Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6658951, 17 pages
https://doi.org/10.1155/2021/6658951

Research Article

Harmonic Differences Method for Robust Fundamental Frequency Detection in Wideband and Narrowband Speech Signals

Cevahir Parlak[1] **and Yusuf Altun**[2]

[1]*Department of Computer Engineering, Institute of Graduate Studies in Science and Engineering, Duzce University, Duzce, Turkey*
[2]*Department of Computer Engineering, Faculty of Engineering, Duzce University, Duzce, Turkey*

Correspondence should be addressed to Cevahir Parlak; cevahirparlak@gmail.com

Received 26 April 2021; Revised 14 August 2021; Accepted 16 August 2021; Published 19 October 2021

Academic Editor: Ali Ahmadian

In this article, a novel pitch determination algorithm based on harmonic differences method (HDM) is proposed. Most of the algorithms today rely on autocorrelation, cepstrum, and lastly convolutional neural networks, and they have some limitations (small datasets, wideband or narrowband, musical sounds, temporal smoothing, etc.), accuracy, and speed problems. There are very rare works exploiting the spacing between the harmonics. HDM is designed for both wideband and exclusively narrowband (telephone) speech and tries to find the most repeating difference between the harmonics of speech signal. We use three vowel databases in our experiments, namely, Hillenbrand Vowel Database, Texas Vowel Database, and Vowels from the TIMIT corpus. We compare HDM with autocorrelation, cepstrum, YIN, YAAPT, CREPE, and FCN algorithms. Results show that harmonic differences are reliable and fast choice for robust pitch detection. Also, it is superior to others in most cases.

1. Introduction

Pitch is an extraordinarily complicated and distinct feature of human speech and plays a major role in the perception of human conversations as well as in human-computer interactions. Pitch detection has a strong and disputed background spanning more than a century. Myriad of methods have been proposed, but it is still a formidable task especially in narrowband (telephone), noisy, multipitch, and multitalker speech with reasonable resolution and fast implementation due to extremely complicated structure of frequency spectrum. Pitch helps us to identify some of the important cues about the speaker, such as the identity, gender, emotional state, or about the tones of a musical instrument. It has a wide range of applications in emotion and gender recognition, speech synthesis, human-computer interaction, and detection of symptoms of pathological disorder at early stages. Fundamental frequency is the quantity of pitch and is measured on the periodic signals (musical tones) or quasiperiodic signals (speech). Pitch detection algorithms (PDA) and pitch tracking algorithms are extensively used to extract the fundamental frequency of a person's speech or of a musical tone. Fundamental frequency is the frequency of vocal cord oscillation, and it can highly vary among the men, women, boys, and girls. Therefore, exact calculation is a crucial factor in variety of applications spanning from human-computer interaction to early detection of pathological symptoms.

This article is organized as follows: in Section 1, we discuss the foundations and importance of pitch detection and tracking, Section 2 deals with literature overview, historical background, algorithms, novelties of this article, datasets, ground truth methods, error measures, difficulties, application areas, and related algorithm domains, Section 3 describes the novel HDM algorithm, Section 4 delineates datasets used in this article and experimental setup, Section 5 presents the

results of wide and narrowband experiments, Section 6 is devoted to gender detection results, and finally, in Section 7, we conclude this work with the evaluations and future studies.

2. Literature Review

2.1. Historical Background. The saga of pitch determination and tracking begins with the dispute between August Seebeck and George Simon Ohm on the mysterious missing fundamental concept. In 1841, August Seebeck showed that the pitch of a sound did not depend on the tone having a fundamental frequency component of the pitch frequency. The debate over missing fundamental began shortly after this [1–3]. In 1843, Ohm severely rejected this idea and stated that the quality of a tone depends solely on the number and relative strength of its partial simple tones [4]. This law is championed by Helmholtz in his masterpiece work [5].

Rayleigh, in his famous study, The Theory of Sound (1877), claimed that pitch could not be simply associated with period using his experiments with sirens. But similar to loudness and timbre, pitch is not a thing to be measured directly [6]. In 1924, Harvey Fletcher proved that even if several lower harmonics of a waveform were removed, the pitch remained the same. The pitch was very closely related to the difference in frequency, even though that frequency did not exist in the sound source [7]. This idea is one of the bases of this study to exploit these differences to extract the fundamental frequency.

In 1938, Schouten proved that the missing fundamental effect could not be explained as a nonlinear difference as Helmholtz had claimed and strongly supported the Seebeck [8]. Schouten's theory is known as the residue theory of pitch, periodicity pitch, or virtual pitch [9–11]. Today, scientists agreed that Seebeck disproved the Ohm's idea. Missing fundamental has a practical implementation in our telephone conversation. In telephone signals, the frequency spectrum is limited between 300 Hz and 4000 Hz. But we can still perceive the voice clearly and discriminatively to capture the talker's identity, gender, and even emotional state. For instance, if we have a signal with harmonics at 400 Hz, 600 Hz, 800 Hz, 1000 Hz, and 1200 Hz, the pitch of the signal is still perceived as 200 Hz, and autocorrelation of both signals remains nearly same as depicted in Figure 1 [12].

The autocorrelation model of pitch perception dates back to Licklider's "duplex" and "triplex" models. Licklider solved the dilemma between Seebeck and Ohm definitely in favor of Seebeck. Periodicity pitch theory succeeded place pitch theory after Licklider's "duplex" and "triplex" theories [13–16]. This theory was investigated deeply by Ritsma, and some of its limitations were shown [17]. Today, the debate is far from over, and two mainstream theories compete against each other paving the way for Place Code of Ohm & Helmholtz and Temporal Code Theory of Seebeck & Wever of hearing.

2.2. Algorithms. First works started around 1950s with AMDF because in the early days of computers, the multiplication of autocorrelation function was replaced with subtraction due to the high computational cost of multiplication. The fierce debate on missing fundamental led to the calculation of fundamental frequency, and many algorithms flourished to extract and track the pitch of a signal. Among the most notables are AMDF [18–20], autocorrelation [13–16, 21–25], cepstrum [26–28], harmonic product spectrum, period histograms [29–31], parallel processing methods [32–34], simplified inverse filter tracking (SIFT) [35], comb filters [36], data reduction [37], LPC-based spectral equalization (unpublished), spectral sieves [38], harmonic spacings and structures [39–41], LPC inverse filtering [20], feature based [42], IPTA [43], harmonic pattern recognition [44], envelop analysis, threshold-crossing analysis (ZXABE, TABE, TTABE) [45, 46], subharmonic summation [47], subband processing [48], superresolution [49], two-way mismatch [50], resolution improvement [51], TEMPO [52], RAPT (NCCF) [53], instantaneous frequency [54–57], STRAIGHT-NDF [58, 59], wavelet based [60–62], joint time-frequency analysis (JTFA) [63], PRAAT [64], FIR filter, subharmonic-to harmonic ratio [65], YIN [66], YAAPT [67, 68], HMM for multipitch tracking, PEPSI_LITE, ESACF, PEBS (block sparsity), PEARLS [69–78], chirp transform [79], sawtooth waveform inspired pitch estimator (SWIPE) [80], perturbation spectrum (TANDEM STRAIGHT) [81], ANLS [82], vocal fold vibration (DIO) [83], ensemble empirical mode decomposition and entropy [84, 85], event-based instantaneous fundamental frequency with impulse-like characteristics of excitation in glottal vibrations [86], tandem algorithm [87], summation of residual harmonics (SRH) [88], glottal closure instants, EGG, glottal opening instants (GOI, GCI) [89–91], PEFAC [92], SAFE [93], SAcC [94], BANA [95, 96], adaptive harmonic model (aHM) with adaptive iterative refinement (AIR) [97], MPM [98], BSURE-IR [99], time-warping for fast f0 changes [100], robust harmonic features with multilayer perceptron [101], HARVEST [102], Pyin [103], off-grid [104], harmonics of impulse-like excitation [105], phase space [106], normalized squared difference function, LASSO regression based [107], adaptive total variation penalty, single frequency filtering and temporal envelopes [108, 109], CNN and DNN (CREPE, Deep Pitch, FCN) [110–117] are among these methods.

Schroeder histogram, Brown comb filter, modified Maher–Beauchamp pitch detector, Meddis–Hewitt model, McAulay–Quatieri's method, and cepstral detector are used in pitch detection and tracking of musical signals. The challenges in multipitch tracking are overlapped harmonics (octaves, fifths), noise, threshold of detection, scarcity of labeled datasets, timbre complexity (one pitch played by multiple instruments), and efficiency. F0 is also an important feature in emotional speech recognition [118–120]. There are many reviews about the methods and datasets of pitch detection and pitch tracking [12, 121–134].

2.3. Innovations in This Article. Pitch detection is an extensively studied research field. Today pitch detection methods are successful in wideband, clean human speech. But there are still too many formidable obstacles that need to be resolved particularly in narrowband (telephone speech) [67, 68], multipitch [69–78], multitalker, and noisy speech

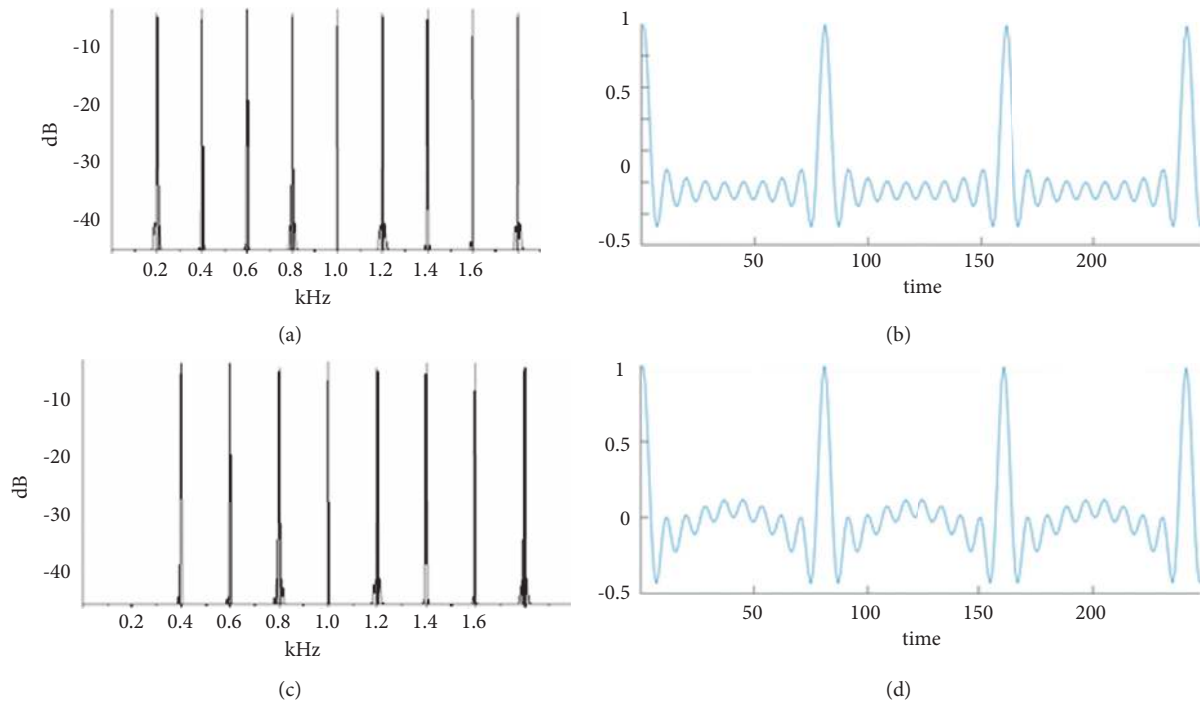

Figure 1: (a) A signal with multiple harmonics and (b) its autocorrelation function. (c) Same signal with the first harmonic removed and (d) its autocorrelation function.

signals [69, 95, 96, 108, 109]. Various techniques have been implemented including autocorrelation, cepstrum, phase based, and lastly, CNN methods. However, an interesting clue for pitch determination is highly neglected and rarely used. Harmonic spacings are strong pieces of evidence for pitch determination even though human speech is extraordinarily quasiperiodic. This method was first used by Seneff [39] for narrowband speech on real-time data between 210 Hz and 1050 Hz band by iteratively assigning weight factors to the pitch candidates obtained from harmonic spacings with LDVT (Lincoln Digital Voice Terminal). Seneff's work is too old, and dataset and results are not clearly presented. There are other works till 21st century. Wu [40] tried harmonic spacings in guitar sound using GCD (Greatest Common Divisor of largest peak) with a table, and Dziubiński and Kostek in various musical instruments with an Artificial Neural Network on a matrix of sets of harmonics. We will apply this method both for wideband and narrowband speech (400 Hz–3400 Hz) signals on large speech datasets, eliminating some of the limitations while spurring the accuracy and speed via determining the most repeated difference between the harmonics using a histogram. This article will revive this technique and demonstrate that harmonic spacings can reliably be used to obtain state-of-the-art results fast and efficiently both in wideband and narrowband telephone speech samples. More detailed explanations are presented in Section 3.

2.4. Datasets. As usual with the other experiments, creation and use of proper datasets is essential in pitch tracking experiments. There are numerous datasets used in this field and among them Keele Studio, PTDB-TUG [135], Keele Telephone [136], TIMIT [137], NTIMIT [138], CSTR [139], FARSDAT [140], Mocha TIMIT [141], RWC Music Database [142], MedleyDB [143], Vowel-CVC [144], NOIZEUS [145], and SPEECON [146] datasets. Some authors used CMU ARCTIC, KED TIMIT [147], APLAWD [148], BACH10 [149], SyncRWC60, Saarland Music Data [150], and Mazurka Dataset [151]. Vowel datasets can also be used for pitch determination. Because f0 tracking is specifically important in music transcription, many music datasets are available in ISMIR (https://www.ismir.net/resources/) web pages.

Pitch detection can be implemented in clean wideband, narrowband telephony, noisy speech, and multipitch musical sounds. In telephony speech, signal is usually band-passed between 300 Hz and 4000 Hz to save the bandwidth. In this study, we applied band-pass filter twice to remove all remains between 400 Hz and 3400 Hz. There are many types of noises that can be added to the pitch datasets, including babble noise, exhibition noise, HF (high-frequency) channel noise, restaurant noise, street noise, white noise, pink, brown, and pub noises. NOISEX [153] dataset is a publicly available noise dataset. Pitch datasets can be recorded in studio, office, living rooms, and car interior environments.

In pitch datasets, speaker profiles, distribution of gender, distribution of age, mother tongue, distribution of dialect, distribution of profession or education, pathologies, number of speakers, contents, speaking style, read speech, answering speech, command and control speech, descriptive speech, nonprompted speech, spontaneous speech, neutral and emotional speech, general recording setup (telephone, on-site, field, wizard-of-oz), annotation,

technical specifications (sampling rate, sample type, number of channels, file formats), corpus structure, release plan and validation procedure, meta data, recording protocol (session id, speaker id, recording date, environmental conditions, technical recording conditions), postprocessing, pronunciation dictionary, and validation are important issues [135].

2.5. Ground Truth Determination of Pitch Datasets. Another important issue for pitch datasets is the exact determination of ground truth. Hand-editing, EGG, and Laryngograph are widely used as ground truth methods. SIGMA, Hilbert Envelope-based detection (HE), the Zero Frequency Resonator-based method (ZFR), the Dynamic Programming Phase Slope Algorithm (DYPSA), the Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS), and Yet Another GCI Algorithm (YAGA) are some other methods used for finding ground truth of pitch samples [154–158]. They have some advantages and disadvantages, and we usually need hand-editing by an expert. In some cases, detecting f0 manually by human experts can be quite difficult. EGG, laryngogram, particularly differentiated laryngogram provides a signal, which makes automatic f0 calculation easier using available methods as shown in Figure 2.

2.6. Error Measures and Pitch Refinement Techniques. Gross Pitch Error (GPE), Fine Pitch Error (FPE), Voicing Decision Error (VDE), f0 Frame Error (FFE) [88, 159, 160], MOS (Mean Opinion Score) [161], Diagnostic Rhyme Test (DRT) [162], Perceptual Evaluation of Speech Quality (PESQ) [163], and Perceptual Objective Listening Quality Analysis (POLQA) [164] are prominent metrics used to evaluate pitch detection algorithms.

Pitch estimation derived from the discrete Fourier spectrum can be improved using techniques, such as Spectral Reassignment (phase based) [165–167], super sampling, interpolation (Grandke, quadratic, Gaussian) [168–171], Matching pursuit [172], Synchrosqueezing [173–176], and Empirical Mode Decomposition [177], to compensate for the resolution problem of FFT.

Vocal Tract Model, Uniform Lossless Tube, and Two-Tube Model can be used to model the f0 and other formant frequency structures [178–186]. Using the laws of conservation of mass, momentum, and energy, it can be proved that sound wave propagation inside a lossless tube satisfies the equations:

$$-\frac{\partial \mathbf{p}}{\partial \mathbf{x}} = \boldsymbol{\rho}\frac{\partial(\mathbf{u}/\mathbf{A})}{\partial \mathbf{t}}, \qquad -\frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{1}{\boldsymbol{\rho}\mathbf{c}^2}\frac{\partial(\mathbf{pA})}{\partial \mathbf{t}} + \frac{\partial A}{\partial \mathbf{t}}, \tag{1}$$

where $p = p\,(x, t)$ is the pressure of sound at position x and time t, $u = u\,(x, t)$ is the velocity at position x and time t, ρ is air density inside the tube, c is sound velocity, and $A = A\,(x, t)$ is the cross-sectional area function of the distance and time.

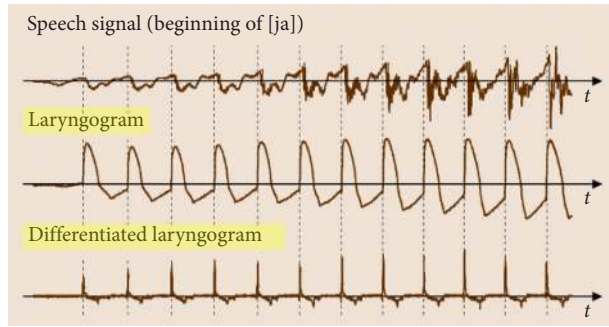

Figure 2: Speech signal, laryngogram, and differentiated laryngogram [156].

2.7. Difficulties in Pitch Tracking. It may be helpful at this point to summarize the difficulties in f0 tracking and voiced detection:

(i) F0 may change in each glottal period
(ii) Subharmonics and spurious harmonics
(iii) Inharmonicity and quasiharmonicity
(iv) Doubling or halving of f0
(v) Formants and band-limiting
(vi) Irregularity of voicing at voice onset and offset
(vii) Even experts cannot always agree on the locations of voice onset and offset
(viii) Narrowband filtering of unvoiced excitation may create periodic signals
(ix) The amplitude of voiced speech may vary in a wide range
(x) Background noise
(xi) Some voiced speech may have very little glottal pulse duration
(xii) Determining the start and end of each pitch period in voiced speech parts
(xiii) Shimmer: amplitude variation from one cycle to the next
(xiv) Jitter: frequency variation from one cycle to the next
(xiv) Breathy voice
(xvi) Inherent quasiperiodicity of human speech
(xvii) Multiple periodicities in music signals
(xviii) Multitalker signals
(xix) Transient regions
(xx) Estimation of pitch with low energy can be difficult
(xxi) Low formant F1 may interfere with high f0 of females and children

(xxii) In telephony systems, nonlinear effects such as phase distortion, crosstalk, clipping of high-level sounds, and amplitude modulation

2.8. Application Areas. Today determination of fundamental frequency is an important factor used in a wide range of solutions:

(i) Emotion recognition and human-computer interaction
(ii) Gender determination (male/female/boy/girl)
(iii) Detection of the pathological characteristics of the voice
(iv) Speech understanding systems
(v) Prosody analysis
(vi) Speaker identification and separation
(vii) Data-driven speech synthesis
(viii) Digital hearing prostheses
(ix) Computer-aided intonation teaching

2.9. Domains of Methods. Pitch detection algorithms can be classified as time, frequency, and hybrid methods. Some of these algorithms use temporal smoothing (median filter, HMM, Viterbi etc.).

2.9.1. Time-Domain Methods. Many methods rely only on the time-related information for pitch calculation.

(i) Threshold-crossing pitch detectors (ZXABE, TABE, and TTABE, Zero-Crossing Rate, Peak rate, Slope event rate)
(ii) Parallel processing method
(iii) Average Magnitude Difference Function (AMDF)
(iv) Envelope analysis methods
(v) Autocorrelation
(vi) YIN
(vii) Phase space
(viii) Deep learning and convolutional neural networks (end to end).

2.9.2. Frequency-Domain Methods. Frequency spectrum is widely used to detect the pitch value of a signal.

(i) Optimum comb filter
(ii) Tunable IIR filter
(iii) FIR filter
(iv) Cepstrum
(v) Harmonics
(vi) Multiresolution methods
(vii) Statistical frequency-domain methods
(viii) Neural networks
(ix) Maximum likelihood estimators

2.9.3. Time- and Frequency-Domain (Hybrid) Pitch Detectors. Some algorithms combine time and frequency domain to calculate fundamental frequency:

(i) Schroeder histogram
(ii) Brown comb filter method
(iii) Maher–Beauchamp pitch detector
(iv) Meddis–Hewitt model
(v) McAulay–Quatieri method
(vi) Neural networks
(vii) YAAPT

3. HDM Algorithm

The novel harmonic differences method (HDM) tries to exploit the differences between the harmonics of power spectrum of the signal with the goal of finding the most repeating difference. Harmonic spacings have been tried by Seneff [39] for telephony speech on real-time data with LDVT, Wu on guitar sound [40], and Dziubiński and Kostek [187] for various musical instrument sounds. Seneff used the area under the hump in frequency spectrum, creating a list of 8 estimates and applies temporal smoothing using median filter. We try to find the most repeating difference without temporal smoothing. We will extend these works by eliminating many steps in the implementation and generalizing the method to all kinds of sound waves, improving the speed and the accuracy. Matlab 2019 student version is used for HDM, autocorrelation, cepstrum, YIN, and YAAPT. Python 3.6.5 is used for CREPE and FCN implementations. General outline and pseudocode of the algorithm is as follows:

(1) Take the discrete Fourier transform and run the peak picking algorithm between the minimum f0 and maximum threshold frequency F_{th}. Length of the samples is taken as 1024 and Hamming windowed before the processing to smooth the signal and avoid the edge effects. Windowing is necessary due to the DFT's vulnerability to discontinuities. Windowed Fourier Transform is given as follows:

$$H_k = \sum_{i=0}^{N-1} x_i w_i e^{-(j2\pi ki/N)}. \tag{2}$$

Sampling rate is 16000 Hz yielding a frequency resolution of 15.625 Hz. Despite this low FFT resolution, the results of HDM are very promising. During this search, the minimum power of a partial should be at least $(1/\gamma)$ times the magnitude of the largest partial. Here, γ is a prespecified constant such that

$$|H_k| > \frac{|H_{\max}|}{\gamma}, \tag{3}$$

$\alpha, \beta, \gamma, f0_{\min}, F_{\text{th}}$ (upper-limit threshold frequency) are determined empirically with more than 7 million experiments on Hillenbrand dataset. There is no limit in the number of peaks to collect, but we set a

threshold frequency value F_{th}, and we collect all peaks below that threshold. This threshold frequency is also determined empirically.

(2) To decide a true peak, its amplitude must be larger than the previous and next Fast Fourier coefficient.

J $\longleftarrow 0$
if $\widehat{f}_i > f0_{\min}$ and $\widehat{f}_i < F_{\text{th}}$ **then**
  if $A_{i-1} < A_i > A_{i+1}$ **then**
    if $A_i > (|H_{\max}|/\gamma)$ **then**
      $\widehat{\mathscr{F}}_i \longleftarrow \widehat{f}_i$, $J = J + 1$
    end if
  end if
end if

(3) In this list, remove the unnecessary peaks that are closer than minimum f0 value to one another.

if $\widehat{f}_i < \widehat{f}_{i-1} + f0_{\min}$ and $\widehat{A}_i < \widehat{A}_{i-1}$ then
  remove $\widehat{f}_i$ from $\widehat{\mathscr{F}}$, $J \longleftarrow J - 1$
end if
if $\widehat{f}_i > \widehat{f}_{i+1} - f0_{\min}$ and $\widehat{A}_i < \widehat{A}_{i+1}$ **then**
  remove $\widehat{f}_i$ from $\widehat{\mathscr{F}}$, $J \longleftarrow J - 1$
end if

(4) In this list, remove the spurious peaks that are weaker than a predefined empirical ratio of the previous and next spectrum value.

if $A_i < (A_{i+1}/\alpha)$ or **if** $A_i < (A_{i-1}/\alpha)$ **then**
  remove $\widehat{f}_i$ from $\widehat{\mathscr{F}}$, $J \longleftarrow J - 1$
end if

(5) Handle some special cases such as frequencies below 110 so that if the amplitude of first entry is less than 1/5 of the second entry remove first entry.

if $\widehat{f}_1 < 110$ Hz and $A_1 < (A_2/\beta)$ **then**
  remove $\widehat{f}_1$ from $\widehat{\mathscr{F}}$, $J \longleftarrow J - 1$
end if

(6) Find the differences between the adjacent entries in the list

$$\widehat{D}\colon \left\{\widehat{d}_i \mid \widehat{d}_i = \widehat{f}_i - \widehat{f}_{i+1}, \widehat{d}_i > f0_{\min}, \quad i = 1, 2, \ldots, N\right\}. \tag{4}$$

(7) Sort the list according to the repetition counts with a histogram.

//$f0_{wb}$: **wideband fundamental frequency**.
//$f0_{nb}$: **narrowband fundamental frequency**.
//$A0_{wb}$: **amplitude of wideband fundamental frequency**.
//$A0_{nb}$: **amplitude of narrowband fundamental frequency.**

$$\begin{aligned} \text{list} &= \text{sort}(\text{list}, \text{repeat count}),\\ f0_{nb} &= \text{list}(1,1),\\ \widehat{D} &= \widehat{\mathscr{F}} - f0_{nb},\\ \widehat{D} &= [\widehat{\mathscr{F}}, \widehat{D}],\ \widehat{d}_i > 0. \end{aligned} \tag{5}$$

(8) Sort the final list according to the differences. First entry is the most repeating difference

//**remove zero elements if any**

$$\begin{aligned} \text{list} &= \text{sort}\left(\text{list}, \widehat{d}_i\right),\\ f0_{wb} &= \text{list}(1,1),\\ A0_{wb} &= \text{list}(1,2),\\ A0_{nb} &= A0_{wb}, \end{aligned} \tag{6}$$

if $abs(f0_{wb} - f0_{nb}) < (f0_{\min}/2)$ **then**
$f0_{wb} = f0_{nb}$
end if
else
$f0_{wb} \longleftarrow 0$, $f0_{nb} \longleftarrow 0$
return $\{f0_{wb}, f0_{nb}\}$

HDM does not use temporal smoothing. Although it is possible to obtain better results using different parameters for each dataset, we use the same parameter set for all 3 datasets and for all wide and narrowband experiments. Our goal is to find a global parameter set that can achieve best results for wideband and narrowband telephone speech in all datasets.

The relation between GPE and α is shown in Figure 3. In the small values of α, GPE goes too high, and after 5, it remains nearly constant. A typical value for α is between 6 and 12. This parameter is used to eliminate the spurious peaks in the list $\widehat{\mathscr{F}}$ of candidate pitches. Elimination of redundant peaks is a key point in capturing the correct spacing between the harmonics.

β parameter is used to handle some special cases such as frequencies below 110 Hz. Due to nonlinearity and complexity of equal loudness curves, different regions of human speech spectrum have different loudness. Therefore, we used β parameter for handling low-frequency components. The effect of α and β parameters on GPE is depicted in Figure 4. Linear region of human auditory system is usually considered between 20 and 1000 Hz (we do believe that a value between 1100 and 1200 Hz is a better choice for upper limit), and logarithmic region is the rest of the audible spectrum [188–190].

Another important parameter is the value of minimum f0, and it heavily affects the GPE as shown in Figure 5. Hillenbrand and Texas datasets provide ground truth value for f0, and minimum fundamental frequency value is 82 Hz.

4. Datasets and Experimental Setup

In this study, we employ 3 vowel datasets: Hillenbrand dataset (http://homepages.wmich.edu/~hillenbr/voweldata.html) [191–193], Texas Vowel dataset (https://personal.utdallas.edu/~assmann/KIDVOW1/North_Texas_vowel_database.html) [194, 195], and vowel part of TIMIT [137] dataset including the SA samples. Hillenbrand dataset is a collection of 1668 vowels, including boy, girl, man, and woman speech samples. Total length of the samples is

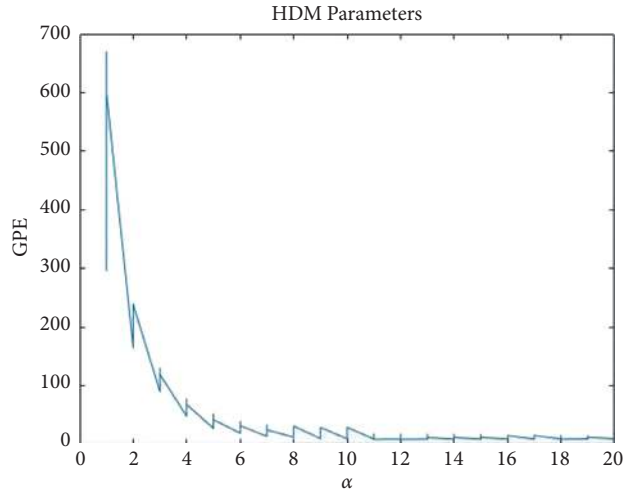

Figure 3: GPE vs. α.

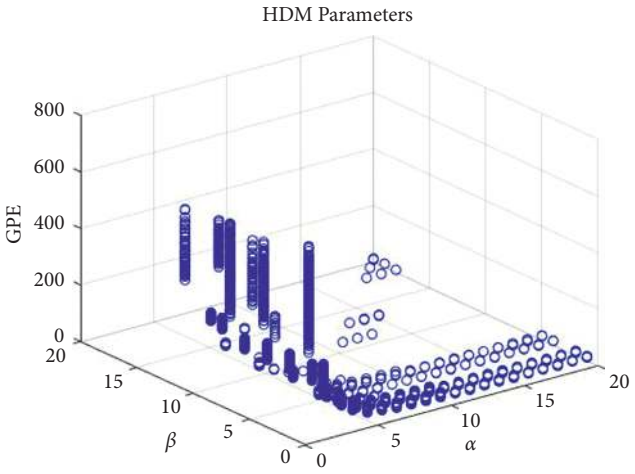

Figure 4: GPE vs. α and β.

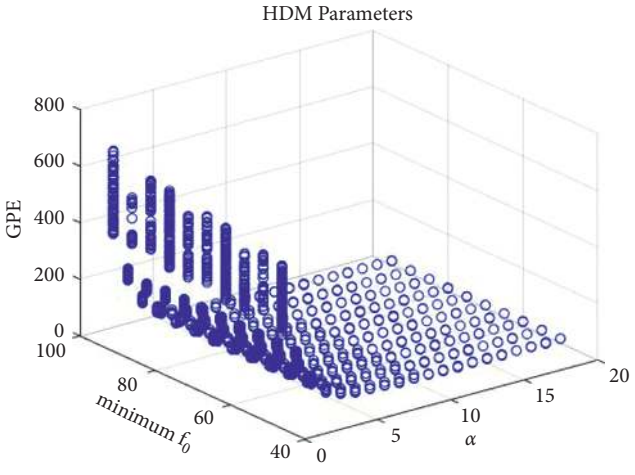

Figure 5: GPE vs. α and minimum f_0.

223.878 seconds and length of files ranges from 2049 to 6872 with mean 2147.515 and standard deviation 489.218. Texas dataset is a collection of 3314 vowels with kid, female, and male vowels. Total length is 698.4711 seconds and length of files ranges from 1110 to 24257 with mean 3372.22 and standard deviation 1010.025. TIMIT is a large collection of words and phones. Usually, it is divided into training, development, and test sets. In this work, we use all vowels, including SA samples. SA samples are not used in classification applications due to overfitting concerns; however, this is not relevant in our case. TIMIT vowel set comprises 78374 samples, including 24017 female and 54357 male voices. Total length is 7521.265 seconds, length of files ranges from 74 to 7735 with mean 1535.461, and standard deviation of 778.7452. Detailed information is shown in Table 1.

In Hillenbrand and Texas datasets, there are 12 American English vowel classes /*i*/ (heed), / / (hid), /ɛ/ (head), /æ/ (had), /*e*/ (hayed), / / (hod), / / (hawed), /*o*/ (hoed), / / (hud), / / (hood), /*u*/ (who'd), / / (heard). Vowels are obtained from the /hVd/ syllable context. In TIMIT dataset, 20 vowels are included: /*i*/ (beet), / / (bit), /ɛ/ (bet), /æ/ (had), /*e*/ (hayed), / / (bott), / / (bought), /*o*/ (boat), / / (but), / / (hood), /*u*/ (boot), / / (herd) with additional /a / (bite), / / (about), / / (boy), [ü] (toot), [ ] (debit), [^{h}] (suspect), [a^{w}] (bout), and [ ] (butter). TIMIT vowel classes can be reduced to 14 from 20 by eliminating the stress and semivowel intonations [137, 196, 197].

In Hillenbrand dataset, fundamental frequency ground truth is calculated using autocorrelation followed by hand-editing. In Texas Vowel dataset, fundamental frequency ground truth is calculated by visual inspection together with semiautomatic LPC analysis. In the TIMIT corpus, transcriptions have been hand-verified. Transcriptions are obtained using the program SPIRE of MIT and then hand-verified by experienced acoustics phoneticians. But there is no fundamental frequency ground truth for TIMIT vowels. Therefore, we used average f0 values found by HDM, autocorrelation, cepstrum, YAAPT, CREPE, and FCN methods in a consistent manner in conjunction with the gender labels provided by the TIMIT dataset. In some samples of TIMIT dataset, finding the ground truth is very difficult even by visual expert inspection on the frequency spectrum. The harmonics and spacing between them can be spread nearly randomly. Such a sample from TIMIT dataset is depicted in Figure 6.

Autocorrelation of a signal with the symmetry property can be obtained using the following equation:

$$A_{ac}(k) = \sum_{i=k}^{N-1} x_i x_{i-k}, \tag{7}$$

where k is the lag number.

Cepstrum has the complexity of $O(N \log N)$, and power cepstrum can be calculated using the following formula:

$$C_p = \left|\mathcal{F}^{-1}\left\{\log\left(\left|\mathcal{F}\{f(t)\}\right|^2\right)\right\}\right|^2. \tag{8}$$

As can be seen from Table 2, HDM is the fastest method followed by cepstrum and autocorrelation. For autocorrelation and cepstrum, we used Naotoshi SEO's (http://note.sonots.com/SciSoftware/Pitch.html) implementations, but

Table 1: Datasets used in this work.

	Boy	Girl	Male	Female
Hillenbrand	324	228	540	576
Texas	1232		972	1110
TIMIT			54357	24017

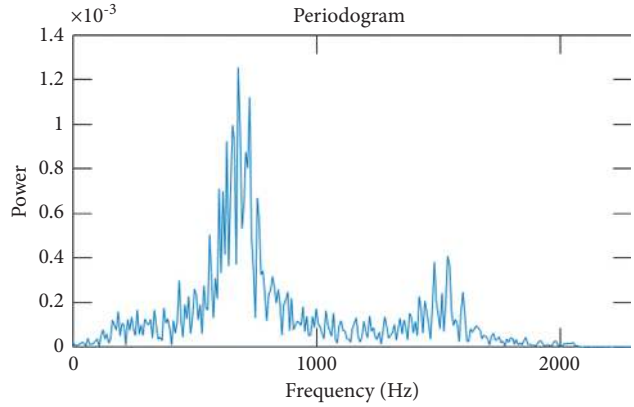

Figure 6: The near random distribution of harmonics and the spacing between them can make extraction of fundamental frequency too complicated.

Table 2: Average speeds of used methods on the Hillenbrand dataset in microseconds.

HDM (2021)	AC (2008)	CEPS (2008)	YIN (2002)	YAAPT (2016)	CREPE (2018)	FCN (2019)
108	192	132	14544	8327	263965	82522

in cepstrum, we changed the maximum pitch value as 500 Hz to obtain better results. Timings are average values of 10 consecutive runs on Hillenbrand dataset for a single speech sample. YIN, YAAPT (version 2016), FCN (2019), and especially CREPE (2018) are too far away from HDM, cepstrum, and autocorrelation in terms of speed. YIN [66], YAAPT [67, 68], and FCN [113] employ temporal smoothing, whereas HDM, autocorrelation, cepstrum, and CREPE [110] do not use temporal smoothing. HDM produces nearly identical results without hamming windowing, but the rest of the algorithms have worse results without hamming window, particularly, cepstrum doubles the error margin in TIMIT dataset. No other preprocessing was applied to the waveforms. In cepstrum and autocorrelation, we imposed an upper-limit frequency of 500 Hz. Without this limit, their performances are going only worse. This makes them unsuitable for the detection of high pitch values. In angry emotional speech samples, it is possible to see up to 700 Hz pitch values.

5. Results

Many error measures are used to evaluate the pitch detection algorithms. Gross Pitch Error (GPE) is the average error, voicing detection error (VDE) if applicable are among the most used. Additionally, we introduce two different error measures. The first is the e_{10}, which denotes the number of samples with more than 10% sway from the ground truth value. Another useful application of pitch estimation is the gender detection. Gender detection error can be very useful for the evaluation of these algorithms. GPE and e_{10} error are, respectively, defined as

$$\begin{aligned} \text{GPE} &= \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{f}_i - f_i\right|, \\ e_{10} &= \#\left\{i \mid i = 1, \ldots, N, \left|\widehat{f}_i - f_i\right| > \frac{f_i}{10}\right\}. \end{aligned} \tag{9}$$

In Hillenbrand and Texas Vowel datasets, ground truth pitch values are given, and we can make a solid comparison with our predictions. In Figure 7, such a comparison is depicted for Hillenbrand dataset. Boy, girl, man, and woman samples can clearly be seen to make an intuition over the frequency regions of these samples. Male voices have definitely lower f0 values compared with the boys, girls, and woman. It is nearly impossible to separate boy, girl, and woman voices using only f0 values. In Hillenbrand dataset, average f0 is 236.0 Hz for boys, 238.35 Hz for girls, 131.21 Hz for man, and 220.40 Hz for woman. Maximum f0 for boy, girl, man, and woman are 320, 303, 224, and 307 Hz, respectively. Minimum f0 for boy, girl, man, and woman are 183, 188, 90, and 149 Hz, respectively. There is no boy or girl sample with f0 value of lower than 180 Hz. Only 7 males have pitch values of greater than 190, and only 5 women have pitch values of lower than 160 Hz. In Texas dataset, there are kid, man, and woman classes. In Texas dataset, the difference

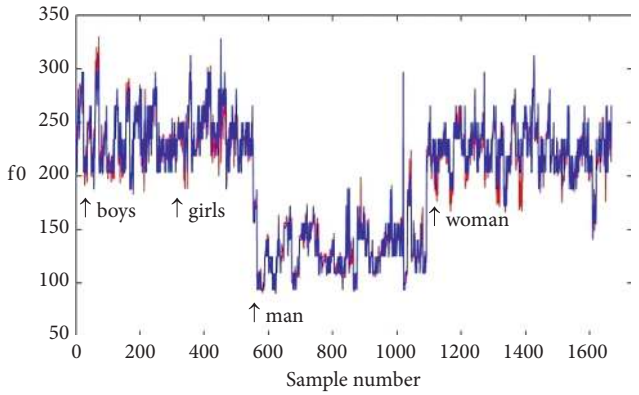

Figure 7: $f0$ ground truth (red) and HDM estimations (blue) in Hillenbrand Vowel dataset.

between woman and kid classes is clearer than Hillenbrand dataset. In Texas dataset, we have 3-, 5-, and 7-year-old kid samples. This suggests that as age decreases, pitch values tend to be higher. Maximum f0 for kid, man, and woman samples are 392, 202, and 271 Hz, respectively. Minimum f0 for kid, man, and woman samples are 105 (sample k7bree05, this is a strange value for a kid so that we felt to check it manually), 82, and 141 Hz, respectively. In Texas Vowel dataset, only 1 kid sample has f0 value of lower than 180 Hz (abovementioned sample k7bree05), and only 3 males have f0 value of greater than 140 Hz. Only 7 woman samples have f0 value of lower than 160 Hz. In TIMIT dataset, we have only male and female classes, and average values conform well to the other two datasets as shown in Table 3. In TIMIT dataset, there are no supplied ground truth pitch values.

As seen from Table 4, in wideband Hillenbrand dataset, the proposed HDM has the smallest GPE, and AC has the smallest e_{10} error. Although autocorrelation is an old technique, it is highly successful in this dataset. In Texas dataset, FCN has the smallest GPE error, and AC has the smallest e_{10} error. In TIMIT dataset, FCN is the best performing method followed closely by HDM and cepstrum in GPE and the proposed HDM is the best method in e_{10} error. YIN produces too many outliers, and for YIN, we tried many different parameters to find better results. A box plot of the abovementioned algorithms is depicted in Figure 8. We need to emphasize that in the AC and cepstrum implementations, we imposed an upper-limit frequency of 500 Hz; otherwise, these methods produce worse results. No upper limit is applied in the remaining methods. Convolutional neural network methods are quite successful in wideband implementations, but as we will see right now, they are nearly blind in narrowband telephone speech data.

From now on, we extend our experiments to the telephony speech. For this purpose, we will apply band-pass filter to our datasets twice to completely remove the frequencies below 400 Hz and above 3400 Hz. In some telephony speech, this bandwidth can be applied between 300 Hz and 4000 Hz. We selected the low frequency as 400 Hz because the highest f0 value is 392 Hz in our datasets, and by selecting 400 Hz as threshold value, we remove the

Table 3: Average $f0$ values for Hillenbrand, Texas, and TIMIT datasets by gender.

	Boy	Girl	Man	Woman
Hillenbrand	236	238.35	131.21	220.40
Texas		245.76	110.86	217.23
TIMIT			119.07	207.21

Table 4: Experimental results on wideband Hillenbrand, Texas, and TIMIT datasets.

	Hillenbrand		Texas		TIMIT	
	GPE	e_{10}	GPE	e_{10}	GPE	e_{10}
HDM	**7.65**	6.41	9.46	11.41	6.39	**4.68**
AC	8.17	**4.68**	10.36	**7.09**	22.74	8.79
CEPS	7.86	6.00	14.53	11.38	11.27	7.45
YIN	16.71	8.63	16.67	8.90	43.88	12.58
YAAPT	17.18	13.55	12.87	14.76	16.00	15.13
CREPE	8.97	9.53	8.45	8.12	12.33	6.84
FCN	9.10	8.75	**8.17**	7.42	**5.82**	5.29

Bold values denote the best performance in the specified dataset and error type.

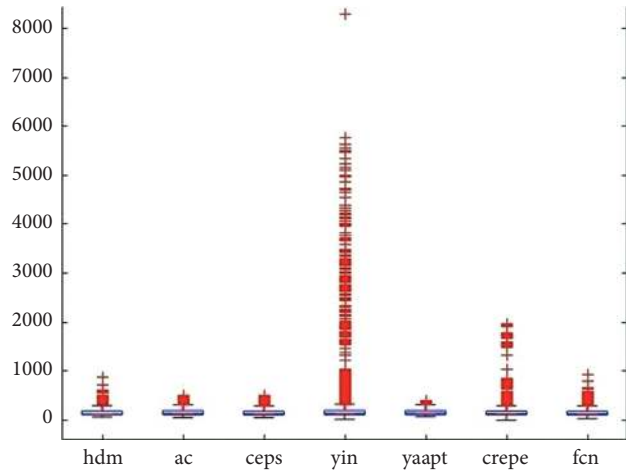

Figure 8: Box plot of all algorithms for wideband TIMIT dataset.

fundamental frequency from all samples in all datasets. This is one of the objectives of this algorithm.

In band-passed Hillenbrand dataset, cepstrum remarkably is the most successful algorithm in all error types as shown in Table 5. HDM is the second in GPE, and AC is the second in e_{10} error measures. Performance of CREPE and FCN is very disappointing in band-passed speech. In narrowband Texas dataset, cepstrum is the most successful algorithm in all error types as shown in Table 5. HDM is the second in GPE, and AC is the second in e_{10} error type.

In narrowband TIMIT dataset, our novel HDM algorithm is superior to all other methods in GPE and e_{10} error measures as seen in Table 5. YAAPT is the second best in GPE, and AC is the second in e_{10} error. YAAPT is primarily designed for telephone speech. In TIMIT dataset, 22670 samples are shorter than 1024 in length. This may explain the failure of cepstrum in narrowband TIMIT dataset. To further clarify the underlying essence of this failure, we removed the samples that are shorter than 1024 and rerun cepstrum on

Table 5: Experimental results on narrowband Hillenbrand, Texas, and TIMIT datasets.

	Hillenbrand		Texas		TIMIT	
	GPE	e_{10}	GPE	e_{10}	GPE	e_{10}
HDM	34.72	19.30	39.42	29.06	**15.26**	**17.88**
AC	37.33	15.53	55.71	22.24	69.47	23.30
CEPS	**11.23**	**7.37**	**37.89**	**17.50**	75.17	25.36
YIN	72.64	25.12	130.29	34.85	160.97	40.83
YAAPT	51.93	50.18	45.83	42.18	27.10	28.14
CREPE	111.52	42.51	134.87	46.11	162.44	50.66
FCN	185.53	70.08	191.81	71.73	194.79	67.28

Bold values denote the best performance in the specified dataset and error type.

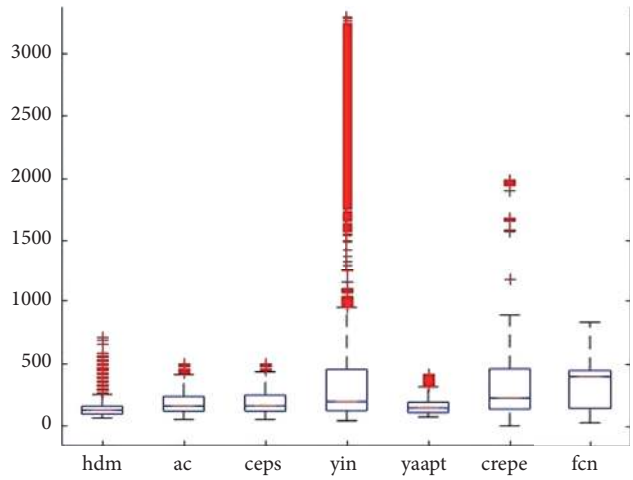

Figure 9: Box plot of all algorithms for narrowband TIMIT dataset.

Table 6: Gender detection results on wideband TIMIT dataset.

	Male incorrect	Female incorrect	Error (%)
HDM	832	1476	**2.94**
AC	5331	2403	9.86
CEPS	2999	3219	7.93
YIN	8136	2674	13.79
YAAPT	5475	2483	10.15
CREPE	3435	2295	7.31
FCN	1302	1724	3.86

Bold value denotes the best performance.

Table 7: Gender detection results on narrowband TIMIT dataset.

	Male incorrect	Female incorrect	Error (%)
HDM	363	6561	**8.83**
AC	15091	3919	24.25
CEPS	16258	4248	26.16
YIN	24393	7359	40.51
YAAPT	6129	5124	14.35
CREPE	29155	10331	50.38
FCN	34398	17717	66.49

Bold value denotes the best performance.

narrowband TIMIT data. In this scheme, cepstrum GPE error dropped to 49.85 from 75.17 and e_{10} error to 17.41 from 25.36. Arguably, we can conclude that zero padding can severely affect the performance of the cepstrum. Box plot of these algorithms for narrowband TIMIT dataset are shown in Figure 9. CREPE and FCN are nearly useless in narrowband speech. This may be due to the fact that their training is done only on the wideband speech. They need to be trained for narrowband speech as well. Convolutional neural networks are quite successful in wideband speech data, however, we must keep in our mind that they are extremely slow compared with HDM, AC, and cepstrum as shown in Table 2. HDM is the fastest method in these experiments.

6. Gender Detection Implementations

Gender detection is an application of fundamental frequency detection. Although gender is not restricted to pitch value, it is highly related to its value. Pitch has specific ranges between men, women, boys, and girls. Therefore, gender evaluation of the $f0$ algorithms is a good measure for the robustness. Here, we present the gender detection errors for wideband and band-passed TIMIT dataset. In TIMIT dataset, the gender information is given with the first letter of the name of the speech sample.

As seen from Table 6, in wideband TIMIT dataset, HDM is the best method in gender detection by a significant margin, FCN is the second, and cepstrum is the third. Cepstrum's success is well known in male speech samples. In the TIMIT dataset, there are 24017 female and 54357 male samples. This may explain the success of cepstrum in this large dataset. In Table 7, we can clearly conclude that HDM has no match specifically in male samples.

7. Conclusions

The experimental results show that proposed harmonic differences can safely be used to detect fundamental frequency in wideband and narrowband telephony speech. The new algorithm shows great success particularly in the large TIMIT dataset. Fast Fourier Transform has a natural resolution problem, but in this article, despite the low resolution of the implementation, the results are satisfactory. It is robust to band-limiting and moderate inharmonicity. HDM algorithm is the fastest method, and further speed improvements can be expected. FCN and CREPE are performing remarkably well in wideband data, but they are too slow compared with the other methods. Therefore, they cannot be used for real-time applications, but they can be helpful in ground truth determination. An interesting

finding is the highly disappointing results of the FCN and CREPE algorithms in narrowband speech. Although they are quite successful in wideband speech datasets, they produced low success rates in all band-passed datasets. We should bear in mind that FCN and CREPE are end-to-end algorithms, and they take the raw waveform as input without using the frequency-domain descriptors. Most of the useful pitch information is hidden in the low part (0–400 Hz) of band-passed signal and without this data, FCN and CREPE are unable to extract the necessary features for pitch determination. The cepstrum algorithm is very old compared to YIN, YAAPT, CREPE, and FCN, but in some cases, it can present better predictions. FCN and CREPE are CNN-based methods; FCN is using temporal smoothing, whereas CREPE does not use temporal smoothing but FCN is still much faster than CREPE.

In the future works, we plan to implement temporal smoothing in HDM. Temporal smoothing can be quite efficient in f0 detection and is used by many algorithms, including YIN, YAAPT, and FCN. Another future direction is testing the ability of HDM in noisy environments and musical sounds that needs to be handled. Pitch refinement is another technique that can be incorporated inside HDM.

Data Availability

Some of the data are available in the following sites: https://github.com/cevparlak/f0-detection; TIMIT dataset: https://catalog.ldc.upenn.edu/LDC93S1; Hillenbrand Vowel dataset: http://homepages.wmich.edu/~hillenbr/voweldata.html; and Texas Vowel Dataset: https://personal.utdallas.edu/~assmann/KIDVOW1/North_Texas_vowel_database.html. Other data can be accessed from the corresponding author via e-mail.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] J. K. Bates, "Edited selections from J. F. Schouten's "the residue revisited"," 2007, https://web.archive.org/web/20070927224928/http://home.computer.net/~jkbates/schoutpd.pdf.

[2] R. S. Turner, "The Ohm-Seebeck dispute, Hermann von Helmholtz, and the origins of physiological acoustics," *British Journal for the History of Science*, vol. 10, no. 1, pp. 1–24, 1977.

[3] A. Seebeck, "Beobachtungen über einige bedingungen der entstehung von tönen," *Annalen der Physik und Chemie*, vol. 53, pp. 417–436, 1841.

[4] G. S. Ohm, "Über die definition des tones, nebst daran geknüpfter theorie der sirene und ähnlicher tonbildender vorrichtungen," *Annalen der Physik und Chemie*, vol. 59, pp. 513–565, 1843.

[5] H. L. F. Von Helmholtz, *Die Lehre von den Tonempfindungen als Physiologische Grundlage fu r die Theorie der Musik*, Vieweg & Sohn, Braunschweig, Germany, 1863.

[6] L. Rayleigh, *Theory of Sound*, Macmillan and Co., London, UK, 1st edition, 1877.

[7] F. L. Wightman and D. M. Green, "The perception of pitch: the pitch of a sound wave is closely related to its frequency or periodicity-but the exact nature of that relation remains a mystery," *American Scientist*, vol. 62, no. 2, pp. 208–215, 1974.

[8] J. F. Schouten, "The perception of subjective tones," *Proceedings of the Koninklijke Nederlandse Akademie van Wetenscbappen*, vol. 41, pp. 1083–1093, 1938.

[9] E. Terhardt, "Zur tonhöhenwahrnehmung von klängen I. Psychoakustische grundlagen," *Acoustica*, vol. 26, pp. 173–186, 1972.

[10] E. Terhardt, "Pitch, consonance and harmony," *Journal of the Acoustical Society of America*, vol. 55, pp. 1061–1069, 1974.

[11] J. F. Schouten, "The residue: a new component in subjective sound analysis," *Proceedings of Koniklijke Nederlandsche Akademie van Wetenschappen*, vol. 43, no. 3, pp. 356–365, 1940.

[12] J. Q. Feng, "Music in terms of science," 2012, https://arxiv.org/pdf/1209.3767.pdf.

[13] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–134, 1951.

[14] J. C. R. Licklider, "Auditory frequency analysis," in *Information Theory*, C. Cherry, Ed., pp. 253–268, Butterworth, London, UK, 1956.

[15] J. C. R. Licklider, "Three auditory theories," in *Psychology: A Study of Science*, S. Koch, Ed., Vol. 1, Mc Graw-Hill, New York, NY, USA, 1959.

[16] J. C. R. Licklider, "Periodicity pitch and related auditory process models," *International Audiology*, vol. 1, pp. 11–36, 1962.

[17] R. J. Ritsma, "Existence region of the tonal residue, I," *Journal of the Acoustical Society of America*, vol. 34, pp. 1224–1229, 1962.

[18] R. L. Miller and E. S. Weibel, "Measurements of the fundamental period of speech using a delay line," *Journal of the Acoustical Society of America*, vol. 28, 1956.

[19] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-22, pp. 353–362, 1974.

[20] C. K. Un and S. C. Yang, "A pitch extraction algorithm based on LPC inverse filtering and AMDF," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-25, pp. 565–572, 1977.

[21] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, pp. 2–8, 1976.

[22] M. Staudacher, V. Steixner, A. Griessner, and C. Zierhofer, "Fast fundamental frequency determination via adaptive autocorrelation," *EURASIP Journal on Audio Speech and Music Processing*, vol. 2016, no. 1, pp. 1–8, 2016.

[23] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-25, no. 1, pp. 24–33, 1977.

[24] Q. Lin and Y. Shao, "A novel normalization method for autocorrelation function for pitch detection and for speech activity detection," in *Proceedings of the 2018 Interspeech*, pp. 2097–2101, Hyderabad, India, 2018.

[25] S. Kumar, "Performance measurement of a novel pitch detection scheme based on weighted autocorrelation for speech signals," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 885–892, 2019.

[26] B. Bogert, J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the 1963 Symposium on Time Series Analysis*, Wiley & Sons, Providence, RI, USA, 1963.

[27] M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *Journal of the Acoustical Society of America*, vol. 36, pp. 296–302, 1964.

[28] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, pp. 293–309, 1967.

[29] M. R. Schroeder, "Period histogram and product spectrum: new methods for fundamental frequency measurement," *Journal of the Acoustical Society of America*, vol. 43, pp. 829–834, 1968.

[30] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," in *Proceedings of the Symposium on Computer Processing Communications*, Las Vegas, NV, USA, 1969.

[31] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *Journal of the Acoustical Society of America*, vol. 47, pp. 634–648, 1970.

[32] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *Journal of the Acoustical Society of America*, vol. 46, pp. 442–448, 1969.

[33] R. Sukkar, "*A parallel processing pitch detector for LPC*," M.S. thesis, Illinois Institute of Technology, Chicago, IL, USA, 1985.

[34] R. A. Sukkar, J. L. LoCicero, and J. W. Picone, "Design and implementation of a robust pitch detector based on a parallel processing technique," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 441–451, 1988.

[35] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-20, pp. 367–377, 1972.

[36] J. A. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 330–338, 1974.

[37] N. J. Miller, "Pitch detection by data reduction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-23, pp. 72–79, 1975.

[38] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," *Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1568–1580, 1982.

[39] S. Seneff, "Real-time harmonic pitch detector," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 358–365, 1978.

[40] L. Wu, *Guitar Sound Analysis and Pitch Detection*, Stanford University, Stanford, CA, USA, 2017.

[41] A. D. Shapiro and C. Wang, "A versatile pitch tracking algorithm: from human speech to killer whale vocalizations," *Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 451–459, 2009.

[42] M. S. Phillips, "A feature-based time domain pitch tracker," *Journal of the Acoustical Society of America*, vol. 77, no. S1, pp. S9–S10, 1985.

[43] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proceedings of the 1983 IEEE Conference on Acoustics, Speech, and Signal Processing*, Boston, MA, USA, April 1983.

[44] R. J. Sluyter, H. J. Kotmans, and A. V. Leeuwarden, "A novel method for pitch extraction from speech and a hardware model applicable to vocoder systems," in *Proceedings of the 1980 IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 80, pp. 45–48, Denver, CO, USA, 1980.

[45] W. H. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer & Verlag, Berlin, Germany, 1983.

[46] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Vol. 3, Springer Science & Business Media, Berlin, Germany, 2012.

[47] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, pp. 257–264, 1988.

[48] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification," *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2881, 1991.

[49] S. Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, pp. 40–48, 1991.

[50] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[51] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, 1986.

[52] H. Kawahara, "STRAIGHT-TEMPO: a universal tool to manipulate linguistic and para-linguistic speech information," in *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 2, pp. 1620–1625, Orlando, FL, USA, 1997.

[53] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 495–518, Elsevier, Amsterdam, Netherlands, 1995.

[54] T. Abe, T. Kobayashi, and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," in *Proceedings of 4th International Conference on Spoken Language Processing ICSLP'96*, vol. 2, pp. 1277–1280, Philadelphia, PA, USA, 1996.

[55] H. Kawahara, H. Katayose, A. D. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999.

[56] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[57] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[58] H. Kawahara, J. Estill, and O. Fujimura, "A periodicity extraction and control using mixed mode excitation and group delay manipulation for a high-quality speech analysis, modification and synthesis system STRAIGHT," in *Proceedings of the 2nd International Workshop on Models and*

Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), pp. 59–64, Florence, Italy, 2001.

[59] H. Kawahara, A. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight," in *Proceedings of the Interspeech 2005*, pp. 537–540, Lisbon, Portugal, 2005.

[60] C. Wendt and A. P. Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform," in *Proceedings of the 1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World ISCAS 96*, vol. 2, pp. 45–48, Atlanta, GA, USA, 1996.

[61] J. P. de León and J. R. Beltrán, "A complex wavelet based fundamental frequency estimator in single channel polyphonic signals," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, 2013.

[62] Y. H. Goh, Y. H. Ko, Y. K. Lee, and Y. L. Goh, "Fast wavelet-based pitch period detector for speech signals," in *Proceedings of the 2016 International Conference on Computer Engineering and Information Systems*, Shanghai, China, 2016.

[63] D. J. Liu and C. T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 609–621, 2001.

[64] P. Boersma, "Praat: a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2002.

[65] S. Xuejing, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 13–17, Orlando, FL, USA, May 2002.

[66] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[67] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Orlando, FL, USA, 2002.

[68] S. A. Zahorian, P. Dikshit, and H. Hu, "A spectral-temporal method for pitch tracking," in *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, 2006.

[69] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, 2003.

[70] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[71] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2011.

[72] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, 2008.

[73] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[74] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, ""Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[75] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 138–150, 2014.

[76] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[77] L. Su and Y.-H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, p. 10, 2015.

[78] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Processing*, vol. 127, pp. 56–70, 2016.

[79] M. Képesi and L. Weruaga, "High-resolution noise-robust spectral-based pitch estimation," in *Proceedings of the 2005 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005.

[80] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[81] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proceedings of the ICASSP 2008*, pp. 3933–3936, Las Vegas, NV, USA, 2008.

[82] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[83] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Proceedings of the AES 35th International Conference: Audio for Games*, London, UK, 2009.

[84] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "A new algorithm for instantaneous F0 speech extraction based on ensemble empirical mode decomposition," in *Proceedings of the 2009 17th European Signal Processing Conference*, pp. 2347–2351, Glasgow, UK, 2009.

[85] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies," in *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, pp. 984–987, Munich, Germany, 2009.

[86] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio Speech and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[87] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[88] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proceedings of the 2011 Interspeech*, Florence, Italy, 2011.

[89] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Transactions on Audio*

Speech and Language Processing, vol. 20, no. 3, pp. 994–1006, 2011.

[90] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, 2006.

[91] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: from analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[92] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proceedings of the 2011 19th European Signal Processing Conference*, pp. 451–455, Barcelona, Spain, 2011.

[93] W. Chu and A. Alwan, "SAFE: a statistical approach to F0 estimation under clean and noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 933–944, 2011.

[94] B. S. Lee and D. P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proceedings of the 2012 INTERSPEECH*, Portland, OR, USA, 2012.

[95] H. Ba, N. Yang, I. Demirkol, and W. Heinzelman, "BaNa: a hybrid approach for noise resilient pitch detection," in *Proceedings of the 2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 369–372, Ann Arbor, MI, USA, 2012.

[96] H. Yedla, R. R. Kishore, and M. N. Yadav, "Hybrid high noise resiliency pitch detection algoritm," *International Journal of Current Engineering and Scientific Research (IJCESR)*, vol. 3, pp. 1–4, 2015.

[97] V. Morfi, G. Degottex, and A. Mouchtaris, "A computationally efficient refinement of the fundamental frequency estimate for the adaptive harmonic model," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1478–1482, Florence, Italy, 2014.

[98] P. McLeod and G. Wyvill, "A smarter way to find pitch," in *Proceedings of International Computer Music Conference ICMC 2005*, Barcelona, Spain, 2005.

[99] J. Fang, F. Wang, Y. Shen, H. Li, and R. S. Blum, "Super-resolution compressed sensing for line spectral Estimation: an iterative reweighted approach," *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4649–4662, 2016.

[100] S. Stone, P. Steiner, and P. Birkholz, "A time-warping pitch tracking algorithm considering fast F0 changes," in *Proceedings of the 2017 INTERSPEECH*, Stockholm, Sweden, 2017.

[101] D. Wang, C. Yu, and J. H. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 952–964, 2017.

[102] M. Morise, "Harvest: a high-performance fundamental frequency estimator from speech signals," in *Proceedings of the 2017 INTERSPEECH*, Stockholm, Sweden, 2017.

[103] M. Mauch and S. Dixon, "PYIN: a fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 659–663, Florence, Italy, 2014.

[104] J. Swärd, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 296–303, 2017.

[105] S. R. Kadiri and B. Yegnanarayana, "Estimation of fundamental frequency from singing voice using harmonics of impulse-like excitation source," in *Proceedings of the 2018 INTERSPEECH*, Hyderabad, India, 2018.

[106] E. Loweimi, J. Barker, and T. Hain, "On the usefulness of the speech phase spectrum for pitch extraction," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018, pp. 696–700, Hyderabad, India, 2018.

[107] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, "Traditional machine learning for pitch detection," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1745–1749, 2018.

[108] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *Proceedings of the 2016 INTERSPEECH*, San Francisco, CA, USA, 2016.

[109] 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*, vol. 25, no. 4, pp. 829–838, 2017.

[110] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: a convolutional representation for pitch estimation," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

[111] L. Watts, "DeepPitch: wide-range monophonic pitch estimation using deep convolutional neural networks," 2018, http://www.lloydwatts.com/images/2018-06-14_MonophonicPitchPaper.pdf.

[112] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 estimation for DNN-based ultrasound silent speech interfaces," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

[113] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *Proceedings of the Insterspeech 2019*, Graz, Austria, 2019.

[114] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017.

[115] A. Elowsson, "Polyphonic pitch tracking with deep layered learning," *Journal of the Acoustical Society of America*, vol. 148, no. 1, pp. 446–468, 2020.

[116] A. Elowsson and A. Friberg, "Modeling music modality with a key-class invariant pitch chroma CNN," 2019, https://arxiv.org/abs/1906.07145.

[117] C. Hernandez-Olivan, I. Z. Pinilla, C. Hernandez-Lopez, and J. R. Beltran, "A comparison of deep learning methods for timbre analysis in polyphonic automatic music transcription," *Electronics*, vol. 10, no. 7, p. 810, 2021.

[118] A. Paeschke, M. Kienast, and W. F. Sendlmeier, "F0-contours in emotional speech," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 2, pp. 929–932, San Francisco, CA, USA, 1999.

[119] K. Hirose, K. Sato, and N. Minematsu, "Emotional speech synthesis with corpus-based generation of F0 contours using generation process model," in *Proceedings of the International Conference on Speech Prosody 2004*, Nara, Japan, 2004.

[120] J. James, H. Mixdorff, and C. I. Watson, "Quantitative model-based analysis of f0 contours of emotional speech," in *Proceedings of the 2019 International Congress of Phonetic Sciences*, Melbourne, Australia, 2019.

[121] M. M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, pp. 262–266, 1968.

[122] L. Rabiner, M. Cheng, A. Rosenberg, and C. Mc Gonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[123] K. Oh and C. Un, "A performance comparison of pitch extraction algorithms for noisy speech," in *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, pp. 85–88, Los Alamitos, CA, USA, 1984.

[124] W. Hess, S. Furui, and M. Sondhi, "Pitch and voicing determination," *Advances in Speech Signal Processing*, pp. 3–48, Taylor & Francis, New York, NY, USA, 1992.

[125] W. J. Hess, "Pitch and voicing determination of speech with an extension toward music signals. Springer handbook of speech processing, 181–212. 2012 version Hess, W. (2012)," *Pitch Determination of Speech Signals: Algorithms and Devices*, vol. 3, Springer Science & Business Media, Berlin, Germany, 2008.

[126] P. C. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Doctoral dissertation, University of Edinburgh, Edinburgh, UK, 1994.

[127] D. Gerhard, *Pitch Extraction and Fundamental Frequency: History and Current Techniques*, Department of Computer Science, University of Regina, Regina, Canada, 2003.

[128] Y. B. Wang, *Development of a Real Time Hearing Enhancement Algorithm for Crowded Social Environments*, University of Toronto, Toronto, Canada, 2014.

[129] S. Strömbergsson, "Today's most frequently used F0 estimation methods, and their accuracy in estimating male and female pitch in clean speech," in *Proceedings of the 2016 INTERSPEECH*, San Francisco, CA, USA, 2016.

[130] D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO)*, 2017.

[131] B. Wang, "Development of a real time hearing enhancement algorithm for crowded social environments," Doctoral dissertation, University of Toronto, Toronto, Canada, 2014.

[132] E. Verteletskaya and B. Šimák, *Performance Evaluation of Pitch Detection Algorithms*, 2009.

[133] M. Szczerba and A. Czyzewski, "Pitch detection enhancement employing music prediction," *Journal of Intelligent Information Systems*, vol. 24, no. 2-3, pp. 223–251, 2005.

[134] V. P. Karunaimathi, D. Gladis, and D. Balakrishnan, "An analogy of F0 estimation algorithms using sustained vowel," in *Proceedings of the 3rd International Symposium on Women in Computing and Informatics*, Kochi, India, 2015.

[135] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 2011.

[136] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid, Spain, 1995.

[137] J. S. Garofolo, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download*, Linguistic Data Consortium, Philadelphia, PA, USA, 1993.

[138] W. Fisher, G. Doddington, K. Goudie-Marshall et al., "NTIMIT," 1993, https://hdl.handle.net/11272.1/AB2/AXQJUZ.

[139] J. Yamagishi, C. Veaux, and K. MacDonald, *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit*, University of Edinburgh, Edinburgh, UK, 2019.

[140] M. Bijankhan, J. SheikHzadegan, and M. R. Roohani, "FARSDAT-The speech database of Farsi spoken language," in *Proceedings of the 1994 Australian Conference on Speech Science and Technology*, Perth, Australia, 1994.

[141] A. Wrench, "The MOCHA-TIMIT articulatory database," 1999, http://www.cstr.ed.ac.uk/articmocha.html.

[142] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: popular, classical and jazz music databases," in *Proceedings of the 2002 ISMIR*, vol. 2, pp. 287-288, Paris, France, 2002.

[143] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: a multitrack dataset for annotation-intensive MIR research," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, 2014.

[144] H. L. Storkel, "A corpus of consonant-vowel-consonant real words and nonwords: comparison of phonotactic probability, neighborhood density, and consonant age of acquisition," *Behavior Research Methods*, vol. 45, no. 4, pp. 1159–1167, 2013.

[145] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[146] D. Iskra, B. Grosskopf, K. Marasek, H. Heuvel, F. Diehl, and A. Kiessling, "Speecon—speech databases for consumer devices: database specification and validation," in *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Spain, 2002.

[147] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, Pittsburgh, PA, USA, 2004.

[148] G. Lindsey, A. Breen, and S. Nevard, *SPAR'S Archivable Actual-Word Databases*, University College London, London, UK, 1987.

[149] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and nonpeak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[150] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland music data (SMD)," in *Proceedings of the 2011 International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA, 2011.

[151] C. Sapp, "The mazurka project," 2017, http://www.mazurka.org.uk.

[152] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunningham, and Y. Hao, "Ten years of MIREX: reflections, challenges and opportunities," in *Proceedings of the ISMIR 2014*, pp. 657–662, Taipei, Taiwan, 2014.

[153] Rice University Digital Signal (DSP) Group, "Noisex92 noise database," 1995, http://spib.linse.ufsc.br/noise.html.

[154] E. R. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: a tutorial," *Clinical Linguistics & Phonetics*, vol. 3, no. 3, pp. 281–296, 1989.

[155] B. Lindblom and J. Sundberg, "The human voice in speech and singing," in *Springer Handbook of Acoustics*, pp. 703–746, Springer, New York, NY, USA, 2014.

[156] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin, Germany, 2007.

[157] C. Mooshammer, "Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal

effort in German," *Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1047–1058, 2010.

[158] M. R. Thomas and P. A. Naylor, "The SIGMA algorithm for estimation of reference-quality glottal closure instants from electroglottograph signals," in *Proceedings of the 2008 16th European Signal Processing Conference*, pp. 1–5, Lausanne, Switzerland, 2008.

[159] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3969–3972, Taipei, Taiwan, 2009.

[160] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7815–7819, Vancouver, Canada, 2013.

[161] Recommendation, I. T. U. T., "ITU-T P. 800.1," Mean Opinion Score (MOS) Terminology (2003).

[162] Recommendation ITU-T. P.807, *Subjective Test Methodology for Assessing Speech Intelligibility*, ITU, Geneva, Switzerland, 2016.

[163] Recommendation ITU-T. P.862, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU, Geneva, Switzerland, 2001.

[164] Recommendation ITU-T. P.863.1, *Methods for Objective and Subjective Assessment of Speech and Video Quality*, ITU, Geneva, Switzerland, 2019.

[165] P. Flandrin, F. Auger, and E. Chassande-Mottin, "Time-frequency reassignment: from principles to algorithms," *Applications in Time-Frequency Signal Processing*, vol. 5, p. 102, 2003.

[166] V. Bruni, M. Tartaglione, and D. Vitulano, "A fast and robust spectrogram reassignment method," *Mathematics*, vol. 7, no. 4, p. 358, 2019.

[167] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[168] K. Duda and S. M. Salih, "Interpolation algorithms of DFT for parameters estimation of sinusoidal and damped sinusoidal signals," in *Fourier Transform-Signal Processing*, vol. 32, p. 3, InTechOpen, London, UK, 2012.

[169] E. Jacobsen, "On local interpolation of DFT outputs," 1994, http://www.ericjacobsen.org/FTinterp.pdf.

[170] X. Wu and A. Wang, "Harmonic signal processing method based on the windowing interpolated DFT algorithm," *Journal of Information Science and Engineering*, vol. 31, no. 3, pp. 787–798, 2015.

[171] M. Gasior and J. L. Gonzalez, *Improving FFT Frequency Measurement Resolution by Parabolic and Gaussian Interpolation (No. AB-Note-2004-021). CERN-AB-Note-2004-021*, CERN, Geneva, Switzerland, 2004.

[172] W. M. Gui, R. F. Liu, G. Y. Bai, and X. Shao, "A novel approach to improve the frequency resolution based on sparse representation," in *Advanced Materials Research*, vol. 756, pp. 3336–3340, Trans Tech Publications Ltd, Zurich, Switzerland, 2013.

[173] I. Daubechies, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets in Medicine and Biology*, pp. 527–546, Routledge, London, UK, 1996.

[174] G. Thakur and H. T. Wu, "Synchrosqueezing-based recovery of instantaneous frequency from nonuniform samples," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 5, pp. 2078–2095, 2011.

[175] C. Li and M. Liang, "A generalized synchrosqueezing transform for enhancing signal time–frequency representation," *Signal Processing*, vol. 92, no. 9, pp. 2264–2274, 2012.

[176] T. Oberlin, S. Meignen, and V. Perrier, "The Fourier-Based synchrosqueezing transform," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 315–319, Florence, Italy, 2014.

[177] L. Li, H. Cai, Q. Jiang, and H. Ji, "An empirical signal separation algorithm for multicomponent signals based on linear time-frequency analysis," *Mechanical Systems and Signal Processing*, vol. 121, pp. 791–809, 2019.

[178] N. Saika, E. Maeda, N. Usuki, T. Arai, and Y. Murahara, "Developing mechanical models of the human vocal tract for education in speech science," in *Proceedings of the 2002 Forum Acusticum Sevilla, Spain*, 2002.

[179] S. Dabbaghchian, "Computational modeling of the vocal tract: applications to speech production," Doctoral dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2018.

[180] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1735–1745, 2007.

[181] L. Rabiner and R. Schafer, "Digital speech processing," *The Froehlich/Kent Encyclopedia of Telecommunications*, vol. 6, pp. 237–258, 2011.

[182] K. N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, MA, USA, 1998.

[183] E. Maeda, N. Usuki, T. Arai, N. Saika, and Y. Murahara, "Comparing the characteristics of the plate and clynder type vocal tract models," *Acoustical Science and Technology*, vol. 25, no. 1, 2004.

[184] T. Arai, "Sliding three-tube model as a simple educational tool for vowel production," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 384–388, 2006.

[185] T. Arai, "Education in acoustics and speech science using vocal-tract models," *Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2444–2454, 2012.

[186] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, Netherlands, 1960.

[187] M. Dziubiński and B. Kostek, "High accuracy and octave error immune pitch detection algorithms," *Archives of Acoustics*, vol. 29, 2004.

[188] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, pp. 374–388, Academic Press, New York, NY, USA, 1976.

[189] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[190] J. S. Bridle and M. D. Brown, "An experimental automatic word-recognition system," JSRU Report No. 1003, Joint Speech Research Unit, London, UK, 1974.

[191] J. Hillenbrand and R. T. Gayvert, "Vowel classification based on fundamental frequency and formant frequencies," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 4, pp. 694–700, 1993.

[192] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels,"

Journal of the Acoustical Society of America, vol. 97, no. 5, pp. 3099–3111, 1995.

[193] J. M. Hillenbrand, R. T. Gayvert, and M. J. Clark, "Phonetics exercises using the Alvin experiment-control software," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 2, pp. 171–184, 2015.

[194] P. F. Assmann and T. M. Nearey, "Perception of front vowels: the role of harmonics in the first formant region," *Journal of the Acoustical Society of America*, vol. 81, no. 2, pp. 520–534, 1987.

[195] P. F. Assmann and W. F. Katz, "Time-varying spectral change in the vowels of children and adults," *Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1856–1866, 2000.

[196] A. K. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1999.

[197] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.