 1, pages 601–604, Cheju Island, South Korea, 1999. IEEE.

[107] A. Choi. Real-time fundamental frequency estimation by least-square fitting. *IEEE Transactions on Speech and Audio Processing*, 5(2):201–205, March 1997.

[108] Wen-Sheng Chou, Kah-Meng Cheong, and Tai-Shih Chi. A binaural algorithm for space and pitch detection. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4976–4979, Prague, Czech Republic, May 2011. IEEE.

[109] Jeiran Choupan, Seyed Ghorshi, Mohammad Mortazavi, and Farshid Sepehrband. Pitch extraction using dyadic wavelet transform and modified higher order moment. In *2010 IEEE 12th International Conference on Communication Technology*, pages 833–836, Nanjing, China, November 2010. IEEE.

[110] Heidi Christensen, Ning Ma, Stuart N Wrigley, and Jon Barker. Integrating Pitch and Localisation Cues at a Speech Fragment Level. *INTERSPEECH 2007*, page 4, 2007.

[111] Mads G. Christensen. Pitch Estimation. In *Introduction to Audio Processing*, pages 179–192. Springer International Publishing, Cham, 2019.

[112] Mads G. Christensen, Pedro Vera-Candeas, Samuel D. Somasundaram, and Andreas Jakobsson. Robust subspace-based fundamental frequency estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 101–104, Las Vegas, NV, USA, March 2008. IEEE.

[113] Mads Graesboll Christensen. A method for low-delay pitch tracking and smoothing. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 345–348, Kyoto, Japan, March 2012. IEEE.

[114] Mads Graesboll Christensen. Multi-channel maximum likelihood pitch estimation. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 409–412, Kyoto, Japan, March 2012. IEEE.

[115] Mads Graesboll Christensen. Accurate Estimation of Low Fundamental Frequencies From Real-Valued Measurements. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2042–2056, October 2013.

[116] Mads Graesboll Christensen, Petre Stoica, Andreas Jakobsson, and Soren Holdt Jensen. The Multi-Pitch Estimation Problem: some New Solutions. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages III–1221–III–1224, Honolulu, HI, April 2007. IEEE.

[117] Mads Grcesboll Christensen and Jesper Rindom Jensen. Pitch estimation for non-stationary speech. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 1400–1404, Pacific Grove, CA, November 2014. IEEE.

[118] Mads Groesboll Christensen. On the estimation of low fundamental frequencies. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 169–172, New Paltz, NY, USA, October 2011. IEEE.

[119] Mads Groesboll Christensen. An exact subspace method for fundamental frequency estimation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6802–6806, Vancouver, BC, Canada, May 2013. IEEE.

[120] Mads Groesboll Christensen, Andreas Jakobsson, and Soren Holdt Jensen. Multi-Pitch Estimation Using Harmonic Music. In *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pages 521–524, Pacific Grove, CA, USA, 2006. IEEE.

[121] Mads Groesboll Christensen, Andreas Jakobsson, and Soren Holdt Jensen. Fundamental Frequency Estimation using the Shift-Invariance Property. In *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, pages 631–635, Pacific Grove, CA, USA, November 2007. IEEE.

[122] Mads Grsbll Christensen, Andreas Jakobsson, and Sren Holdt Jensen. Joint High-Resolution Fundamental Frequency and Order Estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1635–1644, July 2007.

[123] Mads Grsbll Christensen, Jesper Hjvang Jensen, Andreas Jakobsson, and Sren Holdt Jensen. On Optimal Filter Designs for Fundamental Frequency Estimation. *IEEE Signal Processing Letters*, 15:745–748, 2008.

[124] Mads Græsbøll Christensen, Jesper Lisby Højvang, Andreas Jakobsson, and Søren Holdt Jensen. Joint fundamental frequency and order estimation using optimal filtering. *EURASIP Journal on Advances in Signal Processing*, 2011(1), December 2011.

[125] Wei Chu and Abeer Alwan. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3969–3972. IEEE, 2009.

[126] Wei Chu and Abeer Alwan. SAFE: A Statistical Algorithm for F0 Estimation for Both Clean and Noisy Speech. *INTERSPEECH 2010*, page 4, 2010.

[127] Chunghsin Yeh, Axel Roebel, and Xavier Rodet. Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1116–1126, August 2010.

[128] Chunmao Zhang, Kun Tang, Huijuan Cui, Wen Du, and Jing Li. Efficient pitch predictor algorithm in ITU-T G.723.1. In *International Conference on Communication Technology Proceedings, 2003. ICCT 2003.*, volume 2, pages 1727–1729, Beijing, China, 2003. Beijing Univ. Posts & Telecommun. Press.

[129] You Chunyan and Bai Sen. Robust Information Hiding in Speech Signal Based on Pitch Period Prediction. In *2010 International Conference on Computational and Information Sciences*, pages 533–536, Chengdu, China, December 2010. IEEE.

[130] Chunyan Li, V. Cuperman, and A. Gersho. Robust closed-loop pitch estimation for harmonic coders by time scale modification. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 257–260 vol.1, Phoenix, AZ, USA, 1999. IEEE.

[131] Supasit Chuwatthananurux and Dittaya Wanvarie. Improving noise estimation with RAPT pitch voice activity detection under low SNR condition. In *2016 8th International Conference on Knowledge and Smart Technology (KST)*, pages 77–82, Chiangmai, Thailand, February 2016. IEEE.

[132] Amelia Ciobanu, Tudor Catalin Zorila, Cristian Negrescu, and Dumitru Stanomir. Maximum Voiced Frequency Estimation for Voice Conversation Used in Text-To-Speech Systems. In *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Piscataway, NJ, 2012. IEEE. OCLC: 855872892.

[133] Pascal Clark, Sri Harish Mallidi, Aren Jansen, and Hynek Hermansky. Frequency offset correction in speech without detecting pitch. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7020–7024, Vancouver, BC, Canada, May 2013. IEEE.

[134] H. Clergeot, S. Tressens, and A. Ouamri. Performance of high resolution frequencies estimation methods compared to the Cramer-Rao bounds. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1703–1720, November 1989.

[135] L. Cnockaert, F. Grenez, and J. Schoentgen. Fundamental Frequency Estimation and Vocal Tremor Analysis by means of Morlet Wavelet Transforms. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 393–396, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[136] Voinea Radu Cociu and Livia Cociu. Optimization of a method to evaluate the fundamental frequency in real time. In *2017 International Conference on Electromechanical and Power Systems (SIELMEN)*, pages 157–162, Iasi, October 2017. IEEE.

[137] Vincent Colotte and Yves Laprie. Higher precision pitch marking for TD-PSOLA. *2002 11th European Signal Processing Conference*, page 4, 2002.

[138] Zhijun Cui. Pitch extraction based on weighted autocorrelation function in speech signal processing. In *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, pages 2158–2162, Changchun, China, December 2012. IEEE.

[139] V. Cuperman and R. Pettigrew. Robust low-complexity backward adaptive pitch predictor for low-delay speech coding. *IEE Proceedings I Communications, Speech and Vision*, 138(4):338, 1991.

[140] Ryunosuke Daido and Yuji Hisaminato. A Fast and Accurate Fundamental Frequency Estimator Using Recursive Moving Average Filters. In *INTERSPEECH 2016*, pages 2160–2164, September 2016.

[141] H R Dajani, D Purcell, W Wong, H Kunov, and T W Picton. Recording human evoked potentials that follow the pitch contour of a natural vowel. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 52(9):5, 2005.

[142] Manuel Davy. Multiple Fundamental Frequency Estimation Based on Generative Models. In Anssi Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription*, pages 203–227. Springer US, Boston, MA, 2006.

[143] Paul De Palma and Mark VanDam. Using automatic speech processing to analyze fundamental frequency of child-directed speech stored in a very large audio corpus. In *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, pages 1–6, Otsu, Japan, June 2017. IEEE.

[144] Gilles Degottex and Daniel Erro. A uniform phase representation for the harmonic model in speech synthesis applications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):38, 2014.

[145] Boyuan Deng, Denis Jouvet, Yves Laprie, Ingmar Steiner, and Aghilas Sini. Towards confidence measures on fundamental frequency estimations. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5605–5609, New Orleans, LA, March 2017. IEEE.

[146] Rohit S. Deo and Pallavi S. Deshpande. Pitch contour modelling and modification for expressive Marathi speech synthesis. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2455–2458, Delhi, India, September 2014. IEEE.

[147] Der-Jenq Liu and Chin-Teng Lin. Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure. *IEEE Transactions on Speech and Audio Processing*, 9(6):609–621, September 2001.

[148] O. Deshmukh, C.Y. Espy-Wilson, A. Salomon, and J. Singh. Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):776–786, September 2005.

[149] O. Deshmukh, J. Singh, and C. Espy-Wilson. A novel method for computation of periodicity, aperiodicity and pitch of speech signals. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–117–20, Montreal, Que., Canada, 2004. IEEE.

[150] Johanna Devaney and Michael Mandel. An evaluation of score-informed methods for estimating fundamental frequency and power from polyphonic audio. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185, New Orleans, LA, March 2017. IEEE.

[151] Johanna C. Devaney, Michael I. Mandel, and Ichiro Fujinaga. Characterizing singing voice fundamental frequency trajectories. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 73–76, New Paltz, NY, USA, October 2011. IEEE.

[152] Jitendra Kumar Dhiman, Nagaraj Adiga, and Chandra Sekhar Seelamantula. A Spectro-Temporal Demodulation Technique for Pitch Estimation. In *Interspeech 2017*, pages 2306–2310. ISCA, August 2017.

[153] P. Dikshit, S.A. Zahorian, and S. Nagulapati. An Algorithm for Locating Fundamental Frequency Markers in Speech Signals. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 233–236, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[154] Hui Ding, Bo Qian, Yanping Li, and Zhenmin Tang. A Method Combining LPC-Based Cepstrum and Harmonic Product Spectrum for Pitch Detection. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, pages 537–540, Pasadena, CA, USA, December 2006. IEEE.

[155] Sascha Disch and Bernd Edler. Frequency selective pitch transposition of audio signals. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 29–32, Prague, Czech Republic, May 2011. IEEE.

[156] C.-T. Do, D. Pastor, and A. Goalic. On Normalized MSE Analysis of Speech Fundamental Frequency in the Cochlear Implant-Like Spectrally Reduced Speech. *IEEE Transactions on Biomedical Engineering*, 57(3):572–577, March 2010.

[157] M. C. Dogan and J. M. Mendel. Real-time robust pitch detector. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 129–132 vol.1, March 1992.

[158] Bin Dong. Characterizing resonant component in speech: A different view of tracking fundamental frequency. *Mechanical Systems and Signal Processing*, 88:318–333, May 2017.

[159] Minghui Dong and Kim-Teng Lua. Pitch Contour Model for Chinese Text-to-Speech Using CART and Statistical Model. *7th International Conference on Spoken Language Processing (ICSLP2002)*, page 4, 2002.

[160] Mingye Dong, Jie Wu, and Jian Luan. Vocal Pitch Extraction in Polyphonic Music Using Convolutional Residual Network. In *Interspeech 2019*, pages 2010–2014. ISCA, September 2019.

[161] Dong-Yan Huang, Weisi Lin, and Susanto Rahardja. Speech pitch detection in noisy environment using multi-rate adaptive lossless FIR filters. In *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*, pages III–429–32, Vancouver, BC, Canada, 2004. IEEE.

[162] B. Doval and X. Rodet. Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 221–224 vol.1, Minneapolis, MN, USA, 1993. IEEE.

[163] Yaron Doweck, Alon Amar, and Israel Cohen. Fundamental Initial Frequency and Frequency Rate Estimation of Random-Amplitude Harmonic Chirps. *IEEE Transactions on Signal Processing*, 63(23):6213–6228, December 2015.

[164] James Droppo and Alex Acero. Maximum a Posteriori Pitch Tracking. *5th International Conference on Spoken Language Processing (ICSLP 98)*, page 4, 1998.

[165] Jasha Droppo and Alex Acero. A Fine Pitch Model for Speech. *INTERSPEECH 2007*, page 4, August 2007.

[166] Thomas Drugman and Abeer Alwan. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. *INTERSPEECH 2011*, page 4, 2011.

[167] Thomas Drugman, Goeric Huybrechts, Viacheslav Klimkov, and Alexis Moinet. Traditional Machine Learning for Pitch Detection. *IEEE Signal Processing Letters*, 25(11):1745–1749, November 2018.

[168] Sicong Du, Yosuke Sugiura, and Tetsuya Shimamura. Combining Zero Replacement Speech Enhancement with Lag Window Method for Pitch Detection. In *2018 IEEE 3rd International Conference on Communication and Information Systems (ICCIS)*, pages 53–57, Singapore, Singapore, December 2018. IEEE.

[169] Zhiyao Duan, Jinyu Han, and Bryan Pardo. Song-level multi-pitch tracking by heavily constrained clustering. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, Dallas, TX, USA, 2010. IEEE.

[170] Jean-Louis Durrieu and Jean-Philippe Thiran. Source/Filter Factorial Hidden Markov Model, With Application to Pitch and Formant Tracking. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2541–2553, December 2013.

[171] Valentin Emiya, Bertrand David, and Roland Badeau. A Parametric Method for Pitch Estimation of Piano Tones. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages I–249–I–252, Honolulu, HI, April 2007. IEEE.

[172] Taoufik En-Najjary, Olivier Rosec, and Thierry Chonavel. A New Method for Pitch Prediction from Spectral Envelope and its Application in Voice Conversion. *EUROSPEECH 2003*, page 4, 2003.

[173] A. Erell and M. Weintraub. Estimation of noise-corrupted speech DFT-spectrum using the pitch period. *IEEE Transactions on Speech and Audio Processing*, 2(1):1–8, January 1994.

[174] Zhe-Cheng Fan, Jyh-Shing Roger Jang, and Chung-Li Lu. Singing Voice Separation and Pitch Extraction from Monaural Polyphonic Audio Music via DNN and Adaptive Pitch Tracking. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, pages 178–185, Taipei, Taiwan, April 2016. IEEE.

[175] Luoyang Fang, Dongliang Duan, Liuqing Yang, and Louis Scharf. Error floor elimination for DFT-based frequency estimators. *21st European Signal Processing Conference (EUSIPCO 2013)*, page 5, 2013.

[176] Mohamed Hesham Farouk. Spectral Analysis of Speech Signal and Pitch Estimation. In *Application of Wavelets in Speech Processing*, pages 37–39. Springer International Publishing, Cham, 2014.

[177] Mohamed Hesham Farouk. Spectral Analysis of Speech Signal and Pitch Estimation. In *Application of Wavelets in Speech Processing*, pages 23–28. Springer International Publishing, Cham, 2018.

[178] H. Farsi. A Novel Method to Improve Pitch and Voicing Strength Estimation for Low Bit Rate Speech Coding. In *2006 2nd International Conference on Information & Communication Technologies*, volume 1, pages 1281–1286, Damascus, Syria, 2006. IEEE.

[179] S. A. Fattah, W.-P. Zhu, and M. O. Ahmad. A time-frequency domain formant frequency estimation scheme for noisy speech signals. In *2009 IEEE International Symposium on Circuits and Systems*, pages 1201–1204, Taipei, Taiwan, May 2009. IEEE.

[180] Feng Huang and Tan Lee. Pitch Estimation in Noisy Speech Using Accumulated Peak Spectrum and Sparse Estimation Technique. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):99–109, January 2013.

[181] Attila Ferencz, Jeong-Su Kim, Yong-Beom Lee, and Jae-Won Lee. Automatic Pitch Marking and Reconstruction of Glottal Closure Instants from Noisy and Deformed Electro-Glotto-Graph Signals. *INTERSPEECH 2004*, page 4, October 2004.

[182] P. Fernandez-Cid and F.J. Casajus-Quiros. Multi-pitch estimation for polyphonic musical signals. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 6, pages 3565–3568, Seattle, WA, USA, 1998. IEEE.

[183] A. Ferrari, G. Alengrin, and C. Theys. Estimation of the fundamental frequency of a noisy sum of cisoids with harmonic related frequencies. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 517–520 vol.5, March 1992.

[184] Carlos Ferrer, Diana Torres, and María E Hernández-Díaz. Using Dynamic Time Warping of T0 Contours in the Evaluation of Cycle-to-Cycle Pitch Detection Algorithms. *Iberoamerican Congress on Pattern Recognition (CIARP 2008)*, page 8, 2008.

[185] Federico Flego, Maurizio Omologo, and Luca Armani. On the Use of a Weighted Autocorrelation Based Fundamental Frequency Estimation for a Multidimensional Speech Input. *INTERSPEECH 2004*, page 4, 2004.

[186] H. Fujisaki, S. Ohno, and O. Tomita. Automatic parameter extraction of fundamental frequency contours of speech based on a generative model. In *Proceedings of Third International Conference on Signal Processing (ICSP'96)*, volume 1, pages 729–732, Beijing, China, 1996. IEEE.

[187] Hiroya Fujisaki, Keikichi Hirose, and Shigenobu Seto. Proposal and Evaluation of a New Scheme for Reliable Pitch Extraction of Speech. In *First International Conference on Spoken Language Processing (ICSLP 90)*, Kobe, Japan, November 1990.

[188] Keiichi Funaki. On Evaluation of the F_0 Estimation Based on Time-Varying Complex Speech Analysis. *INTERSPEECH 2010*, page 4, 2010.

[189] Thibaut Fux, Gang Feng, and Veronique Zimpfer. Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of speech. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4964–4967, Prague, Czech Republic, May 2011. IEEE.

[190] M. Gainza, B. Lawlor, and E. Coyle. Multi pitch estimation by using modified IIR comb filters. In *47th International Symposium ELMAR, 2005.*, pages 233–236, Zadar, Croatia, 2005. IEEE.

[191] Jovan Galic and Tatjana Pesic-Brdanin. The voice fundamental frequency statistical parameters under noisy conditions with the cepstrum method. In *2011 10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS)*, pages 769–772, Nis, October 2011. IEEE.

[192] P M B Gambino and I S Burnett. Low Delay Pitch Detection using Dynamic-Programming/Viterbi Techniques. *International Symposium on Signal Processing and its Applications (ISSPA)*, page 4, August 1996.

[193] Chunxian Gao and Hui Liu. Diving helmet noise spectral estimation based on pitch tracking. In *2012 International Conference on Systems and Informatics (ICSAI2012)*, pages 1673–1676, Yantai, China, May 2012. IEEE.

[194] Jun Gao and Dan Xu. Noise-robust pitch detection algorithm based on AMDF with clustering analysis picking peaks. In *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, pages 1144–1148, Chongqing, China, May 2016. IEEE.

[195] Yongwei Gao, Bilei Zhu, Wei Li, Ke Li, Yongjian Wu, and Feiyue Huang. Vocal Melody Extraction via DNN-based Pitch Estimation and Salience-based Pitch Refinement. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1000–1004, Brighton, United Kingdom, May 2019. IEEE.

[196] N. Garcia, E. Macias-Toro, J.F. Vargas-Bonilla, J.M. Daza, and J.D. Lopez. Segmentation of bio-signals in field recordings using fundamental frequency detection. In *3rd IEEE International Work-Conference on Bioinspired Intelligence*, pages 86–92, Liberia, Costa Rica, July 2014. IEEE.

[197] N. Garcia, J. C. Vasquez-Correa, J. F. Vargas-Bonilla, J. D. Arias-Londono, and J. R. Orozco-Arroyave. Evaluation of the effects of speech enhancement algorithms on the detection of fundamental frequency of speech. In *2014 XIX Symposium on Image, Signal Processing and Artificial Vision*, pages 1–5, Armenia, Colombia, September 2014. IEEE.

[198] Philip N. Garner, Milos Cernak, and Petr Motlicek. A Simple Continuous Pitch Estimation Algorithm. *IEEE Signal Processing Letters*, 20(1):102–105, January 2013.

[199] Inge Gavat, Matei Zirra, and Bogdan Sabac. Pitch Estimation by Block and Instantaneous Methods. *International Journal of Speech Technology*, page 11, 2002.

[200] Saeed Gazor, Habib Hajimolahoseini, Hamid Soltanian-Zadeh, and Rassoul Amirfattahi. Instantaneous fundamental frequency estimation of non-stationary periodic signals using non-linear recursive filters. *IET Signal Processing*, 9(2):143–153, April 2015.

[201] Edouard Geoffrois. The Multi-Lag-Window Method for Robust Extended-Range F0 Determination. *4th International Conference on Spoken Language Processing (ICSLP 96)*, page 5, 1996.

[202] Branislav Gerazov and Zoran Ivanovski. Analysis of extracted pitch contours across speakers for intonation modelling in TTS synthesis. In *2012 5th International Symposium on Communications, Control and Signal Processing*, pages 1–4, Roma, Italy, May 2012. IEEE.

[203] David Gerhard. Multiresolution Pitch Analysis of Talking, Singing, and the Continuum Between. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Dominik Ślęzak, JingTao Yao, James F. Peters, Wojciech Ziarko, and Xiaohua Hu, editors, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, volume 3642, pages 294–303. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[204] Timo Gerkmann, Rainer Martin, and Derya Dalga. Multi-microphone maximum a posteriori fundamental frequency estimation in the cepstral domain. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4505–4508, Taipei, Taiwan, April 2009. IEEE.

[205] Stephan Gerlach, Jörg Bitzer, Stefan Goetze, and Simon Doclo. Joint estimation of pitch and direction of arrival: improving robustness and accuracy for multi-speaker scenarios. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):1–17, 2014.

[206] S. Ghaemmaghami, M. Deriche, and B. Boashash. A new approach to pitch and voicing detection through spectrum periodicity measurement. In *TENCON '97 Brisbane - Australia. Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (Cat. No.97CH36162)*, volume 2, pages 743–746, Brisbane, Qld., Australia, 1997. IEEE.

[207] S. Ghaemmghami and M. Deriche. A new approach to efficient interpolative determination of pitch contour using temporal decomposition. In *Proceedings of Digital Processing Applications (TENCON '96)*, volume 1, pages 125–130, Perth, WA, Australia, 1996. IEEE.

[208] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2494–2498. IEEE, 2014.

[209] M. Ghazvini, N. Movahedinia, and A. Vafaei. Pitch period detection using second generation wavelet transform. In *ICSES 2010 International Conference on Signals and Electronic Circuits*, pages 53–56, September 2010.

[210] Prasanta Kumar Ghosh, Antonio Ortega, and Shrikanth S Narayanan. Pitch Period Estimation Using Multipulse Model and Wavelet Transform. *INTERSPEECH 2007*, page 4, 2007.

[211] J.D. Gibson and I. Lee. Robust backward adaptive pitch prediction for speech coder. *Electronics Letters*, 31(7):536–538, March 1995.

[212] Keith D. Gilbert and Karen L. Payton. Source enumeration of speech mixtures using pitch harmonics. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 89–92, New Paltz, NY, USA, October 2009. IEEE.

[213] Wei-Guo Gong, Li-Ping Yang, and Di Chen. Pitch Synchronous Based Feature Extraction for Noise-Robust Speaker Verification. In *2008 Congress on Image and Signal Processing*, pages 295–298, Sanya, China, 2008. IEEE.

[214] Sira Gonzalez and Mike Brookes. A Pitch Estimation Filter robust to high levels of noise (PEFAC). *19th European Signal Processing Conference (EUSIPCO 2011)*, page 5, 2011.

[215] Sira Gonzalez and Mike Brookes. Speech active level estimation in noisy conditions. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6684–6688, Vancouver, BC, Canada, May 2013. IEEE.

[216] Sira Gonzalez and Mike Brookes. PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):518–530, February 2014.

[217] K. Gopalan. Pitch estimation using a modulation model of speech. In *WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, volume 2, pages 786–791, Beijing, China, 2000. IEEE.

[218] Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Low-Complexity Pitch Estimation Based on Phase Differences Between Low-Resolution Spectra. In *Interspeech 2017*, pages 2316–2320. ISCA, August 2017.

[219] A. A. Grigoryan. Method of Measuring the Fundamental Frequency of a Complex Periodic Signal. *Measurement Techniques*, 46(11):1084–1087, November 2003.

[220] Francois Grondin and Francois Michaud. Robust speech/non-speech discrimination based on pitch estimation for mobile robots. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1650–1655, Stockholm, May 2016. IEEE.

[221] Y. H. Gu. HMM-based noisy-speech pitch contour estimation. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 21–24 vol.2, March 1992.

[222] Guan Tian, Feng Shu, Huang Shengyang, and Ye Datian. Application of lifting scheme in pitch extraction for speech processing of cochlear implants based on characteristics of chinese language. In *Proceedings 7th International Conference on Signal Processing, 2004. Proceedings. ICSP '04. 2004.*, volume 3, pages 2222–2225, Beijing, China, 2004. IEEE.

[223] Guan Tian and Ye Datian. Application of Lifting Scheme in Pitch Extraction for Cochlear Implant. In *2004 2nd IEEE/EMBS International Summer School on Medical Devices and Biosensors*, pages 40–42, Hong Kong, China, 2004. IEEE.

[224] Kang Guangyu and Guo Shize. Improving AMDF for pitch period detection. In *2009 9th International Conference on Electronic Measurement & Instruments*, pages 4–283–4–286, Beijing, China, August 2009. IEEE.

[225] Guoning Hu and DeLiang Wang. Speech segregation based on pitch tracking and amplitude modulation. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pages 79–82, New Platz, NY, USA, 2001. IEEE.

[226] Guoning Hu and DeLiang Wang. A Tandem Algorithm for Pitch Estimation and Voiced Speech Segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2067–2079, November 2010.

[227] Guotong Zhou and G.B. Giannakis. Harmonics in Gaussian multiplicative and additive noise: Cramer-Rao bounds. *IEEE Transactions on Signal Processing*, 43(5):1217–1231, May 1995.

[228] Tania Habib, Marian Kepesi, and Lukas Ottowitz. Experimental evaluation of the joint position-pitch estimation (POPI) algorithm in noisy environments. In *2008 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, pages 369–372, Darmstadt, Germany, July 2008. IEEE.

[229] Hae Young Kim, Jae Sung Lee, Myung-Whun Sung, Kwang Hyun Kim, and Kwang Suk Park. Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, volume 6, pages 3162–3164, Hong Kong, China, 1998. IEEE.

[230] Haiyan Guo, Xi Shao, and Zhen Yang. An improved phase-space voicing-state classification for co-channel speech based on pitch detection. In *2008 9th International Conference on Signal Processing*, pages 680–683, Beijing, China, October 2008. IEEE.

[231] Haiyun Yang, Lunji Qui, and Soo-Ngee Koh. Application of instantaneous frequency estimation for fundamental frequency detection. In *Proceedings of IEEE-SP International Symposium on Time- Frequency and Time-Scale Analysis*, pages 616–619, Philadelphia, PA, USA, 1994. IEEE.

[232] Habib Hajimolahoseini, Rassoul Amirfattahi, Saeed Gazor, and Hamid Soltanian-Zadeh. Robust Estimation and Tracking of Pitch Period Using an Efficient Bayesian Filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1, 2016.

[233] Habib Hajimolahoseini, Saeed Gazor, and Rassoul Amirfattahi. A robust and fast method for estimating and tracking the instantaneous fundamental frequency of audio signals. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4, Vancouver, BC, Canada, May 2016. IEEE.

[234] Md. Ekramul Hamid and Md. Khademul Islam Molla. A Collelogram based Pitch and Voiced/Unvoiced Classification Method for Real-Time Speech Analysis in Noisy Environment. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pages 93–98, Nadi, December 2017. IEEE.

[235] Sajad Hamzenejadi, Seyed Amir Yousef Hosseini Goki, and Mahdieh Ghazvini. Extraction of Speech Pitch and Formant Frequencies using Discrete Wavelet Transform. In *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pages 1–5, Bojnord, Iran, January 2019. IEEE.

[236] Kun Han and DeLiang Wang. Neural Network Based Pitch Tracking in Very Noisy Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):2158–2168, December 2014.

[237] Kun Han and DeLiang Wang. Neural networks for supervised pitch tracking in noise. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1488–1492, Florence, Italy, May 2014. IEEE.

[238] Hank Chang-Han Huang and F. Seide. Pitch tracking and tone features for Mandarin speech recognition. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1523–1526, Istanbul, Turkey, 2000. IEEE.

[239] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Graesboll Christensen. Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 186–190, New Orleans, LA, March 2017. IEEE.

[240] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Graesboll Christensen. Estimation of Fundamental Frequencies in Stereophonic Music Mixtures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):296–310, February 2019.

[241] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Grasboll Christensen. Multi-pitch estimation of audio recordings using a codebook-based approach. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 983–987, Budapest, August 2016. IEEE.

[242] K. Harisudha, S. Dhanalakshmi, and M. Madhusoodhanan. Implementation of sub band coding and pitch extraction using cumulative impulse strength. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 1125–1128, Chennai, March 2017. IEEE.

[243] J.D. Harris and D. Nelson. Glottal pulse alignment in voiced speech for pitch determination. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 519–522 vol.2, Minneapolis, MN, USA, 1993. IEEE.

[244] K. Hasan, C. Shahnaz, and S.A. Fatath. Determination of pitch of noisy speech using dominant harmonic frequency. In *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, volume 2, pages II–556–II–559, Bangkok, Thailand, 2003. IEEE.

[245] Md. Kamrul Hasan and Md. Lutful Kabir. Minimization of Error in Pitch Detection algorithm using Discrete Fractional Cosine Transform. In *2008 Australasian Telecommunication Networks and Applications Conference*, pages 403–406, Adelaide, Australia, December 2008. IEEE.

[246] Mirza A. F. M. Rashidul Hasan, Rubaiyat Yasmin, Dipankar Das, and M. S. Rahman. Correlation based pitch extraction method in speech signal. In *2014 9th International Forum on Strategic Technology (IFOST)*, pages 140–143, Cox's Bazar, Bangladesh, October 2014. IEEE.

[247] Hiroya Hashimoto, Keikichi Hirose, and Nobuaki Minematsu. Improved Automatic Extraction of Generation Process Model Commands and Its use for Generating Fundamental Frequency Contours for Training HMM-based Speech Synthesis. *INTERSPEECH 2012*, page 4, 2012.

[248] Jean-Paul Haton. Speech analysis for automatic speech recongnition: A review. In *2009 Proceedings of the 5-th Conference on Speech Technology and Human-Computer Dialogue*, pages 1–5, Constanta, Romania, June 2009. IEEE.

[249] He Ba, Na Yang, Ilker Demirkol, and Wendi Heinzelman. BaNa: A hybrid approach for noise resilient pitch detection. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 369–372, Ann Arbor, MI, USA, August 2012. IEEE.

[250] M. Heckmann, C. Glaser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick. Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1699–1704, Nice, September 2008. IEEE.

[251] M Heckmann, F Joublin, and K Nakadai. Pitch extraction in Human-Robot interaction. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1482–1487, Taipei, October 2010. IEEE.

[252] Martin Heckmann, Claudius Glaser, Frank Joublin, and Kazuhiro Nakadai. Applying Geometric Source Separation for Improved Pitch Extraction in Human-Robot Interaction. *INTERSPEECH 2010*, page 4, 2010.

[253] Martin Heckmann, Frank Joublin, and Christian Goerick. Combining Rate and Place Information for Robust Pitch Extraction. *INTERSPEECH 2007*, page 4, August 2007.

[254] P. Hedelin and D. Huber. Pitch period determination of aperiodic speech signals. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 361–364, Albuquerque, NM, USA, 1990. IEEE.

[255] Christian T. Herbst and Jacob C. Dunn. Fundamental Frequency Estimation of Low-quality Electroglottographic Signals. *Journal of Voice*, May 2018.

[256] W J Hess. Pitch and Voicing Determination of Speech with an Extension Toward Music Signals. In *Springer Handbook of Speech Processing*, page 31. 2008.

[257] R. Heyman, R. J. Bird, R. L. Heyman, and J. Harding. Programs for the estimation of fundamental frequency, amplitude, and voicing of speech. *Behavior Research Methods & Instrumentation*, 13(6):760–760, November 1981.

[258] K. Hirose, H. Fujisaki, and S. Seto. A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 149–152 vol.1, March 1992.

[259] Keikichi Hirose. Modeling of fundamental frequency contours for HMM-based speech synthesis: Representation of fundamental frequency contours for statistical speech synthesis. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pages 171–176, Chengdu, China, November 2016. IEEE.

[260] Keikichi Hirose, Hiroya Hashimoto, Jun Ikeshima, and Nobuaki Minematsu. Use of generation process model for synthesizing fundamental frequency contours in HMM-based speech synthesis. In *2012 IEEE 11th International Conference on Signal Processing*, pages 575–578, Beijing, China, October 2012. IEEE.

[261] Keikichi Hirose, Hiroya Hashimoto, Daisuke Saito, and Nobuaki Minematsu. Superpositional modeling of fundamental frequency contours for HMM-based speech synthesis. In *Speech Prosody 2016*, pages 771–775, May 2016.

[262] F. Hiroya and O. Sumio. A preliminary study on the modeling of fundamental frequency contours of Thai utterances. In *6th International Conference on Signal Processing, 2002.*, pages 516–519, Beijing, China, 2002. IEEE.

[263] Daniel Hirst, Hyongsil Cho, Sunhee Kim, and Hyunji Yu. Evaluating Two Versions of the Momel Pitch Modelling Algorithm on a Corpus of Read Speech in Korean. *INTERSPEECH 2007*, page 4, 2007.

[264] Jan Hlavnicka, Roman Cmejla, Jiri Klempir, Evzen Ruzicka, and Jan Rusz. Acoustic Tracking of Pitch, Modal, and Subharmonic Vibrations of Vocal Folds in Parkinson's Disease and Parkinsonism. *IEEE Access*, 7:150339–150354, 2019.

[265] L. Hodgson, M.E. Jernigan, and B.L. Wills. Nonlinear multiplicative cepstral analysis for pitch extraction in speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 257–260, Albuquerque, NM, USA, 1990. IEEE.

[266] Hojung Nam, Hyoung-Soo Kim, Y. Kwon, and Sung-Il Yang. Speaker verification system using hybrid model with pitch detection by wavelets. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380)*, pages 153–156, Pittsburgh, PA, USA, 1998. IEEE.

[267] Harvey Holmes and Weihua Zhang. INVESTIGATION ON THE SPECTRAL ENVELOPE ESTIMATOR(SEEV0c) AND REFINED PITCH ESTIMATION BASED ON THE SINUSOIDAL SPEECH MODEL. *Speech and Image Technologies for Computing and Telecommunications*, page 4, 1997.

[268] John N Holmes. Robust Measurement of Fundamental Frequency and Degree of Voicing. *5th International Conference on Spoken Language Processing (ICSLP 98)*, page 4, 1998.

[269] Hong Hong, Xiao-hua Zhu, Wei-min Su, Run-tong Geng, and Xin-long Wang. Detection of time varying pitch in tonal languages: an approach based on ensemble empirical mode decomposition. *Journal of Zhejiang University SCIENCE C*, 13(2):139–145, February 2012.

[270] Jung Ook Hong and Patrick J Wolfe. Model-Based Estimation of Instantaneous Pitch in Noisy Speech. *INTERSPEECH 2009*, page 4, 2009.

[271] Jung Ook Hong and Patrick J Wolfe. Robust and Efficient Pitch Estimation Using an Iterative ARMA Technique. *INTERSPEECH 2010*, page 4, 2010.

[272] Hong Hong, Zhengmin Zhao, Xinlong Wang, and Zhiyong Tao. Detection of Dynamic Structures of Speech Fundamental Frequency in Tonal Languages. *IEEE Signal Processing Letters*, 17(10):843–846, October 2010.

[273] Hong Zhang, Taiyi Huang, and Junshou Song. A new method of fundamental frequency extraction in frequency domain. In *ICSP '98. 1998 Fourth International Conference on Signal Processing (Cat. No.98TH8344)*, pages 690–693, Beijing, China, 1998. IEEE.

[274] John-Paul Hosom. F0 Estimation for Adult and Children's Speech. *INTERSPEECH 2005*, page 4, 2005.

[275] Kazushi Hotta and Keiichi Funaki. On a robust F0 estimation of speech based on IRAPT using robust TV-CAR analysis. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–4, Chiang Mai, Thailand, December 2014. IEEE.

[276] Chao-Ling Hsu, DeLiang Wang, and Jyh-Shing Roger Jang. A trend estimation algorithm for singing pitch detection in musical recordings. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 393–396, Prague, Czech Republic, May 2011. IEEE.

[277] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu. A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1482–1491, July 2012.

[278] Guoning Hu and DeLiang Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–553–I–556, Orlando, FL, USA, May 2002. IEEE.

[279] H.T. Hu. Robust pitch estimation based on modified comb filtering approach. *Electronics Letters*, 43(25):1471, 2007.

[280] Hwai-Tsu Hu, Chu Yu, and Chih-Hang Lin. Usefulness of the Comb Filtering Output for Voiced/Unvoiced Classification and Pitch Detection. In *2009 International Conference on Signal Processing Systems*, pages 135–139, Singapore, 2009. IEEE.

[281] Hu Weiping, Liang Yaling, Du Minghui, and Wang Xiuxin. A Novel Pitch Period Detection Algorithm Bases on HHT with Application to Normal and Pathological Voice. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4541–4544, Shanghai, China, 2005. IEEE.

[282] Kanru Hua. Improving YANGsaf F0 Estimator with Adaptive Kalman Filter. In *Interspeech 2017*, pages 2301–2305. ISCA, August 2017.

[283] Kanru Hua. Nebula: F0 Estimation and Voicing Detection by Modeling the Statistical Properties of Feature Extractors. In *Interspeech 2018*, pages 337–341. ISCA, September 2018.

[284] Feng Huang and Peter Balazs. Dictionary learning for pitch estimation in speech signals. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Tokyo, September 2017. IEEE.

[285] Feng Huang and Tan Lee. Pitch Estimation in Noisy Speech Based on Temporal Accumulation of Spectrum Peaks. *INTERSPEECH 2010*, page 4, 2010.

[286] Feng Huang and Tan Lee. Robust Pitch Estimation Using l1-regularized Maximum Likelihood Estimation. *INTERSPEECH 2012*, page 4, 2012.

[287] Feng Huang and Tan Lee. Sparsity-based confidence measure for pitch estimation in noisy speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4601–4604, Kyoto, Japan, March 2012. IEEE.

[288] Feng Huang and Tan Lee. Multipitch tracking based on linear programming relaxation and sparsity-based pitch candidate estimation. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 331–335, Singapore, Singapore, September 2014. IEEE.

[289] Feng Huang, Yu Ting Yeung, and Tan Lee. Evaluation of pitch estimation algorithms on separated speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6807–6811, Vancouver, BC, Canada, May 2013. IEEE.

[290] Qinghua Huang and Dongmei Wang. Multi-pitch Estimation for Speech Mixture Based on Multi-length Windows Harmonic Model. In *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, pages 345–348, Kunming and Lijiang City, China, April 2011. IEEE.

[291] W.-W. Hung. Use of fuzzy weighted autocorrelation function for pitch extraction from noisy speech. *Electronics Letters*, 38(19):1148, 2002.

[292] H Hussein, M Wolff, O Jokisch, F Duckhorn, G Strecha, and Ruediger Hoffmann. A Hybrid Speech Signal Based Algorithm for Pitch Marking Using Finite State Machines. *INTERSPEECH 2008*, page 4, September 2008.

[293] Hussein Hussein and Oliver Jokisch. Hybrid Electroglottograph and Speech Signal Based Algorithm for Pitch Marking. *INTERSPEECH 2007*, page 4, 2007.

[294] Hsin-Te Hwang, Chen-Yu Chiang, Po-Yi Sung, and Sin-Horng Chen. A Novel Model-Based Pitch Conversion Method for Mandarin Speech. *INTERSPEECH 2009*, page 4, 2009.

[295] Hyung Lae Kim, Dae Ho Kim, Young Sik Ryu, and Yung Kwon Kim. A study on pitch detection using the local peak and valley for Korean speech recognition. In *Proceedings of Digital Processing Applications (TENCON '96)*, volume 1, pages 107–112, Perth, WA, Australia, 1996. IEEE.

[296] Elliot Moore Ii and Juan Torres. Improving Glottal Waveform Estimation Through Rank-Based Glottal Quality Assessment. *INTERSPEECH 2006*, page 4, 2006.

[297] Akira Ikuta, Hisako Orimoto, and Yegui Xiao. A Bayesian approach for noise suppression of speech signal in real environment. *19th European Signal Processing Conference (EUSIPCO 2011)*, page 5, 2011.

[298] Ziba Imani and Seyed Jahanshah Kabudian. A Regularized Least Squares-Based Method for Optimal Fusion of Speech Pitch Detection Algorithms. In *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 119–123, Tehran, Iran, December 2018. IEEE.

[299] Ziba Imani and Seyed Jahanshah Kabudian. A Neural Network-Based Optimal Nonlinear Fusion of Speech Pitch Detection Algorithms. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pages 794–798, Tehran, Iran, February 2019. IEEE.

[300] Carlos T. Ishi, Dong Liang, Hiroshi Ishiguro, and Norihiro Hagita. The effects of microphone array processing on pitch extraction in real noisy environments. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 550–555, San Francisco, CA, September 2011. IEEE.

[301] Tatsuma Ishihara, Hirokazu Kameoka, Kota Yoshizato, Daisuke Saito, and Shigeki Sagayama. Probabilistic Speech F__0 Contour Model Incorporating Statistical Vocabulary Model of Phrase-Accent Command Sequence. *INTERSPEECH 2013*, page 5, 2013.

[302] Pooja Jain and Ram Bilas Pachori. Event-Based Method for Instantaneous Fundamental Frequency Estimation from Voiced Speech Based on Eigenvalue Decomposition of the Hankel Matrix. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1467–1482, October 2014.

[303] L. Janer. New pitch detection algorithm based on wavelet transform. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380)*, pages 165–168, Pittsburgh, PA, USA, 1998. IEEE.

[304] L. Janer, J.J. Bonet, and E. Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 1209–1212, Philadelphia, PA, USA, 1996. IEEE.

[305] Dalwon Jang, Sei-Jin Jang, and Seok-Pil Lee. Test of pitch extraction algorithms for query-by-singing/humming system. In *IEEE international Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–4, Seoul, Korea (South), June 2012. IEEE.

[306] Seung-Jin Jang, Seong-Hee Choi, Hyo-Min Kim, Hong-Shik Choi, and Young-Ro Yoon. Evaluation of Performance of Several Established Pitch Detection Algorithms in Pathological Voices. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 620–623, Lyon, France, August 2007. IEEE.

[307] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad. Fundamental frequency generation for whisper-to-audible speech conversion. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2579–2583, Florence, Italy, May 2014. IEEE.

[308] Matthias Janke and Lorenz Diener. EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2375–2385, December 2017.

[309] Alfredo Esquivel Jaramillo and Jesper Kj. On Optimal Filtering for Speech Decomposition. *2018 26th European Signal Processing Conference (EUSIPCO)*, page 5, 2018.

[310] Alfredo Esquivel Jaramillo, Jesper Kjar Nielsen, and Mads Grasboll Christensen. A Study on How Pre-whitening Influences Fundamental Frequency Estimation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6495–6499, Brighton, United Kingdom, May 2019. IEEE.

[311] Andrei Jefremov and W. Bastiaan Kleijn. Spline-based continuous-time pitch estimation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–337–I–340, Orlando, FL, USA, May 2002. IEEE.

[312] Jesper Rindom Jensen, Mads Christensen, Jacob Benesty, and Soren Jensen. Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1, 2014.

[313] Jesper Rindom Jensen, Mads Groesboll Christensen, and Soren Holdt Jensen. A single snapshot optimal filtering method for fundamental frequency estimation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4272–4275, Prague, Czech Republic, May 2011. IEEE.

[314] Jesper Rindom Jensen, Mads Groesboll Christensen, and Soren Holdt Jensen. Statistically efficient methods for pitch and DOA estimation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3900–3904, Vancouver, BC, Canada, May 2013. IEEE.

[315] Jesper Rindom Jensen, Mads Græsbøll Christensen, and Søren Holdt Jensen. Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):923–933, May 2013.

[316] Jesper Rindom Jensen, George-Othon Glentis, Mads Gr, and Søren Holdt Jensen. Computationally efficient IAA-based estimation of the fundamental frequency. *20th European Signal Processing Conference (EUSIPCO 2012)*, page 5, 2012.

[317] Jesper Rindom Jensen and Mads Gr. Fundamental frequency estimation using polynomial rooting of a subspace-based method. *18th European Signal Processing Conference (EUSIPCO 2010)*, page 5, 2010.

[318] Jesper Rindom Jensen and Mads Gr. Joint DOA and fundamental frequency estimation methods based on 2-D filtering. *18th European Signal Processing Conference (EUSIPCO 2010)*, page 5, 2010.

[319] Jesper Rindom Jensen and Mads Gr. DOA and pitch estimation of audio sources using IAA-based filtering. *22nd European Signal Processing Conference (EUSIPCO 2014)*, page 5, 2014.

[320] Tobias Lindstrom Jensen and Lieven Vandenberghe. Multi-pitch estimation using semidefinite programming. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4192–4196, New Orleans, LA, March 2017. IEEE.

[321] Jhing-Fa Wang, Chuan-I Tu, Ming-Hua Mo, and Shun-Chieh Lin. The design of a CAMDF-based pitch recognition embedded module and its applications for mobile consumer devices. In *TENCON 2007 - 2007 IEEE Region 10 Conference*, pages 1–5, Taipei, Taiwan, October 2007. IEEE.

[322] Du Jia, Chen Yanpu, Luo Hailong, and Yang Junqiang. An Adaptive Pitch Estimation Algorithm Based on AMDF. In Dehuai Zeng, editor, *Advances in Information Technology and Industry Applications*, volume 136, pages 187–194. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[323] Jian-Qi Yin, Xixian Chen, and Chun-lin Qin. Super resolution pitch determination based on cross-correlation and interpolation of speech signals. In *[Proceedings] Singapore ICCS/ISITA '92*, pages 410–414, Singapore, 1990. IEEE.

[324] Linlin Jiang, Shenghui Zhao, Jing Wang, and Jingming Kuang. Pitch prediction in frequency domain for ITU-T G.719 audio codec. In *2012 5th International Congress on Image and Signal Processing*, pages 1606–1610, Chongqing, Sichuan, China, October 2012. IEEE.

[325] Jianling Hu, Sheng Xu, and Jian Chen. A modified pitch detection algorithm. *IEEE Communications Letters*, 5(2):64–66, 2001.

[326] He Jiao, He Zhimi, and Xie Chaocheng. Pitch Detection Algorithm Based on NCCF and CAMDF. In *2012 Fourth International Conference on Computational Intelligence and Communication Networks*, pages 661–664, Mathura, Uttar Pradesh, India, November 2012. IEEE.

[327] Jinfu Ni, Shinsuke Sakai, Tohru Shimizu, and Satoshi Nakamura. CART-based modeling of Chinese tonal patterns with a functional model tracing the fundamental frequency trajectories. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4253–4256, Taipei, Taiwan, April 2009. IEEE.

[328] Wang Jingfang and Xu Huiyan. Morphological filter and Daubechies wavelet pitch detection. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 2135–2138, Ningbo, China, September 2011. IEEE.

[329] Jinhai Cai and Zhi-Qiang Liu. Robust pitch detection of speech signals using steerable filters. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1427–1430, Munich, Germany, 1997. IEEE Comput. Soc. Press.

[330] Wided Jlassi, Aicha Bouzid, and Noureddine Ellouze. A new method for pitch smoothing. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 657–661, Monastir, Tunisia, March 2016. IEEE.

[331] Wided Jlassi, Aicha Bouzid, and Noureddine Ellouze. Pitch Estimation Based on the Cepstrum Analysis by the Multi Scale Product of Clean and Noisy Speech. In Anna Esposito, Marcos Faundez-Zanuy, Antonietta M. Esposito, Gennaro Cordasco, Thomas Drugman, Jordi Solé-Casals, and Francesco Carlo Morabito, editors, *Recent Advances in Nonlinear Speech Processing*, volume 48, pages 219–225. Springer International Publishing, Cham, 2016.

[332] Seokhwan Jo, Sihyun Joo, and Chang D Yoo. Melody Pitch Estimation Based on Range Estimation and Candidate Extraction Using Harmonic Structure Model. *INTERSPEECH 2010*, page 4, 2010.

[333] Wangrae Jo, Jongkuk Kim, and Myung Jin Bae. A Study on Pitch Detection in Time-Frequency Hybrid Domain. *CICLing 2005*, page 4, 2005.

[334] Dominik Joho, Maren Bennewitz, and Sven Behnke. Pitch Estimation using Models of Voiced Speech on Three Levels. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages IV–1077–IV–1080, Honolulu, HI, April 2007. IEEE.

[335] S. Jones, R. Meddis, S.C. Lim, and A.R. Temple. Toward a digital neuromorphic pitch extraction system. *IEEE Transactions on Neural Networks*, 11(4):978–987, July 2000.

[336] JooHun Lee, HongYeol Jeon, MyungJin Bae, and SouGuil Ann. A fast pitch searching algorithm using correlation characteristics in CELP vocoder. In *Proceedings of MILCOM '94*, pages 699–702, Fort Monmouth, NJ, USA, 1994. IEEE.

[337] Joohun Lee, Myungjin Bae, and Hahyoung Yoo. A new fast pitch search algorithm using the abbreviated correlation function in CELP vocoder. In *Proceedings of MILCOM '96 IEEE Military Communications Conference*, volume 2, pages 653–657, McLean, VA, USA, 1996. IEEE.

[338] JooHun Lee, MyungJin Bae, Souguil Ann, and Hahyoung You. The skipping technique: a simple and fast algorithm to find the pitch in CELP vocoder. In *Proceedings of MILCOM '95*, volume 3, pages 1263–1266, San Diego, CA, USA, 1995. IEEE.

[339] Denis Jouvet and Yves Laprie. Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1614–1618, Kos, Greece, August 2017. IEEE.

[340] Zhang Jun and Wang Heping. A New Approach of Pitch Detection Based on Morphology Filter and Wavelet Transform. In *2010 Second International Workshop on Education Technology and Computer Science*, pages 751–753, Wuhan, China, 2010. IEEE.

[341] E. Jung, A. Schwarzbacher, and R. Lawlor. Implementation of real-time AMDF pitch-detection for voice gender normalisation. In *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)*, volume 2, pages 827–830, Santorini, Greece, 2002. IEEE.

[342] S. Kadambe and G.F. Boudreaux-Bartels. A COMPARISON OF A WAVELET TRANSFORM EVENT DETECTION PITCH DETECTOR WITH CLASSICAL PITCH DETECTORS. In *1990 Conference Record Twenty-Fourth Asilomar Conference on Signals, Systems and Computers, 1990.*, volume 2, page 1073, Pacific Grove, CA, 1990. IEEE.

[343] S. Kadambe and G.F. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Transactions on Information Theory*, 38(2):917–924, March 1992.

[344] S. Kadambe and G. F. Bourdeaux-Bartels. A comparison of a wavelet functions for pitch detection of speech signals. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 449–452 vol.1, April 1991.

[345] Shubha Kadambe and Gloria F. Boudreaux-Bartels. A Pitch Detector Based on Event Detection Using the Dyadic Wavelet Transform. In *First International Conference on Spoken Language Processing (ICSLP 90)*, Kobe, Japan, November 1990.

[346] N.A. Kader. Pitch detection algorithm using a wavelet correlation model. In *Proceedings of the Seventeenth National Radio Science Conference. 17th NRSC'2000 (IEEE Cat. No.00EX396)*, pages C33/1–C33/8, Minufiya, Egypt, 2000. Minufiya Univ.

[347] Cansu Kadi, Seda Gokhuseyin, and Yard.Doc.Umut Arioz. A study on compare pitch detection algorithms. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 132–135, Malatya, Turkey, May 2015. IEEE.

[348] Sudarsana Reddy Kadiri and Bayya Yegnanarayana. Estimation of Fundamental Frequency from Singing Voice Using Harmonics of Impulse-like Excitation Source. In *Interspeech 2018*, pages 2319–2323. ISCA, September 2018.

[349] Ozlem Kalinli. Tone and pitch accent classification using auditory attention cues. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5208–5211, Prague, Czech Republic, May 2011. IEEE.

[350] H Kameoka, N Ono, and S Sagayama. Speech Spectrum Modeling for Joint Estimation of Spectral Envelope and Fundamental Frequency. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1507–1516, August 2010.

[351] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Multi-Pitch Trajectory Estimation of Concurrent Speech Based on Harmonic GMM and Nonlinear Kalman Filtering. *INTERSPEECH 2004*, page 4, 2004.

[352] Hirokazu Kameoka, Kota Yoshizato, Tatsuma Ishihara, Yasunori Ohishi, Kunio Kashino, and Shigeki Sagayama. Generative Modeling of Speech F_0 Contours. *INTERSPEECH 2013*, page 5, 2013.

[353] Hirokazu Kameoka, Kota Yoshizato, Tatxsuma Ishihara, Kento Kadowaki, Yasunori Ohishi, and Kunio Kashino. Generative Modeling of Voice Fundamental Frequency Contours. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):1042–1053, June 2015.

[354] Magdalena Kaniewska. On the use of instantaneous complex frequency for pitch and formant tracking. *New Trends in Audio and Video / Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2008*, page 5, 2008.

[355] Magdalena Kaniewska. Speech formant frequency and pitch estimation using instantaneous complex frequency. In *2008 International Conference on Signals and Electronic Systems*, pages 493–496, Krakow, Poland, 2008. IEEE.

[356] Magdalena Kaniewska. Instantaneous complex frequency for pipeline pitch estimation. *Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2010*, page 6, 2010.

[357] Magdalena Kaniewska. Online pitch estimation using instantaneous complex frequency. In *2011 20th European Conference on Circuit Theory and Design (ECCTD)*, pages 393–396, Linkoping, Sweden, August 2011. IEEE.

[358] Sam Karimian-Azari, Andreas Jakobsson, Jesper R. Jensen, and Mads G. Christensen. Multi-pitch estimation and tracking using Bayesian inference in block sparsity. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 16–20, Nice, August 2015. IEEE.

[359] Sam Karimian-Azari, Jesper Rindom Jensen, and Mads Grasboll Christensen. Computationally Efficient and Noise Robust DOA and Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1613–1625, September 2016.

[360] Sam Karimian-Azari, Jesper Rindom Jensen, and Mads Groesboll Christensen. Fundamental frequency and model order estimation using spatial filtering. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5964–5968, Florence, May 2014. IEEE.

[361] Sam Karimian-Azari, Jesper Rindom Jensen, and Mads Gr. Fast joint DOA and pitch estimation using a broadband MVDR beamformer. *21st European Signal Processing Conference (EUSIPCO 2013)*, page 5, 2013.

[362] Sam Karimian-Azari, Jesper Rindom Jensen, and Mads Gr. Robust pitch estimation using an optimal filter on frequency estimates. *22nd European Signal Processing Conference (EUSIPCO 2014)*, page 5, 2014.

[363] Sam Karimian-Azari, Nasser Mohammadiha, Jesper R. Jensen, and Mads G. Christensen. Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4330–4334, South Brisbane, Queensland, Australia, April 2015. IEEE.

[364] M. Karjalainen and T. Tolonen. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 929–932 vol.2, Phoenix, AZ, USA, 1999. IEEE.

[365] Kavita Kasi and Stephen A. Zahorian. Yet Another Algorithm for Pitch Tracking. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–361–I–364, Orlando, FL, USA, May 2002. IEEE.

[366] Akihiro Kato and Tomi Kinnunen. Waveform to Single Sinusoid Regression to Estimate the F0 Contour from Noisy Speech Using Recurrent Deep Neural Networks. In *Interspeech 2018*, pages 327–331. ISCA, September 2018.

[367] Hideki Kawahara. An Instantaneous-Frequency-Based Pitch Extraction Method for High-Quality Speech Transformation: Revised TEMPO in the STRAIGHT-Suite. *5th International Conference on Spoken Language Processing (ICSLP 98)*, page 4, December 1998.

[368] Hideki Kawahara, Yannis Agiomyrgiannakis, and Heiga Zen. Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis. *arXiv:1605.07809 [cs, eess]*, pages 221–228, September 2016. arXiv: 1605.07809.

[369] Hideki Kawahara, Masanori Morise, Ryuichi Nisimura, and Toshio Irino. Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation. *INTERSPEECH 2012*, page 4, 2012.

[370] Hideki Kawahara, Ken-Ichi Sakakibara, Masanori Morise, Hideki Banno, and Tomoki Toda. A Modulation Property of Time-Frequency Derivatives of Filtered Phase and its Application to Aperiodicity and fo Estimation. In *Interspeech 2017*, pages 424–428. ISCA, August 2017.

[371] Tomonori Kawamura, Atsuhiko Kai, and Seiichi Nakagawa. Noise Robust Fundamental Frequency Estimation of Speech using CNN-based discriminative modeling. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 60–65, Krabi, August 2018. IEEE.

[372] Marian Kepesi, Lukas Ottowitz, and Tania Habib. Joint Position-Pitch Estimation for Multiple Speaker Scenarios. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pages 85–88, Trento, Italy, May 2008. IEEE.

[373] Marian Kepesi, Franz Pernkopf, and Michael Wohlmayr. Joint Position-Pitch Tracking for 2-Channel Audio. In *2007 International Workshop on Content-Based Multimedia Indexing*, pages 303–306, Talence, France, June 2007. IEEE.

[374] Marian Kepesi and Luis Weruaga. High-Resolution Noise-Robust Spectral-Based Pitch Estimation. *INTERSPEECH 2005*, page 4, 2005.

[375] S. S. Kharchenko, R. V. Mescheryakov, D. A. Volf, L. N. Balatskaya, and E. L. Choinzonov. Fundamental frequency evaluation subsystem for natural speech rehabilitation software calculation module for cancer patients after larynx resection. In *2015 International Conference on Biomedical Engineering and Computational Technologies (SIBIRCON)*, pages 197–200, Novosibirsk, Russia, October 2015. IEEE.

[376] A.A. Khulage. Extraction of pitch, duration and formant frequencies for emotion recognition system. In *Fourth International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)*, pages 7–9, Bangalore, India, 2012. Institution of Engineering and Technology.

[377] A. Khurshid and S.L. Denham. A Temporal-Analysis-Based Pitch Estimation System for Noisy Speech With a Comparative Study of Performance of Recent Systems. *IEEE Transactions on Neural Networks*, 15(5):1112–1124, September 2004.

[378] Mohammed Kamal Khwaja, Sunil Sivadas, and P. Arulmozhivarman. Pitch tracking in reverberant environments. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 192–196, Abu Dhabi, United Arab Emirates, December 2015. IEEE.

[379] Yusuke Kida, Masaru Sakai, Takashi Masuko, and Akinori Kawamura. Robust F0 Estimation Based on Log-Time Scale Autocorrelation and its Application to Mandarin Tone Recognition. *INTERSPEECH 2009*, page 4, 2009.

[380] Han-Gyu Kim, Gil-Jin Jang, Yung-Hwan Oh, and Ho-Jin Choi. Speech and music pitch trajectory classification using recurrent neural networks for monaural speech segregation. *The Journal of Supercomputing*, February 2019.

[381] Han-Gyu Kim, Gil-Jin Jang, Jeong-Sik Park, and Yung-Hwan Oh. Monaural Speech Segregation Based on Pitch Track Correction Using an Ensemble Kalman Filter. *INTERSPEECH 2013*, page 4, 2013.

[382] Han-Gyu Kim, Jeong-Sik Park, Gil-Jin Jang, and Yung-Hwan Oh. Particle filtering by sigmoidal weight update for speech pitch correction. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2574–2579, Seoul, Korea (South), October 2012. IEEE.

[383] Hyun Soo Kim. Morphological Pre-Processing Technique and Its Applications on Speech Signal. *INTERSPEECH 2007*, page 4, 2007.

[384] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A Convolutional Representation for Pitch Estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, Calgary, AB, April 2018. IEEE.

[385] Jongkuk Kim, Ki Young Lee, and Myung Jin Bae. On a Pitch Detection Method Using Noise Reduction. *CICLing 2005*, page 4, 2005.

[386] D.R. Kipke and K.L. Levey. An application of neural speech processing in the cochlear nucleus to the estimation of fundamental frequency. In *Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society*, volume 2, pages 973–974, Montreal, Que., Canada, 1995. IEEE.

[387] A. Klapuri. Pitch estimation using multiple independent time-frequency windows. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No.99TH8452)*, pages 115–118, New Paltz, NY, USA, 1999. IEEE.

[388] Anssi Klapuri. Auditory Model-Based Methods for Multiple Fundamental Frequency Estimation. In Anssi Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription*, pages 229–265. Springer US, Boston, MA, 2006.

[389] A.P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–816, November 2003.

[390] W. B. Kleijn, R. P. Ramachandran, and P. Kroon. Generalized analysis-by-synthesis coding and its application to pitch prediction. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 337–340 vol.1, March 1992.

[391] W.B. Kleijn. Improved pitch prediction. In *Proceedings., IEEE Workshop on Speech Coding for Telecommunications,*, pages 19–20, Quebec, Canada, 1993. IEEE.

[392] H. Kobayashi and T. Shimamura. A modified cepstrum method for pitch extraction. In *IEEE. APCCAS 1998. 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings (Cat. No.98EX242)*, pages 299–302, Chiangmai, Thailand, 1998. IEEE.

[393] H. Kobayashi and T. Shimamura. A weighted autocorrelation method for pitch extraction of noisy speech. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1307–1310, Istanbul, Turkey, 2000. IEEE.

[394] Karishma Kolhatkar, Mahesh Kolte, and Jyoti Lele. Implementation of pitch detection algorithms for pathological voices. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, pages 1–5, Coimbatore, India, August 2016. IEEE.

[395] A. S. Kolokolov and I. A. Lyubinskii. Measuring the Pitch of a Speech Signal Using the Autocorrelation Function. *Automation and Remote Control*, 80(2):317–323, February 2019.

[396] A. Koretz and J. Tabrikian. Maximum A Posteriori Probability Multiple-Pitch Tracking Using the Harmonic Model. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2210–2221, September 2011.

[397] Martin Krawczyk-Becker and Timo Gerkmann. Fundamental Frequency Informed Speech Enhancement in a Flexible Statistical Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):940–951, May 2016.

[398] Mohamed Krini and Gerhard Schmidt. Spectral Refinement and its Application to Fundamental Frequency Estimation. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 251–254, New Paltz, NY, USA, October 2007. IEEE.

[399] S. Krishnakumar, K.R.P. Kumar, and N. Balakrishnan. Pitch maxima for robust speaker recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 2, pages II–201–4, Hong Kong, China, 2003. IEEE.

[400] Oraphan Krityakien, Keikichi Hirose, and Nobuaki Minematsu. Generation of Fundamental Frequency Contours for Thai Speech Synthesis Using Tone Nucleus Model. *INTERSPEECH 2013*, page 5, 2013.

[401] Ted Kronvall, Stefan Ingi Adalbjornsson, and Andreas Jakobsson. Joint DOA and multi-pitch estimation using block sparsity. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3958–3962, Florence, Italy, May 2014. IEEE.

[402] Ted Kronvall, Stefan Ingi Adalbjornsson, and Andreas Jakobsson. Joint DOA and multi-pitch estimation via block sparse dictionary learning. *22nd European Signal Processing Conference (EUSIPCO 2014)*, page 5, 2014.

[403] Ted Kronvall, Filip Elvander, Stefan Ingi Adalbjornsson, and Andreas Jakobsson. Multi-pitch estimation via fast group sparse learning. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1093–1097, Budapest, Hungary, August 2016. IEEE.

[404] P. Kroon and B. S. Atal. On the use of pitch predictors with high temporal resolution. *IEEE Transactions on Signal Processing*, 39(3):733–735, March 1991.

[405] P. Kroon and B.S. Atal. Pitch predictors with high temporal resolution. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 661–664, Albuquerque, NM, USA, 1990. IEEE.

[406] D. A. Krubsack and R. J. Niederjohn. An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech. *IEEE Transactions on Signal Processing*, 39(2):319–329, February 1991.

[407] Chih-Yi Kuan, Li Su, Yu-Hao Chin, and Jia-Ching Wang. Multi-pitch streaming of interwoven streams. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315, New Orleans, LA, March 2017. IEEE.

[408] M K Prasanna Kumar and R Kumaraswamy. Role of f0 and formant frequencies in unsupervised separation of convolutive speech mixtures. In *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pages 316–320, Davangere, Karnataka, India, October 2015. IEEE.

[409] Sandeep Kumar. Performance Evaluation of Novel AMDF-Based Pitch Detection Scheme. *ETRI Journal*, January 2016.

[410] Sandeep Kumar. Performance measurement of a novel pitch detection scheme based on weighted autocorrelation for speech signals. *International Journal of Speech Technology*, 22(4):885–892, December 2019.

[411] Sandeep Kumar, S. Bhattacharya, Vishal Dhiman, and Shuvashree Mohapatra. Performance evaluation of a wavelet-based pitch detection scheme. *International Journal of Speech Technology*, 16(4):431–437, December 2013.

[412] Sandeep Kumar, S. Bhattacharya, and Premanand Patel. A new pitch detection scheme based on ACF and AMDF. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pages 1235–1240, Ramanathapuram, India, May 2014. IEEE.

[413] Sandeep Kumar, Satish Kumar Singh, and S. Bhattacharya. Performance evaluation of a ACF-AMDF based pitch detection scheme in real-time. *International Journal of Speech Technology*, 18(4):521–527, December 2015.

[414] Balachandra Kumaraswamy and P G Poonacha. Modified square difference function using fourier series approximation for pitch estimation. In *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, pages 1–8, Chennai, February 2017. IEEE.

[415] Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani. Multiple pitch identification using cochlear-like frequency capture and harmonic grouping. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 613–616, Prague, Czech Republic, May 2011. IEEE.

[416] Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani. Auditory-inspired pitch extraction using a Synchrony Capture Filterbank and phase alignment. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5959–5963, Florence, Italy, May 2014. IEEE.

[417] N. Kunieda, T. Shimamura, and J. Suzuki. Robust method of measurement of fundamental frequency by ACLOS: autocorrelation of log spectrum. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 232–235, Atlanta, GA, USA, 1996. IEEE.

[418] Y.-H. Kwon, D.-J. Park, and B.-C. Ihm. Simplified pitch detection algorithm of mixed speech signals. In *2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No.00CH36353)*, volume 3, pages 722–725, Geneva, Switzerland, 2000. Presses Polytech. Univ. Romandes.

[419] S. Kwong, W. Gang, and O.Y.J. Zheng. Fundamental frequency estimation based on adaptive time-averaging Wigner-Ville distribution. In *[1992] Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 413–416, Victoria, BC, Canada, 1992. IEEE.

[420] Lasse Laaksonen and Anssi Ramo. Using noise reduction in mode selection and pitch search. In *2008 2nd International Conference on Signal Processing and Communication Systems*, pages 1–6, Gold Coast, Australia, December 2008. IEEE.

[421] Everton B. Lacerda and Carlos A. B. Mello. A Pitch Extraction System Based on Laryngeal Mechanisms Classification. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, Rio de Janeiro, July 2018. IEEE.

[422] Mahsa Sadat Elyasi Langarani, Esther Klabbers, and Jan van Santen. A novel pitch decomposition method for the generalized linear alignment model. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2584–2588, Florence, Italy, May 2014. IEEE.

[423] Yves Laprie and Vincent Colotte. Automatic pitch marking for speech transformations via TD-PSOLA. *9th European Signal Processing Conference (EUSIPCO 1998)*, page 4, September 1998.

[424] Kornel Laskowski, Mattias Heldner, and Jens Edlund. Exploring the prosody of floor mechanisms in english using the fundamental frequency variation spectrum. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[425] Javier Latorre, M J F Gales, and Heiga Zen. Training a Parametric-Based LogF0 Model with the Minimum Generation Error Criterion. *INTERSPEECH 2010*, page 4, 2010.

[426] Rustam Latypov, Ruslan Nigmatullin, and Evgeni Stolov. Instantaneous frequency and detection of dynamics in speech. In *2017 International Symposium ELMAR*, pages 141–144, Zadar, September 2017. IEEE.

[427] Byung Suk Lee and Daniel P W Ellis. Noise Robust Pitch Tracking by Subband Autocorrelation Classification. *INTERSPEECH 2012*, page 4, 2012.

[428] Jaehyung Lee and Soo-Young Lee. Robust Fundamental Frequency Estimation Combining Contrast Enhancement and Feature Unbiasing. *IEEE Signal Processing Letters*, 15:521–524, 2008.

[429] S W Lee, Frank K Soong, and P C Ching. Harmonic Filtering for Joint Estimation of Pitch and Voiced Source with Single-Microphone Input. *INTERSPEECH 2005*, page 4, 2005.

[430] S. W. Lee, Frank K. Soong, P. C. Ching, and Tan Lee. Pitch Tracking for Model-Based Speech Separation. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4, Kunming, China, December 2008. IEEE.

[431] Adrian Leemann, Keikichi Hirose, and Hiroya Fujisaki. Analysis of Voice Fundamental Frequency Contours of Continuing and Terminating Prosodic Phrases in Four Swiss German Dialects. *INTERSPEECH 2009*, page 4, 2009.

[432] Milan Legat, Jindrich Matousek, and Daniel Tihelka. A Robust Multi-Phase Pitch-Mark Detection Algorithm. *INTERSPEECH 2007*, page 4, August 2007.

[433] Ming Lei, Yijian Wu, Frank K Soong, Zhen-Hua Ling, and Lirong Dai. A Hierarchical F0 Modeling Method for HMM-Based Speech Synthesis. *INTERSPEECH 2010*, page 4, 2010.

[434] Bogu Li, Zhilei Liu, and Jianwu Dang. Study on the relation of fundamental and formant frequencies for affective speech synthesis. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5, Tianjin, China, October 2016. IEEE.

[435] Dixi Li, Shiwei Tang, Xiaoxue Han, Feng Yu, and Li Zhao. A fundamental frequency tracking and note segmentation algorithm tailored for Karaoke autoscoring. In *2010 3rd International Congress on Image and Signal Processing*, pages 3519–3523, Yantai, China, October 2010. IEEE.

[436] Ming Li, Chuan Cao, Di Wang, Ping Lu, Qiang Fu, and Yonghong Yan. Cochannel Speech Separation Using Multi-Pitch Estimation and Model Based Voiced Sequential Grouping. *INTERSPEECH 2008*, page 4, September 2008.

[437] Ru-wei Li, Chang-chun Bao, and Hui-jing Dou. Pitch detection method for noisy speech signals based on pre-filter and weighted wavelet coefficients. In *2008 9th International Conference on Signal Processing*, pages 530–533, Beijing, China, October 2008. IEEE.

[438] Xiao Li, Jonathan Malkin, and Jeff Bilmes. Graphical Model Approach to Pitch Tracking. *INTERSPEECH 2004*, page 4, 2004.

[439] Xingda Li, Yujing Guan, Yingnian Wu, and Zhongbo Zhang. Piano multipitch estimation using sparse coding embedded deep learning. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1), December 2018.

[440] Yujia Li and Tan Lee. Perception-Based Automatic Approximation of F0 Contours in Cantonese Speech. *INTERSPEECH 2010*, page 4, 2010.

[441] Yusheng Li, Biao Xue, Hong Hong, and Xiaohua Zhu. Instantaneous pitch estimation based on empirical wavelet transform. In *2014 19th International Conference on Digital Signal Processing*, pages 250–253, Hong Kong, Hong Kong, August 2014. IEEE.

[442] Li Hui, Bei-Qian Dai, and Lu Wei. A Pitch Detection Algorithm Based on AMDF and ACF. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–377–I–380, Toulouse, France, 2006. IEEE.

[443] LI Jing and BAO Changchun. A pitch detector based on the dyadic wavelet transform and the autocorrelation function. In *6th International Conference on Signal Processing, 2002.*, pages 414–417, Beijing, China, 2002. IEEE.

[444] Liang Wang, Jie Zhu, and Yao Lv. An improved method for predicting fundamental frequency contour in mandarin text-to-speech system with a small corpus. In *TENCON 2010 - 2010 IEEE Region 10 Conference*, pages 751–754, Fukuoka, November 2010. IEEE.

[445] Yu-An S. Lien and Cara E. Stepp. Automated estimation of relative fundamental frequency. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2136–2139, Osaka, July 2013. IEEE.

[446] Jean-Sylvain Lienard, Francois Signol, and Claude Barras. Speech Fundamental Frequency Estimation Using the Alternate Comb. *INTERSPEECH 2007*, page 4, 2007.

[447] S.C. Lim, A.R. Temple, S. Jones, and R. Meddis. VHDL-based design of biologically inspired pitch detection system. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, volume 2, pages 922–927, Houston, TX, USA, 1997. IEEE.

[448] Cheng-Yuan Lin, Chien-Hung Huang, and Chih-Chung Kuo. A simple and effective pitch re-estimation method for rich prosody and speaking styles in HMM-based speech synthesis. In *2012 8th International Symposium on Chinese Spoken Language Processing*, pages 286–290, Kowloon Tong, China, December 2012. IEEE.

[449] Shoufeng Lin. Robust Pitch Estimation and Tracking For Speakers Based on Subband Encoding and The Generalized Labeled Multi-Bernoulli Filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):827–841, April 2019.

[450] Bin Liu, Fuyuan Mo, and Jianhua Tao. Speech enhancement based on analysis-synthesis framework with improved pitch estimation and spectral envelope enhancement. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 461–466, Hangzhou, Zhejiang, China, October 2014. IEEE.

[451] Bin Liu, Jianhua Tao, Dawei Zhang, and Yibin Zheng. A novel pitch extraction based on jointly trained deep BLSTM Recurrent Neural Networks with bottleneck features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340, New Orleans, LA, March 2017. IEEE.

[452] Jian Liu, Thomas Fang Zheng, Jing Deng, and Wenhu Wu. Real-Time Pitch Tracking Based on Combined SMDSF. *INTERSPEECH 2005*, page 4, 2005.

[453] Yuzhou Liu and DeLiang Wang. Speaker-dependent multipitch tracking using deep neural networks. *INTERSPEECH 2015*, 141(2):710–721, 2015.

[454] Yuzhou Liu and DeLiang Wang. Robust pitch tracking in noisy speech using speaker-dependent deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5255–5259, Shanghai, March 2016. IEEE.

[455] Yuzhou Liu and DeLiang Wang. Time and frequency domain long short-term memory for noise robust pitch tracking. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5600–5604, New Orleans, LA, March 2017. IEEE.

[456] Yuzhou Liu and DeLiang Wang. Permutation Invariant Training for Speaker-Independent Multi-Pitch Tracking. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5594–5598, Calgary, AB, April 2018. IEEE.

[457] Lizhi Wang, Guiping Hu, and Zengfu Wang. Pitch detection based on time-frequency analysis. In *Fifth World Congress on Intelligent Control and Automation (IEEE Cat. No.04EX788)*, volume 4, pages 3022–3026, Hangzhou, China, 2004. IEEE.

[458] J. Logan and J. Gowdy. Adaptive pitch period decimation and its application in speech compression. In *Proceedings of SOUTHEASTCON '96*, pages 220–222, Tampa, FL, USA, 1996. IEEE.

[459] Damien Lolive, Nelly Barbot, and Olivier Boeffard. B-Spline Model Order Selection With Optimal MDL Criterion Applied to Speech Fundamental Frequency Stylization. *IEEE Journal of Selected Topics in Signal Processing*, 4(3):571–581, June 2010.

[460] J.V. Lorenzo Ginori and M.E. Hernández-Diaz Huici. Combined algorithm for pitch detection of speech signals. *Electronics Letters*, 31(1):15–16, January 1995.

[461] Erfan Loweimi, Jon Barker, and Thomas Hain. On the Usefulness of the Speech Phase Spectrum for Pitch Extraction. In *Interspeech 2018*, pages 696–700. ISCA, September 2018.

[462] Yang Lu and Philipos C. Loizou. Speech enhancement by combining statistical estimators of speech and noise. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4754–4757, Dallas, TX, USA, 2010. IEEE.

[463] Lu Qin, Qiang Li, and Xin Guan. Pitch extraction for musical signals with modified AMDF. In *2011 International Conference on Multimedia Technology*, pages 3599–3602, Hangzhou, China, July 2011. IEEE.

[464] Iker Luengo, Ibon Saratxaga, Eva Navas, Inma Hernaez, Javier Sanchez, and Iñaki Sainz. Evaluation of pitch detection algorithms under real conditions. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1057. IEEE, 2007.

[465] Lunji Qiu, Haiyun Yang, and Soo Ngee Koh. A fundamental frequency detector of speech signals based on short time Fourier transform. In *Proceedings of TENCON'94 - 1994 IEEE Region 10's 9th Annual International Conference on: 'Frontiers of Computer Technology'*, pages 526–530, Singapore, 1994. IEEE.

[466] Lunji Qiu, Soo-Ngee Koh, and Haiyun Yang. Pitch determination of noisy speech using wavelet transform in time and frequency domains. In *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, pages 337–340, Beijing, China, 1993. IEEE.

[467] H.Y. Luo and P.N. Denbigh. Improved pitch-tracking for the separation of two overlapping voices. In *Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Societ*, pages 1389–1390, San Diego, CA, 1993. IEEE.

[468] H. Maalem and F. Marir. The fourth order cumulant of speech signals applied to pitch estimation. In *2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT '04.*, volume 3, pages 1303–1306, Hammemet, Tunsia, 2004. IEEE.

[469] M.D. Macleod. Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones. *IEEE Transactions on Signal Processing*, 46(1):141–148, January 1998.

[470] Ratko Magjarevic, G. Schlotthauer, M. E. Torres, and H. L. Rufiner. Voice Fundamental Frequency Extraction Algorithm Based on Ensemble Empirical Mode Decomposition and Entropies. In Olaf Dössel and Wolfgang C. Schlegel, editors, *World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany*, volume 25/4, pages 984–987. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[471] Vijay Mahadevan and Carol Y. Espy-Wilson. Maximum likelihood pitch estimation using sinusoidal modeling. In *2011 International Conference on Communications and Signal Processing*, pages 310–314, Kerala, India, February 2011. IEEE.

[472] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh. Determination of pitch range based on onset and offset analysis in modulation frequency domain. In *2010 5th International Symposium on Telecommunications*, pages 604–608, Tehran, Iran, December 2010. IEEE.

[473] A. Mahmoodzadeh, H. R. Abutalebi, H. Soltanian-Zadeh, and H. Sheikhzadeh. Single channel speech separation with a frame-based pitch range estimation method in modulation frequency. In *2010 5th International Symposium on Telecommunications*, pages 609–613, Tehran, Iran, December 2010. IEEE.

[474] J.A. Maidment and M.L.G. Lecumberri. Pitch analysis methods for cross-speaker comparison. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 4, pages 2247–2249, Philadelphia, PA, USA, 1996. IEEE.

[475] N. Malik and W.H. Holmes. Pitch estimation and a measure of voicing from pseudo-spectra. In *ISSPA '99. Proceedings of the Fifth International Symposium on Signal Processing and its Applications (IEEE Cat. No.99EX359)*, volume 1, pages 59–62, Brisbane, Qld., Australia, 1999. Queensland Univ. Technol.

[476] Nicolas Malyska and Thomas F Quatieri. Analysis of Nonmodal Phonation Using Minimum Entropy Deconvolution. *INTERSPEECH 2006*, page 4, 2006.

[477] Mangui Liang, Qi Hu, Jie Peng, and Guoqiao Yu. Pitch Detection with Fuzzy Computation. In *First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06)*, volume 1, pages 620–624, Beijing, China, 2006. IEEE.

[478] Iain Mann and Steve McLaughlin'. POINCAR& MAPS AND PITCH DETECTION IN SPEECH. *IEE Colloquium on Signals Systems and Chaos 1997*, page 5, 1997.

[479] Philippe Martin. Crosscorrelation of Adjacent Spectra Enhances Fundamental Frequency Tracking. *INTERSPEECH 2008*, page 4, September 2008.

[480] H. Martinez-Alfaro and J. L. Contreras-Vidal. A robust real-time pitch detector based on neural networks. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 521–523 vol.1, April 1991.

[481] Tomoko Matsui, Masataka Goto, Jean-Philippe Vert, and Yuji Uchiyama. Gradient-based musical feature extraction based on scale-invariant feature transform. *19th European Signal Processing Conference (EUSIPCO 2011)*, page 5, 2011.

[482] T. Matsuoka, N. Matusdaira, and N. Yamawaki. Application Of Synchronous Phenomenon In A Nervous System To The Pitch Period Detection Of Speech. In *[1990] Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 793–794, November 1990.

[483] Matthias Mauch and Simon Dixon. PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, Florence, Italy, May 2014. IEEE.

[484] Nadhifa Maulida, Wilujeng F. Alfiah, Desty A. Pawestri, Heru Susanto, M. Q. Zaman, and Dhany Arifianto. Fundamental frequency evaluation of infant crying. In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 61–66, Lombok, Indonesia, July 2016. IEEE.

[485] R.J. McAulay and T.F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 249–252, Albuquerque, NM, USA, 1990. IEEE.

[486] Robert John McAulay. Sine-wave based PSOLA pitch scaling with real-time pitch marking. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4, New Paltz, NY, USA, October 2013. IEEE.

[487] Matthew McCallum and Bernard Guillemin. Stochastic-Deterministic Signal Modelling for the Tracking of Pitch in Noise and Speech Mixtures Using Factorial HMMs. *INTERSPEECH 2013*, page 5, 2013.

[488] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*, 39(1):40–48, January 1991.

[489] P. Mermelstein and Y. Qian. Analysis by synthesis speech coding with generalized pitch prediction. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 1–4 vol.1, Phoenix, AZ, USA, 1999. IEEE.

[490] P. Mermelstein, P. Zeng, M. Saikaly, and Y. Qian. Multiband pitch and residual coding of speech signals. In *ICCT'98. 1998 International Conference on Communication Technology. Proceedings (IEEE Cat. No.98EX243)*, volume vol.2, page 4, Beijing, China, 1998. Publising House of Constr. Mater.

[491] M.A.B. Messaoud, A. Bouzid, and N. Ellouze. Using multi-scale product spectrum for single and multi-pitch estimation. *IET Signal Processing*, 5(3):344, 2011.

[492] Mohanmed Anouar Ben Messaoud, Aicha Bouzid, and Noureddine Ellouze. Spectral Multi-Scale Analysis for Multi-Pitch Tracking. In *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, pages 26–31, Marco Island, FL, USA, January 2009. IEEE.

[493] Ben Milner and Xu Shao. Prediction of Fundamental Frequency and Voicing From Mel-Frequency Cepstral Coefficients for Unconstrained Speech Reconstruction. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):24–33, January 2007.

[494] Ben Milner, Xu Shao, and Jonathan Darch. Fundamental Frequency and Voicing Prediction from MFCCs for Speech Reconstruction from Unconstrained Speech. *INTERSPEECH 2005*, page 4, 2005.

[495] Huaiping Ming, Dongyan Huang, Minghui Dong, Haizhou Li, Lei Xie, and Shaofei Zhang. Fundamental frequency modeling using wavelets for emotional voice conversion. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 804–809, Xi'an, China, September 2015. IEEE.

[496] Mingyang Wu, DeLiang Wang, and G.J. Brown. Pitch tracking based on statistical anticipation. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, volume 2, pages 866–871, Washington, DC, USA, 2001. IEEE.

[497] Kenichiro Miwa and Masashi Unoki. Robust Method for Estimating F0 of Complex Tone Based on Pitch Perception of Amplitude Modulated Signal. In *Interspeech 2017*, pages 2311–2315. ISCA, August 2017.

[498] Hansjorg Mixdorff and Hartmut R Pfitzinger. A Quantitative Study of F0 Peak Alignment and Sentence Modality. *INTERSPEECH 2009*, page 4, 2009.

[499] Toshihiro Miyawaki, Naoto Sasaoka, Yoshio Itoh, Kensaku Fujii, and Sumio Tsuiki. A Study on Pitch Detection of Sinusoidal Noise for Noise Reduction System. In *2006 International Symposium on Intelligent Signal Processing and Communications*, pages 311–314, Yonago, Japan, December 2006. IEEE.

[500] Musfir Mohammed. Implementation of an intelligent system for estimation of fundamental frequency of speech. *2011 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, page 5, 2011.

[501] Shuvashree Mohapatra, Vishal Dhiman, Sandeep Kumar, and S. Bhattacharya. A Theoretical Justification for Coincidence of Wavelet Maxima at a Particular Scale Pair in an Event-Based Pitch Detection Method. In *2011 International Conference on Devices and Communications (ICDeCom)*, pages 1–4, Mesra, Ranchi, India, February 2011. IEEE.

[502] Khademul Islam Molla, Keikichi Hirose, Nobuaki Minematsu, and Kamrul Hasan. Pitch Estimation of Noisy Speech Signals Using Empirical Mode Decomposition. *INTERSPEECH 2007*, page 4, 2007.

[503] Md. Khademul Islam Molla, Mahboob Qaosar, and Keikichi Hirose. Instantaneous pitch estimation of noisy speech signal with multivariate SST. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 770–773, Montréal, QC, Canada, May 2016. IEEE.

[504] A. Moreno and J. A. R. Fonollosa. Pitch determination of noisy speech using higher order statistics. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 133–136 vol.1, March 1992.

[505] Veronica Morfi, Gilles Degottex, and Athanasios Mouchtaris. A computationally efficient refinement of the fundamental frequency estimate for the Adaptive Harmonic Model. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1478–1482, Florence, Italy, May 2014. IEEE.

[506] Masanori Morise. Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. In *Interspeech 2017*, pages 2321–2325. ISCA, August 2017.

[507] Masanori Morise and Hideki Kawahara. TUSK: A Framework for Overviewing the Performance of F0 Estimators. In *INTERSPEECH 2016*, pages 1790–1794, September 2016.

[508] Sonia Moussa, Zied Hajaiej, and Ali Garsallah. Proposition of adaptative time-frequency representation of speech signal. In *2016 4th International Conference on Control Engineering & Information Technology (CEIT)*, pages 1–5, Hammamet, Tunisia, December 2016. IEEE.

[509] E. Mousset, W. A. Ainsworth, and J. A. R. Fonollosa. A comparison of several recent methods of fundamental frequency and voicing decision estimation. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 1273–1276 vol.2, October 1996.

[510] J. R. E. Moxham, P. A. Jones, H. J. McDermott, and G. M. Clark. A new algorithm for voicing detection and voice pitch estimation based on the neocognitron. In *Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop*, pages 204–213, August 1992.

[511] Dipesh Mudatkar, S. Adarsh, and D. Govind. Robust pitch estimation in distant speech signals collected from vehicle. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1784–1791, Udupi, September 2017. IEEE.

[512] Ghulam Muhammad. Noise Robust Pitch Detection Based on Extended AMDF. In *2008 IEEE International Symposium on Signal Processing and Information Technology*, pages 133–138, Sarajevo, Bosnia and Herzegovina, December 2008. IEEE.

[513] Sankar Mukherjee and Shyamal Kumar Das Mandal. Generation of F_0 Contour Using Deep Boltzmann Machine and Twin Gaussian Process Hybrid Model for Bengali Language. *INTERSPEECH 2014*, page 5, 2014.

[514] Takahiro Murakami, Munehiro Namba, Tetsuya Hoya, and Yoshihisa Ishida. Pitch Extraction of Speech Signals Using an Eigen-Based Subspace Method. *7th International Conference on Spoken Language Processing (ICSLP2002)*, page 4, 2002.

[515] Hema A. Murthy. Pitch Extraction From Root Cepstrum. In *Third International Conference on Spoken Language Processing (ICSLP 94)*, Yokohama, Japan, September 1994.

[516] Myung Jin Bae, Hwe Yoong Whang, and Hah Young Yoo. On a fast pitch searching by using a simple correlation technique in the CELP vocoder. In *38th Midwest Symposium on Circuits and Systems. Proceedings*, volume 2, pages 1256–1259, Rio de Janeiro, Brazil, 1996. IEEE.

[517] MyungJin Bae and WangRae Jo. On a fast pitch search of CELP type vocoder using decimation technique. In *Proceedings of Digital Processing Applications (TENCON '96)*, volume 1, pages 204–208, Perth, WA, Australia, 1996. IEEE.

[518] Na Yang, He Ba, Weiyang Cai, Ilker Demirkol, and Wendi Heinzelman. BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1833–1848, December 2014.

[519] C. Nadeu, J. Pascual, and J. Hernando. Pitch determination using the cepstrum of the one-sided autocorrelation sequence. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 3677–3680 vol.5, April 1991.

[520] G V S S K R Naganjaneyulu, M Venkata Ramana, and A.V. Narasimhadhan. A novel method for pitch detection via instantaneous frequency estimation using polynomial chirplet transform. In *2016 IEEE Region 10 Conference (TENCON)*, pages 1250–1253, Singapore, November 2016. IEEE.

[521] Keigo Nakamura, Matthias Janke, Michael Wand, and Tanja Schultz. Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 573–576, Prague, Czech Republic, May 2011. IEEE.

[522] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features. *INTERSPEECH 2006*, page 4, 2006.

[523] Tomohiro Nakatani and Toshio Irino. Robust Fundamental Frequency Estimation Against Background Noise and Spectral Distortion. *7th International Conference on Spoken Language Processing (ICSLP2002)*, page 4, 2002.

[524] Swagata Nandi and Debasis Kundu. Estimating the fundamental frequency of a periodic function: Estimating fundamental frequency. *Statistical Methods and Applications*, 12(3):341–360, February 2004.

[525] Antonio Napolitano. Asymptotic normality of cyclic autocorrelation estimate with estimated cycle frequency. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1481–1485, Nice, August 2015. IEEE.

[526] Antonio Napolitano. On cyclic spectrum estimation with estimated cycle frequency. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 160–164, Budapest, Hungary, August 2016. IEEE.

[527] Masatoshi Narita and Tetsuya Shimamura. Exponentiated enhancement for fundamental frequency extraction of noisy speech. In *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 342–346, Bilbao, Spain, December 2011. IEEE.

[528] Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose, and Hiroya Fujisaki. A method for automatic extraction of model parameters from fundamental frequency contours of speech. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–509–I–512, Orlando, FL, USA, May 2002. IEEE.

[529] A. Nehorai and B. Porat. Adaptive comb filtering for harmonic signal enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1124–1138, October 1986.

[530] Daniel Neiberg, G Ananthakrishnan, and Joakim Gustafson. Tracking Pitch Contours Using Minimum Jerk Trajectories. *INTERSPEECH 2011*, page 4, 2011.

[531] L.Y. Ngan, Y. Wu, H.C. So, P.C. Ching, and S.W. Lee. Joint time delay and pitch estimation for speaker localization. In *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03.*, volume 3, pages III–722–III–725, Bangkok, Thailand, 2003. IEEE.

[532] Kim Ngo, Toon van Waterschoot, Mads Gr, Marc Moonen, Søren Holdt Jensen, and Jan Wouters. Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation. *18th European Signal Processing Conference (EUSIPCO 2010)*, page 5, 2010.

[533] Chongjia Ni and Wenju Liu. Durational Characteristics and Pitch Characteristics of the Prosodic Phrase in Mandarin Chinese. In *2008 Chinese Conference on Pattern Recognition*, pages 1–5, Beijing, China, October 2008. IEEE.

[534] Jinfu Ni and Satoshi Nakamura. Use of Poisson Processes to Generate Fundamental Frequency Contours. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages IV–825–IV–828, Honolulu, HI, USA, 2007. IEEE.

[535] Jinfu Ni, Yoshinori Shiga, and Chiori Hori. Extraction of pitch register from expressive speech in Japanese. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4764–4768, South Brisbane, Queensland, Australia, April 2015. IEEE.

[536] Jinfu Ni, Yoshinori Shiga, Chiori Hori, and Yutaka Kidawara. A Targets-Based Superpositional Model of Fundamental Frequency Contours Applied to HMM-Based Speech Synthesis. *INTERSPEECH 2013*, page 5, 2013.

[537] J. K. Nielsen, M. G. Christensen, and S. H. Jensen. Default Bayesian Estimation of the Fundamental Frequency. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):598–610, March 2013.

[538] Jesper Kjaer Nielsen, Mads Graesboll Christensen, and Soren Holdt Jensen. An approximate Bayesian fundamental frequency estimator. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4617–4620, Kyoto, Japan, March 2012. IEEE.

[539] Jesper Kjcer Nielsen, Tobias Lindstrom Jensen, Jesper Rindom Jensen, Mads Graesboll Christensen, and Soren Holdt Jensen. A fast algorithm for maximum likelihood-based fundamental frequency estimation. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 589–593, Nice, August 2015. IEEE.

[540] Jesper Kjø er Nielsen, Tobias Lindstr, Jesper Rindom Jensen, Mads Grø esb, and others. Fast and statistically efficient fundamental frequency estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 86–90. IEEE, 2016.

[541] H. Niemann, J. Denzler, B. Kahles, R. Kompe, A. Kiessling, E. Noth, and V. Strom. Pitch determination considering laryngealization effects in spoken dialogs. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 7, pages 4457–4461, Orlando, FL, USA, 1994. IEEE.

[542] Tommy Nilsson, Stefan I Adalbjornsson, Naveed R Butt, and Andreas Jakobsson. Multi-pitch estimation of inharmonic signals. *21st European Signal Processing Conference (EUSIPCO 2013)*, page 5, 2013.

[543] S. S. Nimbhore, G. D. Ramteke, and R. J. Ramteke. Pitch estimation of Marathi spoken numbers in various speech signals. In *2013 International Conference on Communication and Signal Processing*, pages 405–409, Melmaruvathur, India, April 2013. IEEE.

[544] Anurag Nishad and Ram Bilas Pachori. Instantaneous fundamental frequency estimation of speech signals using tunable-Q wavelet transform. *2018 International Conference on Signal Processing and Communications (SPCOM)*, page 5, 2018.

[545] Anurag Nishad and Ram Bilas Pachori. Instantaneous fundamental frequency estimation of speech signals using tunable-Q wavelet transform. *2018 International Conference on Signal Processing and Communications (SPCOM)*, page 5, 2018.

[546] Sidsel Marie Norholm, Jesper Rindom Jensen, and Mads Graesboll Christensen. Instantaneous Fundamental Frequency Estimation With Optimal Segmentation for Nonstationary Voiced Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2354–2367, December 2016.

[547] Nutthacha Prukkanon, Kosin Chamnongthai, Yoshikazu Miyanaga, and Kohji Higuchi. VT-AMDF, a pitch detection algorithm. In *2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 453–456, Kanazawa, Japan, December 2009. IEEE.

[548] M.S. Obaidat, T. Lee, E. Zhang, G. Khalid, and D. Nelson. Wavelet algorithm for the estimation of pitch period of speech signal. In *Proceedings of Third International Conference on Electronics, Circuits, and Systems*, volume 1, pages 471–474, Rodos, Greece, 1996. IEEE.

[549] Keiko Ochi, Keikichi Hirose, and Nobuaki Minematsu. Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4257–4260, Taipei, Taiwan, April 2009. IEEE.

[550] A. Ogihara, S. Yamashita, and S. Yoneda. A switched capacitor pitch extraction circuit using the autocorrelation function and its application to spectral analysis. In *1991 International Conference on Circuits and Systems*, pages 216–219, Shenzhen, China, 1991. IEEE.

[551] Yasunori Ohishi, Hirokazu Kameoka, Kunio Kashino, and Kazuya Takeda. Parameter Estimation Method of F0 Control Model for Singing Voices. *INTERSPEECH 2008*, page 4, 2008.

[552] Yasunori Ohishi, Hirokazu Kameoka, Daichi Mochihashi, Hidehisa Nagano, and Kunio Kashino. Statistical Modeling of F0 Dynamics in Singing Voices Based on Gaussian Processes with Multiple Oscillation Bases. *INTERSPEECH 2010*, page 4, 2010.

[553] H. Ohmura. Fine pitch contour extraction by voice fundamental wave filtering method. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume ii, pages II/189–II/192, Adelaide, SA, Australia, 1994. IEEE.

[554] P. Ojala, P. Haavisto, A. Lakaniemi, and J. Vainio. A novel pitch-lag search method using adaptive weighting and median filtering. In *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No.99EX351)*, pages 114–116, Porvoo, Finland, 1999. IEEE.

[555] Antonio Origlia, Giovanni Abete, Francesco Cutugno, Iolanda Alfano, Renata Savy, and Bogdan Ludusan. A Divide et impera Algorithm for Optimal Pitch Stylization. *INTERSPEECH 2011*, page 4, 2011.

[556] Aziz Kubilay Ovacıklı, Patrik Paajarvi, James P LeBlanc, and Johan E Carlson. Uncovering harmonic content via skewness maximization - a Fourier analysis. *22nd European Signal Processing Conference (EUSIPCO 2014)*, page 5, 2014.

[557] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, and D.M. Wilkes. Analysis of fundamental frequency for near term suicidal risk assessment. In *SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions' (Cat. No.00CH37166)*, volume 3, pages 1853–1858, Nashville, TN, USA, 2000. IEEE.

[558] Monisankha Pal, Dipjyoti Paul, and Goutam Saha. Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language*, 48:31–50, March 2018.

[559] Shanthini Pandiaraj, Lineeta Gloria, Nisha Rachael Keziah, Synthia Vynothini, and K. R. Shankar Kumar. A proficient vocal training system with pitch detection using SHR. In *2011 3rd International Conference on Electronics Computer Technology*, pages 303–307, Kanyakumari, India, April 2011. IEEE.

[560] Vishala Pannala, G. Aneeja, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Robust Estimation of Fundamental Frequency Using Single Frequency Filtering Approach. In *INTERSPEECH 2016*, pages 2155–2159, September 2016.

[561] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. On the robustness of the Quasi-Harmonic model of speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4210–4213, Dallas, TX, USA, 2010. IEEE.

[562] Jayesh Pate. LOW COMPLEXITY VQ FOR MULTI-TAP PITCH PREDICTOR CODING. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, page 4, 1997.

[563] Tanvina B. Patel and Hemant A. Patil. Effectiveness of fundamental frequency (F0) and strength of excitation (SOE) for spoofed speech detection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5105–5109, Shanghai, March 2016. IEEE.

[564] Alexander Pavlovets and Alexander Petrovsky. Robust HNR-Based Closed-Loop Pitch and Harmonic Parameters Estimation. *INTERSPEECH 2011*, page 4, 2011.

[565] Alipah Pawi, Saeed Vaseghi, and Ben Milner. Pitch extraction using modified higher order moments. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5078–5081, Dallas, TX, USA, 2010. IEEE.

[566] Alipah Pawi, Saeed Vaseghi, Ben Milner, and Seyed Ghorshi. Fundamental Frequency Estimation Using Modified Higher Order Moments and Multiple Windows. *INTERSPEECH 2011*, page 4, 2011.

[567] Robert Peharz, Michael Wohlmayr, and Franz Pernkopf. Gain-robust multi-pitch tracking using sparse nonnegative matrix factorization. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5416–5419, Prague, Czech Republic, May 2011. IEEE.

[568] P.A. Pelle. A Robust Pitch Extraction System Based on Phase Locked Loops. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–249–I–252, Toulouse, France, 2006. IEEE.

[569] Patricia Pelle and Claudio Estienne. A robust pitch detector based on time envelope and individual harmonic information using Phase Locked Loops and consensual decisions. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1483–1487, Florence, Italy, May 2014. IEEE.

[570] Patricia A Pelle and Matias L Capeletto. Pitch Estimation Using Phase Locked Loops. *EUROSPEECH 2003*, page 4, 2003.

[571] Patricia A Pelle and Claudio F Estienne. A Pitch Extraction System Based on Phase Locked Loops and Consensus Decision. *INTERSPEECH 2007*, page 4, August 2007.

[572] Changping Peng and Wenjiu Liu. Pitch Prediction from MFCC Vectors Using Support Vector Regression. In *2007 International Conference on Natural Language Processing and Knowledge Engineering*, pages 342–346, Beijing, China, August 2007. IEEE.

[573] Antonio Pertusa and Jose M. Inesta. Multiple fundamental frequency estimation using Gaussian smoothness. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 105–108, Las Vegas, NV, USA, March 2008. IEEE.

[574] Rico Petrick, Masashi Unoki, Anish Mittal, Carlos Segura, and Ruediger Hoffmann. A Comprehensive Study on the Effects of Room Reverberation on Fundamental Frequency Estimation. *INTERSPEECH 2008*, page 4, September 2008.

[575] M. Petroni, A.S. Malowany, C.C. Johnston, and B.J. Stevens. A new, robust vocal fundamental frequency (F/sub 0/) determination method for the analysis of infant cries. In *Proceedings of IEEE Symposium on Computer-Based Medical Systems (CBMS)*, pages 223–228, Winston-Salem, NC, USA, 1994. IEEE Comput. Soc. Press.

[576] M. Petroni, A.S. Malowany, C.C. Johnston, and B.J. Stevens. A robust and accurate cross-correlation-based fundamental frequency (F/sub 0/) determination method for the improved analysis of infant cries. In *Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society*, volume 2, pages 975–976, Montreal, Que., Canada, 1995. IEEE.

[577] Siripong Potisuk. A system for extracting F0 contours of lexical tones using adaptive IIR notch filter with harmonic suppression. In *2016 International Conference on Asian Language Processing (IALP)*, pages 116–119, Tainan, Taiwan, November 2016. IEEE.

[578] RaviShankar Prasad and B Yegnanarayana. Robust Pitch Estimation in Noisy Speech Using ZTW and Group Delay Function. *INTERSPEECH 2015*, page 4, 2015.

[579] S R M Prasanna and D Govind. Unified pitch markers generation method for pitch and duration modification. In *2013 National Conference on Communications (NCC)*, pages 1–5, New Delhi, India, February 2013. IEEE.

[580] K Pratibha and H M Chandrashekar. Estimation and tracking of pitch for noisy speech signals using EMD based autocorrelation function algorithm. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2071–2075, Bangalore, May 2017. IEEE.

[581] R.E. Prieto and S. Kim. Time delay estimation and adaptive frame length iterations for noise robust pitch extraction. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–164–I–167, Hong Kong, China, 2003. IEEE.

[582] Nutthacha Prukkanon, Kosin Chamnongthai, and Yoshikazu Miyanaga. VT-AMDF pitch detection algorithm and Thai tone recognition system. In *The 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011*, pages 958–961, Khon Kaen, Thailand, May 2011. IEEE.

[583] Yao Qian, Frank K Soong, Miaomiao Wang, and Zhizheng Wu. A Minimum V/U Error Approach to F0 Generation in HMM-Based TTS. *INTERSPEECH 2009*, page 4, 2009.

[584] Qinghua Huang, Dongmei Wang, and Yunfeng Lu. Single channel speech enhancement based on prominent pitch estimation. In *IET International Communication Conference on Wireless Mobile & Computing (CCWMC 2009)*, pages 205–208, Shanghai, China, 2009. IET.

[585] Holger Quast, Olaf Schreiner, and Manfred R. Schroeder. Robust pitch tracking in the car environment. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–353–I–356, Orlando, FL, USA, May 2002. IEEE.

[586] Thomas F Quatieri. 2-D Processing of Speech with Application to Pitch Estimation. *7th International Conference on Spoken Language Processing (ICSLP2002)*, page 4, September 2002.

[587] Stanislaw A. Raczynski, Emmanuel Vincent, and Shigeki Sagayama. Dynamic Bayesian Networks for Symbolic Polyphonic Pitch Modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1830–1840, September 2013.

[588] Stanisław Raczynski, Nobutaka Ono, and Shigeki Sagayama. EXTENDING NONNEGATIVE MATRIX FACTORIZATION—A DISCUSSION IN THE CONTEXT OF MULTIPLE FREQUENCY ESTIMATION OF MUSICAL SIGNALS. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[589] M. H. Radfar, R. M. Dansereau, W.-Y. Chan, and W. Wong. MPtracker: A new multi-pitch detection and separation algorithm for mixed speech signals. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4468–4471, Prague, Czech Republic, May 2011. IEEE.

[590] M. H. Radfar, A. Sayadiyan, and R. M. Dansereau. A new algorithm for two-speaker pitch tracking in single channel paradigm. In *Signal Processing, 2006 8th International Conference on*, volume 1. IEEE, 2006.

[591] M Shahidur Rahman and Tetsuya Shimamura. Pitch characteristics of bone conducted speech. *18th European Signal Processing Conference (EUSIPCO 2010)*, page 5, 2010.

[592] M Shahidur Rahman and Tetsuya Shimamura. Pitch Determination Using Autocorrelation Function in Spectral Domain. *INTERSPEECH 2010*, page 4, 2010.

[593] M Shahidur Rahman, Hirobumi Tanaka, and Tetsuya Shimamura. Pitch Determination Using Aligned AMDF. *INTERSPEECH 2006*, page 4, 2006.

[594] Md. Saifur Rahman, Yosuke Sugiura, and Tetsuya Shimamura. Pitch Determination of Noisy Speech Using Cumulant Based Modified Weighted Function. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 1474–1478, Jeju, Korea (South), October 2018. IEEE.

[595] Md. Saifur Rahman, Yosuke Sugiura, and Tetsuya Shimamura. A Multiple Functions Multiplication Approach for Pitch Extraction of Noisy Speech. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6, Timisoara, Romania, October 2019. IEEE.

[596] Ch Srikanth Raj and T V Sreenivas. Joint Pitch-Analysis Formant-Synthesis framework for CS recovery of speech. *INTERSPEECH 2012*, page 4, 2012.

[597] Rajeev Rajan and Hema A. Murthy. Modified Group Delay Based MultiPitch Estimation in Co-Channel Speech. *arXiv:1603.05435 [cs]*, March 2016. arXiv: 1603.05435.

[598] Rohith Ramachandran and Philipos C. Loizou. Real-time pitch detection on the PDA for cochlear implant applications. In *2007 IEEE Dallas Engineering in Medicine and Biology Workshop*, pages 90–93, Dallas, TX, November 2007. IEEE.

[599] M V Achuth Rao and Prasanta Kumar Ghosh. Pitch prediction from Mel-frequency cepstral coefficients using sparse spectrum recovery. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–6, Chennai, India, March 2017. IEEE.

[600] M V Achuth Rao and Prasanta Kumar Ghosh. Pitch prediction from Mel-generalized cepstrum — a computationally efficient pitch modeling approach for speech synthesis. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1629–1633, Kos, Greece, August 2017. IEEE.

[601] Vishweshwara Rao, S Ramakrishnan, and Preeti Rao. Singing Voice Detection in Polyphonic Music using Predominant Pitch. *INTERSPEECH 2009*, page 4, 2009.

[602] Purshottam Singh Rathore and Ram Bilas Pachori. Instantaneous fundamental frequency estimation of speech signals using DESA in low-frequency region. In *2013 INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND COMMUNICATION (ICSC)*, pages 470–473, Noida, India, December 2013. IEEE.

[603] M. Kiran Reddy and K. Sreenivasa Rao. Robust Pitch Extraction Method for the HMM-Based Speech Synthesis System. *IEEE Signal Processing Letters*, 24(8):1133–1137, August 2017.

[604] Barbara Resch, Mattias Nilsson, Anders Ekman, and W. Bastiaan Kleijn. Estimation of the Instantaneous Pitch of Speech. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):813–822, March 2007.

[605] Sergio Roa, Maren Bennewitz, and Sven Behnke. Fundamental Frequency Estimation Based on Pitch-Scaled Harmonic Filtering. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages IV–397–IV–400, Honolulu, HI, April 2007. IEEE.

[606] Laura Romoli, Stefania Cecchi, Paolo Peretti, and Francesco Piazza. A Mixed Decorrelation Approach for Stereo Acoustic Echo Cancellation Based on the Estimation of the Fundamental Frequency. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):690–698, February 2012.

[607] Laura Romoli, Stefania Cecchi, and Francesco Piazza. A voice activity detection algorithm for multichannel acoustic echo cancellation exploiting fundamental frequency estimation. In *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 244–249, Zagreb, Croatia, September 2015. IEEE.

[608] Andrew Rosenberg and Julia Hirschberg. Detecting Pitch Accent Using Pitch-Corrected Energy-Based Predictors. *INTERSPEECH 2007*, page 4, August 2007.

[609] Julie Rosier and Yves Grenier. Two-pitch estimation for co-channel speakers separation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages IV–4160–IV–4160, Orlando, FL, USA, May 2002. IEEE.

[610] Stephane Rossignol and Jean-Luc Collette. A method for simultaneously extract the fundamental frequency of a speech signal and segment it. In *2008 IEEE International Symposium on Signal Processing and Information Technology*, pages 213–218, Sarajevo, Bosnia and Herzegovina, December 2008. IEEE.

[611] A. B. Rostron and C. P. Welbourn. A computer assisted system for the extraction and visual display of pitch. *Behavior Research Methods & Instrumentation*, 8(5):456–459, September 1976.

[612] Sujan Kumar Roy, Md. Kamrul Hasan, Keikichi Hirose, and Md. Khademul Islam Molla. Dominant harmonic modification with data adaptive filter based algorithm for robust pitch estimation. In *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pages 2417–2420, Rio de Janeiro, Brazil, May 2011. IEEE.

[613] Sujan Kumar Roy, Md. Khademul Islam Molla, Keikichi Hirose, and Md. Kamrul Hasan. Pitch estimation of noisy speech signals using EMD-fourier based hybrid algorithm. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 2658–2661, Paris, France, May 2010. IEEE.

[614] Sujan Kumar Roy and Wei-Ping Zhu. Pitch estimation of noisy speech using ensemble empirical mode decomposition and dominant harmonic modification. In *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4, Toronto, ON, Canada, May 2014. IEEE.

[615] Li Ru-Wei, Cao Long-Tao, and Li Yang. Pitch Detection Method for Noisy Speech Signals Based on Wavelet Transform and Autocorrelation Function. In *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 153–156, Beijing, China, October 2013. IEEE.

[616] Krzysztof Rychlicki-Kicior and Bartłomiej Stasiak. Metaheuristic Optimization of Multiple Fundamental Frequency Estimation. In Dr. Aleksandra Gruca, Tadeusz Czachórski, and Stanisław Kozielski, editors, *Man-Machine Interactions 3*, volume 242, pages 307–314. Springer International Publishing, Cham, 2014.

[617] B. Tarun Sai, Ishwar Chandra Yadav, S. Shahnawazuddin, and Gayadhar Pradhan. Enhancing Pitch Robustness of Speech Recognition System through Spectral Smoothing. In *2018 International Conference on Signal Processing and Communications (SPCOM)*, pages 242–246, Bangalore, India, July 2018. IEEE.

[618] S. Sakai and J. Glass. Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 712–717, St Thomas, VI, USA, 2003. IEEE.

[619] Masaharu Sakamoto and Takashi Saitoh. An Automatic Pitch-Marking Method using Wavelet Transform. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, page 4, October 2000.

[620] E. Sakk, J.C. Belina, J.M. Kuzma, and C. Heegard. A time-frequency model for pitch analysis. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380)*, pages 149–152, Pittsburgh, PA, USA, 1998. IEEE.

[621] O. Salor, M. Demirekler, and U. Orguner. An Efficient Algorithm for Pitch Determination of Speech Signals - Kalman Filter Approach. In *2006 IEEE 14th Signal Processing and Communications Applications*, pages 1–4, Antalya, Turkey, 2006. IEEE.

[622] Sam Kwong, Wei Gang, and C.H. Lee. A pitch detection algorithm based on time-frequency analysis. In *[Proceedings] Singapore ICCS/ISITA '92*, pages 432–436, Singapore, 1990. IEEE.

[623] S.A. Samad, A. Hussain, and Low Kok Fah. Pitch detection of speech signals using the cross-correlation technique. In *2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No.00CH37119)*, volume 1, pages 283–286, Kuala Lumpur, Malaysia, 2000. IEEE.

[624] Gerard Sanchez, Hanna Silen, Jani Nurminen, and Moncef Gabbouj. Hierarchical Modeling of F0 Contours for Voice Conversion. *INTERSPEECH 2014*, page 4, 2014.

[625] Naoto Sasaoka, Toshihiro Miyawaki, Yoshio Itoh, Kensaku Fujii, and Yutaka Fukui. A Study on Pitch Period Detection Based on ALE for Sinusoidal Noise. In *2006 International Symposium on Communications and Information Technologies*, pages 114–117, Bangkok, Thailand, October 2006. IEEE.

[626] A. V. Savchenko and V. V. Savchenko. A Method for Measuring the Pitch Frequency of Speech Signals for the Systems of Acoustic Speech Analysis. *Measurement Techniques*, 62(3):282–288, June 2019.

[627] Gaston Schlotthauer and Maria Eugenia Torres. A NEW ALGORITHM FOR INSTANTANEOUS Fo SPEECH EXTR ACTION BASED ON ENSEMBLE EMPIRIC AL MODE DECOMPOSITION. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[628] Jilt Sebastian, P A Manoj Kumar, and Hema A Murthy. Pitch estimation from speech using Grating Compression Transform on Modified Group-Delay-gram. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, Mumbai, India, February 2015. IEEE.

[629] Erol Seke and Kemal Özkan. A new speech signal denoising algorithm using common vector approach. *International Journal of Speech Technology*, 21(3):659–670, September 2018.

[630] S Chandra Sekhar and Sridhar PiZZi. Novel auditory motivated subband temporal envelope based fundamental frequency estimation algorithm. *14th European Signal Processing Conference (EUSIPCO 2006)*, page 5, 2006.

[631] Gregory Sell. Automatic carrier pitch estimation for coherent demodulation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2119–2123, Florence, Italy, May 2014. IEEE.

[632] P. T. Selvan and V. Vaishnavi. Singing pitch extraction and voice separation from music accompaniment using trend estimation and tandem algorithm. In *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, pages 232–235, Tirunelveli, March 2013. IEEE.

[633] V Sercov and A Petrovsky. The method of pitch frequency detection on the base of tuning to its harmonics. *9th European Signal Processing Conference (EUSIPCO 1998)*, page 4, 1998.

[634] Guruprasad Seshadri and B. Yegnanarayana. Performance of an Event-Based Instantaneous Fundamental Frequency Estimator for Distant Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1853–1864, September 2011.

[635] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Pitch Contour Parameterisation Based on Linear Stylisation for Emotion Recognition. *INTERSPEECH 2009*, page 4, 2009.

[636] A. Shah, R.P. Ramachandran, and M.A. Lewis. Robust pitch estimation using an event based adaptive Gaussian derivative filter. In *2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353)*, pages II–843–II–846, Phoenix-Scottsdale, AZ, USA, 2002. IEEE.

[637] Nirmesh J. Shah, Pramod B. Bachhav, and Hemant A. Patil. A novel filtering-based F0 estimation algorithm with an application to voice conversion. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1528–1531, Kuala Lumpur, December 2017. IEEE.

[638] S Shahnawazuddin, Deepak K. T., Gayadhar Pradhan, and Rohit Sinha. Enhancing noise and pitch robustness of children's ASR. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5225–5229, New Orleans, LA, March 2017. IEEE.

[639] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. An Approach for Pitch Estimation from Noisy Speech. In *2007 Canadian Conference on Electrical and Computer Engineering*, pages 1590–1593, Vancouver, BC, Canada, 2007. IEEE.

[640] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad. A Pitch Detection Method for Speech Signals with Low Signal-to-Noise Ratio. In *2007 International Symposium on Signals, Systems and Electronics*, pages 399–402, Montreal, QC, Canada, July 2007. IEEE.

[641] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad. A robust pitch detection algorithm for speech signals in a practical noisy environment. In *2007 50th Midwest Symposium on Circuits and Systems*, pages 385–388, Montreal, QC, Canada, August 2007. IEEE.

[642] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A Robust Pitch Estimation Algorithm in Noise. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages IV–1073–IV–1076, Honolulu, HI, April 2007. IEEE.

[643] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A method for pitch estimation from noisy speech signals based on a pitch-harmonic extraction. In *2008 International Conference on Neural Networks and Signal Processing*, pages 120–123, Nanjing, China, June 2008. IEEE.

[644] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. On extracting pitch from noisy speech signals based on spectral and temporal enhancement. In *2008 Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference*, pages 77–80, Montreal, QC, Canada, June 2008. IEEE.

[645] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. On the estimation of pitch of noisy speech based on time and frequency domain representations. In *2008 Canadian Conference on Electrical and Computer Engineering*, pages 001819–001822, Niagara Falls, ON, Canada, May 2008. IEEE.

[646] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A pitch extraction algorithm in noise based on temporal and spectral representations. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4477–4480, Las Vegas, NV, USA, March 2008. IEEE.

[647] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A temporal matching method for pitch determination from noisy speech signals. In *2008 51st Midwest Symposium on Circuits and Systems*, pages 938–941, Knoxville, TN, USA, August 2008. IEEE.

[648] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A spectral matching method for pitch estimation from noise-corrupted speech. In *2009 IEEE International Symposium on Circuits and Systems*, pages 1413–1416, Taipei, Taiwan, May 2009. IEEE.

[649] C. Shahnaz, W.-P. Zhu, and M.O. Ahmad. A Robust Pitch Estimation Approach for Colored Noise-Corrupted Speech. In *2005 IEEE International Symposium on Circuits and Systems*, pages 3143–3146, Kobe, Japan, 2005. IEEE.

[650] C. Shahnaz, W.-P. Zhu, and M.O. Ahmad. Robust Pitch Estimation At Very Low SNR Exploiting Time and Frequency Domain Cues. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 389–392, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[651] C. Shahnaz, W.-P. Zhu, and M.O. Ahmad. A spectro-temporal algorithm for pitch frequency estimation from noisy observations. In *2008 IEEE International Symposium on Circuits and Systems*, pages 1704–1707, Seattle, WA, USA, May 2008. IEEE.

[652] Celia Shahnaz and A Fattah. A time-frequency domain approach for pitch-peak picking of noisy speech. *2004 12th European Signal Processing Conference*, page 4, September 2004.

[653] Celia Shahnaz, Wei-Ping Zhu, and M. Omair Ahmad. Pitch Estimation Based on a Harmonic Sinusoidal Autocorrelation Model and a Time-Domain Matching Scheme. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):322–335, January 2012.

[654] Wei Shan and Keiichi Funaki. F0 estimation of speech based on IRAPT using WLP-based TV-CAR analysis. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–4, Tianjin, China, October 2016. IEEE.

[655] Xu Shao and Ben Milner. MAP Prediction of Pitch from MFCC Vectors for Speech Reconstruction. *INTERSPEECH 2004*, page 4, 2004.

[656] Xu Shao, Ben P Milner, and Stephen J Cox. Integrated Pitch and MFCC Extraction for Speech Reconstruction and Speech Recognition Applications. *EUROSPEECH 2003*, page 4, 2003.

[657] Yiwen Shao and Qiguang Lin. Use of Pitch Continuity for Robust Speech Activity Detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5534–5538, Calgary, AB, April 2018. IEEE.

[658] Dushyant Sharma and Patrick A Naylor. Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[659] D.W.N. Sharp and R.L. While. Determining the pitch period of speech using no multiplications. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 527–529 vol.2, Minneapolis, MN, USA, 1993. IEEE.

[660] Shasha Xing, Haishan Han, Kai Liu, and Junli Chen. A robust pitch tracking method in noisy environment. In *IET International Conference on Smart and Sustainable City 2013 (ICSSC 2013)*, pages 145–149, Shanghai, China, 2013. Institution of Engineering and Technology.

[661] S.A. Shedied, M.E. Gadalah, and H.F. VanLundingham. Pitch estimator for noisy speech signals. In *SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions' (Cat. No.00CH37166)*, volume 1, pages 97–100, Nashville, TN, USA, 2000. IEEE.

[662] G.A. Shelby, C.M. Cooper, and R.R. Adhami. A wavelet-based speech pitch detector for tone languages. In *Proceedings of IEEE-SP International Symposium on Time- Frequency and Time-Scale Analysis*, pages 596–599, Philadelphia, PA, USA, 1994. IEEE.

[663] Liming Shi, Jesper K. Nielsen, Jesper R. Jensen, Max A. Little, and Mads G. Christensen. A Kalman-based fundamental frequency estimation algorithm. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 314–318, New Paltz, NY, October 2017. IEEE.

[664] Liming Shi, Jesper Kjaer Nielsen, Jesper Rindom Jensen, Max A. Little, and Mads Graesboll Christensen. Robust Bayesian Pitch Tracking Based on the Harmonic Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1737–1751, November 2019.

[665] Liming Shi, Jesper Kjar Nielsen, Jesper Rindom Jensen, and Mads Grosboll Christensen. A variational EM method for pole-zero modeling of speech with mixed block sparse and Gaussian excitation. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1784–1788, Kos, Greece, August 2017. IEEE.

[666] Shi-Huang Chen and Jhing-Fa Wang. Extraction of pitch information in noisy speech using wavelet transform with aliasing compensation. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 89–92, Salt Lake City, UT, USA, 2001. IEEE.

[667] T. Shimamura and H. Kobayashi. Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Transactions on Speech and Audio Processing*, 9(7):727–730, October 2001.

[668] T. Shimamura and H. Takagi. Noise-robust fundamental frequency extraction method based on exponentiated band-limited amplitude spectrum. In *The 2004 47th Midwest Symposium on Circuits and Systems, 2004. MWSCAS '04.*, volume 2, pages II_141–II_144, Hiroshima, Japan, 2004. IEEE.

[669] Hiroshi Shimodaira and Mitsuru Nakai. Robust Pitch Detection by Narrow Band Spectrum Analysis. In *Second International Conference on Spoken Language Processing (ICSLP'92)*, Banff, Alberta, Canada, October 1992.

[670] Wu Shuhong and Zhang Gang. 8kbit/s LD-aCELP Speech Coding with Backward Pitch Detection. In *2009 Asia-Pacific Conference on Information Processing*, pages 434–437, Shenzhen, China, July 2009. IEEE.

[671] Berrak Ozturk Simsek and Aydin Akan. Frequency estimation for monophonical music by using a modified VMD method. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1873–1876, Kos, Greece, August 2017. IEEE.

[672] Chetan Pratap Singh and T Kishore Kumar. Efficient pitch detection algorithms for pitched musical instrument sounds: A comparative performance evaluation. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1876–1880, Delhi, India, September 2014. IEEE.

[673] M. Slaney and R.F. Lyon. A perceptual pitch detector. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 357–360, Albuquerque, NM, USA, 1990. IEEE.

[674] Christine Smit and Daniel P.W. Ellis. Guided harmonic sinusoid estimation in a multi-pitch environment. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 41–44, New Paltz, NY, USA, October 2009. IEEE.

[675] H.C. So and K.W. Chan. Reformulation of Pisarenko Harmonic Decomposition Method for Single-Tone Frequency Estimation. *IEEE Transactions on Signal Processing*, 52(4):1128–1135, April 2004.

[676] YongJin So, Jia Jia, and LianHong Cai. Analysis and Improvement of Auto-correlation Pitch Extraction Algorithm Based on Candidate Set. In Zhihong Qian, Lei Cao, Weilian Su, Tingkai Wang, and Huamin Yang, editors, *Recent Advances in Computer Science and Information Engineering*, volume 128, pages 697–702. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[677] J. Sola-Soler, R. Jane, J.A. Fiz, and J. Morera. Towards automatic pitch detection in snoring signals. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, volume 4, pages 2974–2976, Chicago, IL, USA, 2000. IEEE.

[678] Bing Song, Chuan-qing Gu, and Jian-jun Zhang. A new pitch detection algorithm based on wavelet transform. *Journal of Shanghai University (English Edition)*, 9(4):309–313, August 2005.

[679] Charlotte Sorensen, Angeliki Xenaki, Jesper B. Boldt, and Mads G. Christensen. Pitch-based non-intrusive objective intelligibility prediction. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 386–390, New Orleans, LA, March 2017. IEEE.

[680] N. Sripriya and T. Nagarajan. Pitch estimation using harmonic product spectrum derived from DCT. In *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*, pages 1–4, Xi'an, China, October 2013. IEEE.

[681] Johannes Stahl and Pejman Mowlaee. A Pitch-Synchronous Simultaneous Detection-Estimation Framework for Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):436–450, February 2018.

[682] Miroslav Stanek and Tomas Smatana. Comparison of fundamental frequency detection methods and introducing simple self-repairing algorithm for musical applications. In *2015 25th International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 217–221, Pardubice, Czech Republic, April 2015. IEEE.

[683] Michael Staudacher, Viktor Steixner, Andreas Griessner, and Clemens Zierhofer. Fast fundamental frequency determination via adaptive autocorrelation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1), December 2016.

[684] Simon Stone, Peter Steiner, and Peter Birkholz. A Time-Warping Pitch Tracking Algorithm Considering Fast f0 Changes. In *Interspeech 2017*, pages 419–423. ISCA, August 2017.

[685] Fabian-Robert Stoter, Nils Werner, Stefan Bayer, and Bernd Edler. Refining fundamental frequency estimates using time warping. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 6–10, Nice, August 2015. IEEE.

[686] Sofia Strömbergsson. Today's Most Frequently Used F0 Estimation Methods, and Their Accuracy in Estimating Male and Female Pitch in Clean Speech. In *INTERSPEECH 2016*, pages 525–529, September 2016.

[687] Hong Su, Hui Zhang, Xueliang Zhang, and Guanglai Gao. Convolutional neural network for robust pitch determination. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 579–583, Shanghai, March 2016. IEEE.

[688] Li Su, Tsung-Ying Chuang, and Yi-Hsuan Yang. EXPLOITING FREQUENCY, PERIODICITY AND HARMONICITY USING ADVANCED TIME-FREQUENCY CONCENTRATION TECHNIQUES FOR MULTIPITCH ESTIMATION OF CHOIR AND SYMPHONY. *17th International Society for Music Information Retrieval Conference*, page 7, 2016.

[689] S. A. Suma and K. S. Gurumurthy. Novel pitch extraction methods using average magnitude difference function (AMDF) for LPC speech coders in noisy environments. In *2010 2nd International Conference on Signal Processing Systems*, pages V1–636–V1–640, Dalian, China, July 2010. IEEE.

[690] Jia Sun, Jilun Lu, Aijun Li, and Yuan Jia. The Pitch Analysis of Imperative Sentences in Standard Chinese. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4, Kunming, China, December 2008. IEEE.

[691] Jiadong Sun, Elias Aboutanios, and David B Smith. Iterative Weighted Least Squares Frequency Estimation for Harmonic Sinusoidal Signal in Power Systems. *2018 26th European Signal Processing Conference (EUSIPCO)*, page 5, 2018.

[692] Jiadong Sun, Shanglin Ye, and Elias Aboutanios. Robust and rapid estimation of the parameters of harmonic signals in three phase power systems. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 408–412, Budapest, Hungary, August 2016. IEEE.

[693] Xuejing Sun. A PITCH DETERMINATION ALGORITHM BASED ON SUBHARMONIC- TO-HARMONIC RATIO. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, page 4, October 2000.

[694] Xuejing Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–333. IEEE, 2002.

[695] Xuejing Sun and Sameer Gadre. Efficient Three-Stage Pitch Estimation for Packet Loss Concealment. *INTERSPEECH 2010*, page 4, 2010.

[696] Johan Sward, Hongbin Li, and Andreas Jakobsson. Off-Grid Fundamental Frequency Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):296–303, February 2018.

[697] Tan Tian Swee, Sheikh Hussain Shaikh Salleh, and Mohd Redzuan Jamaludin. Speech pitch detection using short-time energy. In *International Conference on Computer and Communication Engineering (ICCCE'10)*, pages 1–6, Kuala Lumpur, Malaysia, May 2010. IEEE.

[698] Marek Szczerba and Andrzej Czyzewski. Pitch Detection Enhancement Employing Music Prediction. *Journal of Intelligent Information Systems*, 24(2-3):223–251, March 2005.

[699] Istvan Szekrenyes. ProsoTool, a method for automatic annotation of fundamental frequency. In *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 291–296, Gyor, Hungary, October 2015. IEEE.

[700] Y. Tabata and T. Shimamura. Noise robust pitch extraction based on auto-correlation analysis in the frequency domain. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*, pages 193–196, Hong Kong, China, 2001. IEEE.

[701] J. Tabrikian, S. Dubnov, and Y. Dickalov. Maximum A-Posteriori Probability Pitch Tracking in Noisy Environments Using Harmonic Model. *IEEE Transactions on Speech and Audio Processing*, 12(1):76–87, January 2004.

[702] Joseph Tabrikian, Shlomo Dubnov, and Yulya Dickalov. Speech enhancement by harmonic modeling via map pitch tracking. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–549–I–552, Orlando, FL, USA, May 2002. IEEE.

[703] Hiroshi Takagi and Tetsuya Shimamura. Extraction of fundamental frequency of speech based on exponentiated band-limited spectrum. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages IV–4165–IV–4165, Orlando, FL, USA, May 2002. IEEE.

[704] Kou Tanaka, Hirokazu Kameoka, and Kazuho Morikawa. Vae-Space: Deep Generative Model of Voice Fundamental Frequency Contours. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5779–5783, Calgary, AB, April 2018. IEEE.

[705] Tomohiro Tanaka, Takao Kobayashi, Dhany Arifianto, and Takashi Masuko. Fundamental frequency estimation based on instantaneous frequency amplitude spectrum. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–329–I–332, Orlando, FL, USA, May 2002. IEEE.

[706] S. Tanigawa, T. Kikuchi, T. Yamaoka, and N. Hamada. The estimation of fundamental frequency of speech using microphone array. In *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers (Cat. No.CH37020)*, volume 2, pages 1115–1119, Pacific Grove, CA, USA, 1999. IEEE.

[707] John H Taylor and Ben Milner. Modelling and Estimation of the Fundamental Frequency of Speech Using a Hidden Markov Model. *INTERSPEECH 2013*, page 5, 2013.

[708] Horia-Nicolai Teodorescu. Using local variance, Allan- and Hadamard variances in speech analysis – Pitch analysis. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4, Iasi, Romania, July 2019. IEEE.

[709] Harvey D. Thornburg and Randal J. Leistikow. A New Probabilistic Spectral Pitch Estimator: Exact and MCMC-approximate Strategies. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, and Uffe Kock Wiil, editors, *Computer Music Modeling and Retrieval*, volume 3310, pages 41–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[710] Balint Pal Toth and Tamas Gabor Csapo. Continuous fundamental frequency prediction with deep neural networks. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1348–1352, Budapest, Hungary, August 2016. IEEE.

[711] Ivo Trajkovic, Christoph Reller, Martin Wolf, and Hans-Andrea Loeliger. Modelling and filtering almost periodic signals by time-varying Fourier series with application to near-infrared spectroscopy. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[712] T.H. Tran, Q.P. Ha, and G. Dissanayake. New wavelet-based,pitch detection method for human-robot voice interface. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 1, pages 527–532, Sendai, Japan, 2004. IEEE.

[713] Georgina Tryfou and Maurizio Omologo. A reassigned based singing voice pitch contour extraction method. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–325, New Orleans, LA, March 2017. IEEE.

[714] EnShuo Tsau, Namgook Cho, and C.-C. Jay Kuo. Fundamental frequency estimation for music signals with modified Hilbert-Huang transform (HHT). In *2009 IEEE International Conference on Multimedia and Expo*, pages 338–341, New York, NY, USA, June 2009. IEEE.

[715] Pirros Tsiakoulis, Alexandros Potamianos, and Dimitrios Dimitriadis. Instantaneous frequency and bandwidth estimation using filterbank arrays. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8032–8036, Vancouver, BC, Canada, May 2013. IEEE.

[716] Martin Turi Nagy, Gregor Rozinaj, and Andrej Palenik. A hybrid pitch period estimation method based on HNM model. In *ELMAR 2007*, pages 175–178, Zadar, Croatia, September 2007. IEEE.

[717] Savitha S Upadhya. Pitch detection in time and frequency domain. In *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, pages 1–5, Mumbai, India, October 2012. IEEE.

[718] Abhay Upadhyay and Ram Bilas Pachori. A new method for determination of instantaneous pitch frequency from speech signals. In *2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, pages 325–330, Salt Lake City, UT, USA, August 2015. IEEE.

[719] Ewout van den Berg and Bhuvana Ramabhadran. Dictionary-Based Pitch Tracking with Dynamic Programming. *INTERSPEECH 2014*, page 5, 2014.

[720] Daniel R van Niekerk and Etienne Barnard. Generating Fundamental Frequency Contours for Speech Synthesis in Yoruba. *INTERSPEECH 2013*, page 5, 2013.

[721] Mark VanDam and Paul De Palma. Fundamental frequency of child-directed speech using automatic speech recognition. In *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*, pages 1349–1353, Kita-Kyushu, Japan, December 2014. IEEE.

[722] M. R. Varley and R. J. Simpson. A secondary feature algorithm for pitch determination of speech signals and its implementation in real-time using a Motorola DSP56001 based system. In *IEE Colloquium on Practical Applications of DSP Devices*, pages 6/1–6/5, June 1990.

[723] Ekrem Varoglu and Kadri Hacioglu. NONLINEAR FORMANT-PITCH PREDICTION USING RECURRENT NEURAL NETWORKS. *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, page 4, 1996.

[724] R N J Veldhuis. Consistent Pitch Marking. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, page 4, 2000.

[725] Laura Verde, Giuseppe De Pietro, and Giovanna Sannino. A methodology for voice classification based on the personalized fundamental frequency estimation. *Biomedical Signal Processing and Control*, 42:134–144, April 2018.

[726] Prateek Verma and Ronald W. Schafer. Frequency Estimation from Waveforms Using Multi-Layered Neural Networks. In *INTERSPEECH 2016*, pages 2165–2169, September 2016.

[727] Ekaterina Verteletskaya, Kirill Sakhnov, and Boris Simak. Pitch Detection Algorithms and Voiced/Unvoiced Classification for Noisy Speech. In *2009 16th International Conference on Systems, Signals and Image Processing*, pages 1–5, Chalkida, Greece, June 2009. IEEE.

[728] E. Vincent, N. Bertin, and R. Badeau. Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, March 2010.

[729] Srikanth Vishnubhotla and Carol Y Espy-Wilson. An Algorithm for Multi-Pitch Tracking in Co-Channel Speech. *INTERSPEECH 2008*, page 4, September 2008.

[730] Vladan Vuckovic. Some advances in fundamental frequency analyzing by digital speech processing application ADS v2.1. In *2009 9th International Conference on Telecommunication in Modern Satellite, Cable, and Broadcasting Services*, pages 475–478, Nis, October 2009. IEEE.

[731] György Várallyay. SSM – A Novel Method to Recognize the Fundamental Frequency in Voice Signals. In Luis Rueda, Domingo Mery, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4756, pages 88–95. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[732] Yukoh Wakabayashi, Takahiro Fukumori, Masato Nakayama, Takanobu Nishiura, and Yoichi Yamashita. Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5560–5564, New Orleans, LA, March 2017. IEEE.

[733] Wan-yi Lin and Lin-shan Lee. Improved tone recognition for fluent Mandarin speech based on new inter-syllabic features and robust pitch extraction. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 237–242, St Thomas, VI, USA, 2003. IEEE.

[734] Cheng-Cheng Wang, Zhen-Hua Ling, and Li-Rong Dai. Asynchronous F0 and Spectrum Modeling for HMM-Based Speech Synthesis. *INTERSPEECH 2009*, page 4, 2009.

[735] Dongmei Wang and John H. L. Hansen. F0 estimation for noisy speech by exploring temporal harmonic structures in local time frequency spectrum segment. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6510–6514, Shanghai, March 2016. IEEE.

[736] Dongmei Wang, John H. L. Hansen, and Emily Tobey. F0 estimation for noisy speech based on exploring local time-frequency segment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, New Paltz, NY, USA, October 2015. IEEE.

[737] Dongmei Wang and Philipos C Loizou. Pitch Estimation Based on Long Frame Harmonic Model and Short Frame Average Correlation Coefficient. *INTERSPEECH 2012*, page 4, 2012.

[738] Dongmei Wang, Philipos C Loizou, and John H L Hansen. F0 Estimation in Noisy Speech Based on Long-Term Harmonic Feature Analysis Combined with Neural Network Classification. *INTERSPEECH 2014*, page 5, 2014.

[739] Dongmei Wang, Chengzhu Yu, and John H. L. Hansen. Robust Harmonic Features for Classification-Based Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):952–964, May 2017.

[740] Dusheng Wang, Lizhong Li, and Jiankang Zhang. A Practical Look-back and Look-forth Pitch Tracking and Smoothing Algorithm. In *2006 1ST IEEE Conference on Industrial Electronics and Applications*, pages 1–4, Singapore, May 2006. IEEE.

[741] Miaomiao Wang, Miaomiao Wen, Keikichi Hirose, and Nobuaki Minematsu. Improved Generation of Fundemental Frequency in HMM-Based Speech Synthesis Using Generation Process Model. *INTERSPEECH 2010*, page 4, 2010.

[742] Qiao Wang, Xiaoqun Zhao, and Jingyun Xu. Pitch detection algorithm based on normalized correlation function and central bias function. In *2015 10th International Conference on Communications and Networking in China (ChinaCom)*, pages 617–620, Shanghai, China, August 2015. IEEE.

[743] Tianyu T. Wang and Thomas F. Quatieri. Exploiting temporal change of pitch in formant estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3929–3932, Las Vegas, NV, USA, March 2008. IEEE.

[744] Tianyu T Wang and Thomas F Quatieri. 2-D Processing of Speech for Multi-Pitch Analysis. *INTERSPEECH 2009*, page 4, 2009.

[745] Tianyu T Wang and Thomas F Quatieri. Multi-Pitch Estimation by a Joint 2-D Representation of Pitch and Pitch Dynamics. *INTERSPEECH 2010*, page 4, 2010.

[746] T.T. Wang and T.F. Quatieri. High-Pitch Formant Estimation by Exploiting Temporal Change of Pitch. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):171–186, January 2010.

[747] Yih-Ru Wang, I-Je Wong, and Teng-Chun Tsao. A statistical pitch detection algorithm. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–357–I–360, Orlando, FL, USA, May 2002. IEEE.

[748] Wang Hong and Jin'Gui Pan. An improved SIFT method for pitch estimation of speech. In *The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding*, pages 298–302, Chengdu, China, December 2010. IEEE.

[749] M. Wasserblat, M. Gainza, D. Dorran, and Y. Domb. Pitch tracking and voiced/unvoiced detection in noisy environment using optimal sequence estimation. In *IET Irish Signals and Systems Conference (ISSC 2008)*, pages 43–48, Galway, Ireland, 2008. IEE.

[750] Xiaopeng Wei, Lasheng Zhao, Qiang Zhang, and Jing Dong. A Pitch Estimation Algorithm Based on the Variance Analysis. In *2008 3rd International Conference on Innovative Computing Information and Control*, pages 455–455, Dalian, Liaoning, China, 2008. IEEE.

[751] Wei-Chen Chang and A.W.Y. Su. A novel recurrent network based pitch detection technique for quasi-periodic/pitch-varying signals. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, pages 816–821, Honolulu, HI, USA, 2002. IEEE.

[752] Wei Huang, Jianshu Chao, and Yaxin Zhang. Combination of pitch and MFCC GMM supervectors for speaker verification. In *2008 International Conference on Audio, Language and Image Processing*, pages 1335–1339, Shanghai, China, July 2008. IEEE.

[753] Wei Wei and L. Kilmartin. A pitch analysis technique for automated speech distortion identification in VoIP networks. In *Proceedings of 2002 IEEE 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop.*, pages 47–52, Pine Mountain, GA, USA, 2002. IEEE.

[754] Weihua Zhang, Hyun-Soo Kim, and W.H. Holmes. Investigation of the spectral envelope estimation vocoder and improved pitch estimation based on the sinusoidal speech model. In *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No.97TH8237)*, volume 1, pages 513–516, Singapore, 1997. IEEE.

[755] Wen-Hsing Lai. Pitch modeling for Chinese Speech mixed with English word spelling. In *2008 9th International Conference on Signal Processing*, pages 592–595, Beijing, China, October 2008. IEEE.

[756] C. Wendt and A.P. Petropulu. Pitch determination and speech segmentation using the discrete wavelet transform. In *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96*, volume 2, pages 45–48, Atlanta, GA, USA, 1996. IEEE.

[757] Wenhui Jia and Wai-Yip Chan. Joint pitch and voicing estimation for multiband excitation and sinusoidal speech coders. In *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.*, volume 1, pages 210–213, Pacific Grove, CA, USA, 2002. IEEE.

[758] R.K. Whitman and D.M. Etter. An investigation of estimating pitch periods using a non-linear differential operator. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1237–1241, Pacific Grove, CA, USA, 1994. IEEE Comput. Soc. Press.

[759] Michael Wohlmayr and Marian Kepesi. Joint Position-Pitch Extraction from Multichannel Audio. *INTERSPEECH 2007*, page 4, 2007.

[760] Michael Wohlmayr, Robert Peharz, and Franz Pernkopf. Efficient implementation of probabilistic multi-pitch tracking. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5412–5415, Prague, Czech Republic, May 2011. IEEE.

[761] Michael Wohlmayr and Franz Pernkopf. Multipitch Tracking Using a Factorial Hidden Markov Model. *INTERSPEECH 2008*, page 4, 2008.

[762] Michael Wohlmayr and Franz Pernkopf. Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markov Model. *INTERSPEECH 2009*, page 4, 2009.

[763] Michael Wohlmayr and Franz Pernkopf. EM-Based Gain Adaptation for Probabilistic Multipitch Tracking. *INTERSPEECH 2011*, page 4, 2011.

[764] Michael Wohlmayr and Franz Pernkopf. Model-Based Multiple Pitch Tracking Using Factorial HMMs: Model Adaptation and Inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1742–1754, August 2013.

[765] Hongwei Wu, Yibiao Yu, Heming Zhao, Xueqin Chen, and Chunjuan Wang. Pitch estimation using mean shift algorithm on multitaper spectrum of noisy speech. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 441–444, Shenzhen, China, April 2014. IEEE.

[766] M. Wu, D. Wang, and G. J. Brown. A multi-pitch tracking algorithm for noisy speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–369–I–372, May 2002.

[767] Yi-Jian Wu and Frank Soong. Modeling pitch trajectory by hierarchical HMM with minimum generation error training. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4017–4020, Kyoto, Japan, March 2012. IEEE.

[768] Yuntao Wu, Amir Leshem, Jesper Rindom Jensen, and Guisheng Liao. Joint Pitch and DOA Estimation Using the ESPRIT Method. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):32–45, January 2015.

[769] Xhao Zhijin and Wu Jie. A new pitch detection method. In *WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th World Computer Congress 2000*, volume 2, pages 747–751, Beijing, China, 2000. IEEE.

[770] Johan Xi Zhang, Mads Græsbøll Christensen, Søren Holdt Jensen, and Marc Moonen. A Robust and Computationally Efficient Subspace-Based Fundamental Frequency Estimator. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):487–497, March 2010.

[771] Xiao-Dan Mei, Jengshyang Pan, and Sheng-He Sun. Efficient algorithms for speech pitch estimation. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*, pages 421–424, Hong Kong, China, 2001. IEEE.

[772] Xiaoshu Qian and R. Kumaresan. Joint estimation of time delay and pitch of voiced speech signals. In *Conference Record of The Twenty-Ninth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 735–739, Pacific Grove, CA, USA, 1996. IEEE Comput. Soc. Press.

[773] Xiaoshu Qian and R. Kumaresan. A variable frame pitch estimator and test results. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 228–231, Atlanta, GA, USA, 1996. IEEE.

[774] Xin Xu and Y. Miyanaga. A robust pitch detection in noisy speech with band-pass filtering on modulation spectra. In *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.*, volume 1, pages 266–269, Beijing, China, 2005. IEEE.

[775] Jian-Wu Xu and Jose C. Principe. A Novel Pitch Determination Algorithm based on Generalized Correlation Function. In *2007 IEEE Workshop on Machine Learning for Signal Processing*, pages 270–275, Thessaloniki, Greece, August 2007. IEEE.

[776] Jian-Wu Xu and Jose C. Principe. A Pitch Detector Based on a Generalized Correlation Function. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1420–1432, November 2008.

[777] Shuzhuang Xu and Hiroshi Shimodaira. Direct F0 Estimation with Neural-Network-Based Regression. In *Interspeech 2019*, pages 1995–1999. ISCA, September 2019.

[778] Xin Xu, Tian-qi Zhang, Sui Shi, and Ya-juan Zhang. An improved pitch detection of speech combined with speech enhancement. In *2014 7th International Congress on Image and Signal Processing*, pages 778–782, Dalian, China, October 2014. IEEE.

[779] Xu Gang and Tang Liang-rui. Speech pitch period estimation using circular AMDF. In *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003.*, pages 2452–2455, Beijing, China, 2003. IEEE.

[780] Xu Shao and B. Milner. Pitch prediction from MFCC vectors for speech reconstruction. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–97–100, Montreal, Que., Canada, 2004. IEEE.

[781] Xudong Jiang. Fundamental frequency estimation by higher order spectrum. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 1, pages 253–256, Istanbul, Turkey, 2000. IEEE.

[782] K. Yanagisawa, K. Tanaka, and I. Yamaura. Detection of the fundamental frequency in noisy environment for speech enhancement of a hearing aid. In *Proceedings of the 1999 IEEE International Conference on Control Applications (Cat. No.99CH36328)*, volume 2, pages 1330–1335, Kohala Coast, HI, USA, 1999. IEEE.

[783] Fan Yang, Sun-hua Xu, Ming-hui Liu, and Guo-feng Pan. Research on a new method of preprocessing and speech synthesis pitch detection. In *2010 International Conference On Computer Design and Applications*, pages V1–399–V1–401, Qinhuangdao, China, June 2010. IEEE.

[784] Xu-Kui Yang, Liang He, Dan Qu, and Wei-Qiang Zhang. Voice activity detection algorithm based on long-term pitch information. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1), December 2016.

[785] Yaping Yang and Yaoyao Zhang. Fundamental Frequency Extraction and Tone Recognition of Chinese Continuous Two-character-words. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 668–673, Singapore, June 2018. IEEE.

[786] Zhihua Yang, Daren Huang, and Lihua Yang. A Novel Pitch Period Detection Algorithm Based on Hilbert-Huang Transform. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Stan Z. Li, Jianhuang Lai, Tieniu Tan, Guocan Feng, and Yunhong Wang, editors, *Advances in Biometric Person Authentication*, volume 3338, pages 586–593. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[787] Yang Shao and DeLiang Wang. Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 2, pages II–205–8, Hong Kong, China, 2003. IEEE.

[788] Gao Yanhua and Zheng Guoqiang. Speech Pitch Period Detection Algorithm Based on Wavelet Transform and Spacial Correlation Function. In *2010 International Conference on Electrical and Control Engineering*, pages 5613–5616, Wuhan, China, June 2010. IEEE.

[789] J.-H. Yao, J.J. Shynk, and A. Gersho. Low-delay speech coding with adaptive interframe pitch tracking. In *Proceedings of ICC '93 - IEEE International Conference on Communications*, volume 1, pages 410–414, Geneva, Switzerland, 1993. IEEE.

[790] Q. Yasheng and P. Kabal. Pseudo-three-tap pitch prediction filters. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 523–526 vol.2, Minneapolis, MN, USA, 1993. IEEE.

[791] Y. Yazama, Y. Mitsukura, M. Fukumi, and N. Akamatsu. A Simple Algorithm of Pitch Detection by using Fast Direct Transform. In *2005 International Symposium on Computational Intelligence in Robotics and Automation*, pages 205–209, Espoo, Finland, 2005. IEEE.

[792] B. Yegnanarayana and K.S.R. Murty. Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):614–624, May 2009.

[793] B. Yegnanarayana, S. R. M. Prasanna, and S. Guruprasad. Study of robustness of zero frequency resonator method for extraction of fundamental frequency. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5392–5395, Prague, Czech Republic, May 2011. IEEE.

[794] Tzu-Chun Yeh, Ming-Ju Wu, Jyh-Shing Roger Jang, Wei-Lun Chang, and I-Bin Liao. A hybrid approach to singing pitch extraction based on trend estimation and hidden Markov models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 457–460, Kyoto, Japan, March 2012. IEEE.

[795] Yi-Yuan Wang and Li-Ming Zhao. Pitch frequency estimation of Chinese multi-syllable words based on wavelet transform. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, volume 6, pages 3765–3769, Shanghai, China, 2004. IEEE.

[796] Hui Yin, Climent Nadeu, Volker Hohmann, Xiang Xie, and Jingming Kuang. Order Adaptation of the Fractional Fourier Transform Using the Intraframe Pitch Change Rate for Speech Recognition. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4, Kunming, China, December 2008. IEEE.

[797] Xiang Yin, Ming Lei, Yao Qian, Frank K Soong, Lei He, Zhen-Hua Ling, and Li-Rong Dai. Modeling DCT Parameterized F0 Trajectory at Intonation Phrase Level with DNN or Decision Tree. *INTERSPEECH 2014*, page 5, 2014.

[798] G.S. Ying, L.H. Jamieson, and C.D. Michell. A probabilistic approach to AMDF pitch detection. In *Proceeding of Fourth International Conference on Spoken Language Processing ICSLP 96 ICSLP-96*, pages 1201–1204 vol.2, Philadelphia, PA, USA, 1996. IEEE.

[799] Yingjie Yang, Huanhuan Zhang, and Xiue Guo. A pitch tracking method mixing ACF & AMDF algorithms based on correlations. In *2011 International Conference on Image Analysis and Signal Processing*, pages 553–556, Wuhan, Hubei, China, October 2011. IEEE.

[800] Yipeng Li and DeLiang Wang. Detecting pitch of singing voice in polyphonic audio. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 3, pages 17–20, Philadelphia, Pennsylvania, USA, 2005. IEEE.

[801] F. Ykhlef and L. Bendaouia. Pitch Marking Using the Fundamental Signal for Speech Modifications via TDPSOLA. In *2013 IEEE International Symposium on Multimedia*, pages 118–124, Anaheim, CA, USA, December 2013. IEEE.

[802] Yong Duk Cho, Hong Kook Kim, Moo Young Kim, and Sang Ryong Kim. Pitch estimation using spectral covariance method for low-delay MBE vocoder. In *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, pages 21–22, Pocono Manor, PA, USA, 1997. IEEE.

[803] Kazuyoshi Yoshii and Masataka Goto. Infinite kernel linear prediction for joint estimation of spectral envelope and fundamental frequency. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 463–467, Vancouver, BC, Canada, May 2013. IEEE.

[804] Kota Yoshizato, Hirokazu Kameoka, Daisuke Saito, and Shigeki Sagayama. Hidden Markov Convolutive Mixture Model for Pitch Contour Analysis of Speech. *INTERSPEECH 2012*, page 4, 2012.

[805] An-Tze Yu and Hsiao-chuan Wang. New Harmonicity Measures for Pitch Estimation and Voice Activity Detection. *INTERSPEECH 2004*, page 4, 2004.

[806] R. Yu and E. C. Tan. Comparison of different time-frequency distributions in pitch detection. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 2, pages 817–821. IEEE, 2003.

[807] Yu-Min Zeng, Zhen-Yang Wu, Hai-Bin Liu, and Lin Zhou. Modified AMDF pitch detection algorithm. In *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*, pages 470–473, Xi'an, China, 2003. IEEE.

[808] Wang Yue, Qian Zhihong, and Zhang Ying. A New Pitch Detection Algorithm Based on RCAF. In *2009 WRI World Congress on Computer Science and Information Engineering*, pages 126–130, Los Angeles, California USA, 2009. IEEE.

[809] Yumin Zeng, Yi Zhang, and Ping Li. Improvement of pitch detection in noisy speech based on wavelet transform. In *IET International Conference on Wireless Mobile and Multimedia Networks Proceedings (ICWMMN 2006)*, volume 2006, pages 274–274, Hangzhou, China, 2006. IEE.

[810] P. Yutthagowith and A. Kunakorn. Fast, simple, accurate fundamental frequency estimation. In *12th IET International Conference on AC and DC Power Transmission (ACDC 2016)*, pages 109 (4 .)–109 (4 .), Beijing, China, 2016. Institution of Engineering and Technology.

[811] M.R. Zad-Issa and P. Kabal. A new LPC error criterion for improved pitch tracking. In *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, pages 1–2, Pocono Manor, PA, USA, 1997. IEEE.

[812] Stephen A Zahorian, Princy Dikshit, and Hongbing Hu. A Spectral-Temporal Method for Pitch Tracking. *INTERSPEECH 2006*, page 4, 2006.

[813] L. Zao and R. Coelho. On the Estimation of Fundamental Frequency From Nonstationary Noisy Speech Signals Based on the Hilbert–Huang Transform. *IEEE Signal Processing Letters*, 25(2):248–252, February 2018.

[814] Li Zeng, Liang Chen, and Qiang Xiao. Pitch period estimation base on voiced degree weighted sub-frame octave region dynamic programming. In *2010 International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–5, Suzhou, China, October 2010. IEEE.

[815] Zeng Zhihua and Xiao Zimei. Fast VQ of multi-tap pitch predictor coefficients. In *ICSP '98. 1998 Fourth International Conference on Signal Processing (Cat. No.98TH8344)*, pages 580–583, Beijing, China, 1998. IEEE.

[816] Jihen Zeremdini, Mohamed Anouar Ben Messaoud, and Aicha Bouzid. Multiple fundamental frequencies estimation approaches based on multi-scale product analysis. In *2017 3rd International Conference on Frontiers of Signal Processing (ICFSP)*, pages 55–58, Paris, September 2017. IEEE.

[817] Jihen Zeremdini, Mohamed Anouar Ben Messaoud, and Aicha Bouzid. Multi-pitch estimation based on multi-scale product analysis, improved comb filter and dynamic programming. *International Journal of Speech Technology*, 20(2):225–237, June 2017.

[818] Geliang Zhang and Simon Godsill. Tracking pitch period using particle filters. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, New Paltz, NY, USA, October 2013. IEEE.

[819] Geliang Zhang and Simon Godsill. Fundamental Frequency Estimation in Speech Signals With Variable Rate Particle Filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):890–900, May 2016.

[820] Haojie Zhang and Yong Yang. Fundamental frequency adjustment and formant transition based emotional speech synthesis. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1797–1801, Chongqing, Sichuan, China, May 2012. IEEE.

[821] Hui Zhang, Xueliang Zhang, Shuai Nie, Guanglai Gao, and Wenju Liu. A pairwise algorithm for pitch estimation and speech separation using deep stacking network. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250, South Brisbane, Queensland, Australia, April 2015. IEEE.

[822] Jianshu Zhang, Jian Tang, and Li-Rong Dai. RNN-BLSTM Based Multi-Pitch Estimation. In *INTERSPEECH 2016*, pages 1785–1789, September 2016.

[823] Qi Zhang, Chong Cao, Tiantian Li, Yanlu Xie, and Jinsong Zhang. Pitch Range Estimation with Multi features and MTL-DNN Model. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 939–943, Beijing, China, August 2018. IEEE.

[824] Shiming Zhang, Yosuke Sugiura, Tetsuya Shimamura, and Hisanori Makinae. Fundamental frequency estimation combining air-conducted speech with bone-conducted speech in noisy environment. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 244–247, Cox's Bazar, Bangladesh, February 2017. IEEE.

[825] Wei Zhang, Qi Zhang, Yanlu Xie, and Jinsong Zhang. LSTM-Based Pitch Range Estimation from Spectral Information of Brief Speech Input. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 349–353, Taipei City, Taiwan, November 2018. IEEE.

[826] Wenyao Zhang, Gang Xu, and Yuguo Wang. Pitch estimation based on Circular AMDF. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages I–341–I–344, Orlando, FL, USA, May 2002. IEEE.

[827] Xiaoheng Zhang and Yongming Li. Pitch tracking algorithm based on evolutionary computing with regularisation in very low SNR. *The Journal of Engineering*, 2018(16):1509–1514, November 2018.

[828] Xiu Zhang, Wei Li, and Bilei Zhu. Latent time-frequency component analysis: A novel pitch-based approach for singing voice separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, South Brisbane, Queensland, Australia, April 2015. IEEE.

[829] Xueliang Zhang, Wenju Liu, and Bo Xu. Multi-pitch determination algorithm based on mixture laplacian distribution. In *2010 International Conference on Audio, Language and Image Processing*, pages 37–41, Shanghai, China, November 2010. IEEE.

[830] Xueliang Zhang, Hui Zhang, Shuai Nie, Guanglai Gao, and Wenju Liu. A Pairwise Algorithm Using the Deep Stacking Network for Speech Separation and Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1066–1078, June 2016.

[831] Yiyan Zhang, Wenju Liu, Bo Xu, and Huayun Zhang. Pitch and tone's modeling in parametric trajectory model. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages IV–4160–IV–4160, Orlando, FL, USA, May 2002. IEEE.

[832] Yuhong Zhang, Aaron C. Elkins, and Jay F. Nunamaker. Pitch detection algorithms modifications and implementations towards automated vocal analysis. In *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control*, pages 405–410, Miami, FL, USA, April 2014. IEEE.

[833] Zhenwei Zhang and Yingyun Yang. Comparison of Speech Signal Features between Chinese and English. In *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pages 312–315, Hangzhou, China, August 2016. IEEE.

[834] Zhang Peng, Wang Lihong, and Liu Sheng. On fundamental frequency contour synthesis and control method for Chinese Speech Synthesis. In *2008 27th Chinese Control Conference*, pages 739–742, Kunming, China, July 2008. IEEE.

[835] Zhang Sen and K. Shirai. Visual approach for automatic pitch period estimation. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1339–1342, Istanbul, Turkey, 2000. IEEE.

[836] Zhang Xiongwei and Chen Xianzhi. Pitch determination based on LMS algorithm. In *1991 International Conference on Circuits and Systems*, pages 26–28, Shenzhen, China, 1991. IEEE.

[837] Wanda W Zhao and Tokunbo Ogunfunmi. Formant and Pitch Detection Using Time-Frequency Distribution. *International Journal of Speech Technology*, 3:15, 1999.

[838] Zhao Heming, Zhu Qi, Yu Yibiao, and Chen Xueqin. Pitch detection of nosy speech signal based on Bark wavelet transform. In *6th International Conference on Signal Processing, 2002.*, pages 418–421, Beijing, China, 2002. IEEE.

[839] Zhen-Dong Zhao, Xi-Mei Hu, and Jing-Feng Tian. An effective pitch detection method for speech signals with low signal-to-noise ratio. In *2008 International Conference on Machine Learning and Cybernetics*, pages 2775–2778, Kunming, China, July 2008. IEEE.

[840] Yibin Zheng, Zhengqi Wen, Bin Liu, Ya Li, and Jianhua Tao. An initial research: Towards accurate pitch extraction for speech synthesis based on BLSTM. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pages 165–170, Chengdu, China, November 2016. IEEE.

[841] Zhiyao Duan, B Pardo, and Changshui Zhang. Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, November 2010.

[842] Zhongqiang Ding, I.V. McLoughlin, and E.C. Tan. How to track pitch pulses in LP residual ? - joint time-frequency distribution approach. In *2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE Cat. No.01CH37233)*, volume 1, pages 43–46, Victoria, BC, Canada, 2001. IEEE.

[843] Qunqun Zhou, Yong Ma, Shengqing Wang, Ruijin Lu, and Hongyuan Wang. Adaptive two-direction pitch tracking algorithm for MBE vocoder. *IEEE Transactions on Consumer Electronics*, 57(4):1800–1806, November 2011.

[844] Ruohua Zhou, Joshua D. Reiss, Marco Mattavelli, and Giorgio Zoia. A Computationally Efficient Method for Polyphonic Pitch Estimation. *EURASIP Journal on Advances in Signal Processing*, 2009(1):729494, December 2009.

[845] Zhenhua Zhou, M. G. Christensen, J. R. Jensen, and H. C. So. Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6812–6816, Vancouver, BC, Canada, May 2013. IEEE.

[846] Jian-wei Zhu, Shui-fa Sun, Xiao-li Liu, and Bang-jun Lei. Pitch in Speaker Recognition. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, pages 33–36, Shenyang, China, 2009. IEEE.

[847] M. Zivanovic and J. Schoukens. Time-variant harmonic and transient signal modeling by joint polynomial and piecewise linear approximation. In *2010 18th European Signal Processing Conference*, pages 467–471, August 2010.

[848] Miroslav Zivanovic and Johan Schoukens. Time-variant harmonic signal modeling by using polynomial approximation and fully automated spectral analysis. *17th European Signal Processing Conference (EUSIPCO 2009)*, page 5, 2009.

[849] Miroslav Zivanovic and Johan Schoukens. Constrained time-variant signal modeling for identifying colliding harmonics in sound mixtures. *19th European Signal Processing Conference (EUSIPCO 2011)*, page 5, 2011.

[850] Yuan Zong, Yumin Zeng, Mengchao Li, and Rui Zheng. Pitch detection using EMD-based AMDF. In *2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*, pages 594–597, Beijing, China, June 2013. IEEE.

[851] Alireza Zourmand and Ting Hua Nong. Vowel Classification of Children's Speech Using Fundamental and Formant Frequencies. In *2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation*, pages 282–287, Kuantan, Malaysia, September 2012. IEEE.

[852] Zuzhen Feng, Xuanhao Ding, and Yingchun Jiang. Pitch period estimation of voice signal based on EEMD and Hilbert Transform. In *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010)*, pages 365–368, Wuhan, China, March 2010. IEEE.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig verfasst habe und nur die angegebenen Quellen und Hilfsmittel verwendet habe. Die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichungen, wie sie von den Leitlinien guter wissenschaftlier Praxis der Carl von Ossietzky Universät Oldenburg festgelegt sind, wurden befolgt.

Oldenburg, 8.10.2020

.....
(Bastian Bechtold)