

Pitch extraction using discrete cosine transform based power spectrum method in noisy speech

Humaira Sunzida¹, Nargis Parvin², Jafrin Akter Jeba¹, Sulin Chi³, Md. Shiplu Ali¹, Moinur Rahman¹,
Md. Saifur Rahman¹

¹Department of Information and Communication Technology, Faculty of Engineering, Comilla University, Cumilla, Bangladesh

²Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh

³Department of Information Engineering, Otemon Gakuin University, Osaka, Japan

Article Info

Article history:

Received Jun 8, 2024

Revised Mar 9, 2025

Accepted Jun 8, 2025

Keywords:

Autocorrelation function
Cumulative power spectrum
Discrete cosine transform
Fundamental frequency
Pitch

ABSTRACT

The pitch period is a key component of many speech analysis research projects. In real-world applications, voice data is frequently gathered in noisy surroundings, therefore algorithms must be able to manage background noise well in order to estimate pitch accurately. Despite advancements, many state-of-the-art algorithms struggle to deliver adequate results when faced with low signal-to-noise ratios (SNRs) in processing noisy speech signals. This research proposes an effective concept specifically designed for speech processing applications, particularly in noisy conditions. To achieve this goal, we introduce a fundamental frequency extraction algorithm designed to tolerate non-stationary changes in the amplitude and frequency of the input signal. In order to improve the extraction accuracy, we also use a cumulative power spectrum (CPS) based on discrete cosine transform (DCT) rather than conventional power spectrum. We enhance extraction accuracy of our method by utilizing shorter sub-frames of the input signal to mitigate the noise characteristics present in speech signals. According to the experimental results, our proposed technique demonstrates superior performance in noisy conditions compared to other existing state-of-the-art methods without utilizing any kind of post-processing techniques.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Md. Saifur Rahman

Department of Information and Communication Technology, Faculty of Engineering, Comilla University
Kotbari, Cumilla, Bangladesh

Email: saifurice@cou.ac.bd

1. INTRODUCTION

The vocalized form of human communication, known as speech, is defined as the movement of different speech organs, to produce sounds. In other words, speech can be defined as a series of sounds arranged in a sequence. A symbolic representation of information that needs to be transmitted between people or between people and machines is sound. The speech signal, represented acoustically as fluctuations in air pressure, convey information between individuals or between individuals and machines. Speech may take the form being voiced, unvoiced, or silent, reflecting different approaches to vocalization and sound generation. A voiced sound occurs when the speaker's vocal cords vibrate during sound production, while an unvoiced sound is produced without vocal cord vibration and when nothing is coming out from mouth is considered as a silence part. When a person speaks, their vocal cords vibrate, and the pitch is determined by how long it takes for the cords to open and close, known as the pitch period. This periodicity defines the fundamental frequency, which is

also represented as the pitch. In voiced sounds, the perceived pitch is determined by the apparent periodicity of vocal cord vibrations. Essentially, "pitch" in speech corresponds to the frequency of vocal cord vibrations during voiced sounds [1]. Pitch level correlates with the fundamental frequency: lower frequencies correspond to lower pitches, while higher frequencies indicate higher pitches [2]. Children and females capable of reaching frequencies up to 500 Hz, while males typically have a lower fundamental frequency around 60 Hz [3].

Pitch, or fundamental frequency (F_0), is vital in speech production, reflecting the rate of vocal fold vibration and influencing intonation and emotion perception. Accurate pitch estimation is essential across multiple fields like speech processing and music, enabling tasks such as music analysis, speech prosody understanding, and telecommunications. Precision in pitch extraction significantly impacts the effectiveness of applications like music synthesis, speech processing, and voice modulation [4], [5].

Up till now, a variety of pitch recognition methods have been covered. Pitch detection algorithm (PDA) is the term used to describe these techniques, which were founded on various mathematical principles [6]. PDAs can be used in three different ways: in the frequency domain, in the time domain, or in combination of the two [7]. Some pitch detection methods focus on identifying and timing specific features in the time domain. Pitch estimators in the time domain usually have three parts: a basic estimator, a post processor for error correction, and a preprocessor for signal simplification. Within this domain, various techniques, such as autocorrelation function (ACF) [8], average magnitude difference function (AMDF) [9], average squared mean difference function (ASMDF) [10], weighted autocorrelation function (WAF) [11], and YIN [12].

The autocorrelation approach is the most often used method for figuring out a voice signal's pitch period. The correlation between the input signal and a time-delayed version of itself is indicated by the ACF. AMDF, known for showcasing low points at integral multiples of the pitch period, is often utilized for pitch estimation [13]. AMDF stands as an alternative approach to autocorrelation analysis, presenting a simplified version compared to ACF.

With AMDF, as opposed to ACF, the delayed speech is subtracted from the original to create a difference signal, and the absolute magnitude is then determined at each delay value. In the WAF method, the periodicity property shared with ACF and AMDF is utilized. The WAF is characterized by employing the ACF as its numerator and the AMDF as its denominator. An algorithm called the YIN technique analyzes the traditional ACF [14].

In frequency domain techniques, various techniques have been developed to analyze the frequency domain cepstrum coefficients or spectrum of periodic signals in order to extract pitch. The cepstrum (CEP) [15] method is one of the most well-known methods. This method, relies on spectral characteristics. CEP is able to distinguish vocal tract features from periodic components. However, its performance is significantly compromised in a noisy environment, where the presence of noise has a pronounced impact on the log-amplitude spectrum.

Enhancements to the cepstrum method are tackled in the modified cepstrum (MCEP) [16]. Features from both windowless autocorrelation function (WLACF) and cepstral analysis are included in the cepstrum technique known as WLACF-CEP. WLACF reduces noise in the speech signal without compromising its periodicity. Pitch estimation filter with amplitude compression (PEFAC) utilizes summations of sub-harmonics in the log frequency domain. To improve its resilience to noise, the PEFAC incorporates an amplitude compression technique [17].

Using both logarithmic and power functions, [18] reduces the effect of formants and utilizes the Radon transform to provide a novel method for estimating pitch in noisy speech conditions. It also incorporates the Viterbi algorithm for pitch pattern refinement. Mnasri *et al.* [19] based on establishing a pragmatic relationship between the instantaneous frequency (F_i) and the fundamental frequency (F_0). It determines whether speech areas are voiced or unvoiced and extracts the F_0 contour by approximating it as a smoothed envelope of remaining F_i values. To estimate pitch by comparing the temporal accumulations of clean and noisy speech samples, the topology-aware intra-operator parallelism strategy searching (TAPS) algorithm, as described in [20], trains a set of peak spectrum exemplars. To understand how noise affects the locations and amplitudes within the spectrum of clear speech, Chu and Alwan developed the statistical algorithm for F0 estimation (SAFE) model [21]. Pitch estimation is enhanced using self-supervised pitch estimation (SPICE), as stated in [22], by refining the acquired data and training a constant Q transform of signals. To accommodate pitches with varying noise levels, Deep F_0 [23] expands the network's receptive range. It has been demonstrated that Harmo F_0 outperforms Deep F_0 in pitch estimation by employing a range of dilated convolutions. On the other hand, BaNa [24] opts for the initial five amplitude spectral peaks from the speech signal's spectrum on average for both male

and female speakers.

Existing methods often struggle with accuracy in noisy conditions, particularly when the signal-to-noise ratio (SNR) is low. In a novel approach, the study explores using discrete cosine transform (DCT) [25] instead of fast Fourier transform (FFT) [26], which proves effective in noisy signals but susceptible to vocal tract effects, resulting in some inconsistencies. However, when DCT was applied directly in power spectrum, detection accuracy decreased. To mitigate noise impact and improve accuracy, the study introduces a novel method combining cumulative power spectrum (CPS) with DCT features.

Instead of the conventional power spectrum, the proposed technique employs CPS based on DCT. CPS emphasizes the shorter sub-frames which is more effective to reduce the noise characteristics as well as mitigate the effect of vocal tract. Therefore, the proposed approach outperforms traditional pitch extraction methods in noisy speech signals by effectively suppressing noise components, demonstrating superior efficacy in fundamental frequency extraction under noisy conditions.

2. PROPOSED METHOD

Assuming that $y(n)$ represents a speech signal impacted by noise, as specified by (1),

$$y(n) = s(n) + w(n) \quad (1)$$

Where $w(n)$ is additive noise and $s(n)$ is a clean speech signal. The CPS approach's block diagram is displayed in Figure 1. The initial step involves dividing the noise corrupted speech signal $y(n)$ into frames.

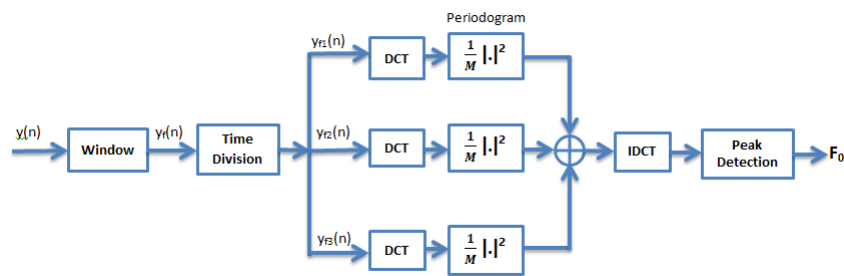


Figure 1. Block diagram of DCT based CPS

In this approach, framing is accomplished by employing a rectangular window function. In our experiments, the input signal needs to be partitioned into frames, each comprising 800 samples (equivalent to 50 [ms]). The signal framed as $y_f(n)$, where $0 \leq n \leq N - 1$, is partitioned into three sub-frames using a time division approach. These sub-frames are part as (2)-(4).

$$y_{f,1}(n) = y_f(n), 0 \leq n \leq M - 1 \quad (2)$$

$$y_{f,1}(n - D) = y_f(n), D \leq n \leq D + M - 1 \quad (3)$$

$$y_{f,1}(n - 2D) = y_f(n), 2D \leq n \leq 2D + M - 1 \quad (4)$$

In this context, where M represents an integer indicating the sub-frame length and D denotes the frame shift in samples, the goal is typically to set $2D + M - 1$ to be equal to N . In section 3, the values for the lengths of M and D are specified as 30 [ms] and 10 [ms], respectively. The signal $y_f(n)$, where $0 \leq n \leq N - 1$, undergoes frequency domain transformation through Periodogram computation using DCT. We examine the $y_f(n)$ based power spectrum to obtain information about the basic frequencies regarding the DCT.

DCT is a Fourier-related transform that uses only real values, much similar to discrete Fourier transform (DFT) [27]. The DCT was favored over the DFT in the transformation of actual signals, like an acoustic signal. Different kinds of DCT and inverse discrete cosine transform (IDCT) pairings can be used for implementation purposes. The DFT changes a complicated signal within its intricate spectrum. On the other hand, half of the data is redundant and half of the computation is wasted if the signal is real, as it is in the majority of applications. DCT tends to concentrate signal energy in a smaller number of coefficients compared to DFT.

The DFT provides a complex spectrum for a real signal, thereby wasting over half of the data. On the other hand, the DCT eliminates the need to compute redundant data by producing a true spectrum of real signals. DCT gathers most of the signal's information and sends it to the signal's lower-order coefficients, resulting in a large reduction in processing costs [28]. DCT avoids superfluous data and computation by producing a real spectrum of a real signal as a real transform. DCT has a further benefit in that it requires a straightforward phase unwrapping procedure because it is a real function. Furthermore, as DCT is derived from DFT, all of DFT's advantageous characteristics are retained, and a quick algorithm is available. Because DCT is a fully real transform and doesn't require complex variables or arithmetic, it is computationally more efficient than DFT. Taking into account the benefits of DCT for actual signals, the DCT $Y_f(k)$ of $y_f(n)$ is chosen and derived as (5).

$$Y_f(k) = c_d(k) \sum y_f(n) \cos \left[\frac{\pi(2n-1)(k-1)}{2N} \right] \quad (5)$$

In (5), k represents the frequency bin index, and the coefficient $c_d(k)$ can be found as follows:

Here, $c_d(k) = \sqrt{\frac{1}{N}}$ for $k=1$, and $c_d(k) = \sqrt{\frac{2}{N}}$ for $2 \leq k \leq N$. Therefore, the $Y_f(k)$ is obtained. The fundamental frequency and higher harmonics are represented as sharper, higher amplitude peaks in the DCT spectrum. DCT's downsampled or compressed spectra allow for the location of the higher harmonics at the fundamental frequency. The resultant spectrum, identified as the power spectrum of $y_f(n)$, is denoted as $P_f^y(k)$, where k corresponds to the frequency bin number associated with a discrete representation of w represented by w_k . For each sub-frame $y_{f,1}(n)$, where $j = 1, 2, 3$ and $0 \leq n \leq M - 1$, the power spectra are computed as $P_{f,1}^y(k)$, $P_{f,2}^y(k)$, and $P_{f,3}^y(k)$. The accumulations of these three power spectra are performed for each frequency bin as (6).

$$\bar{P}_f^y(k) = \sum_{j=1}^3 P_{f,j}^y(k) \quad (6)$$

The obtained power spectrum undergoes an IDCT. By identifying the maximum location in the resulting ACF, the fundamental frequency of $y_f(n)$ is detected. Figure 2 displays the output waveforms of the noisy speech signal, the conventional ACF approach, DCT-based ACF, and DCT-based CPS.

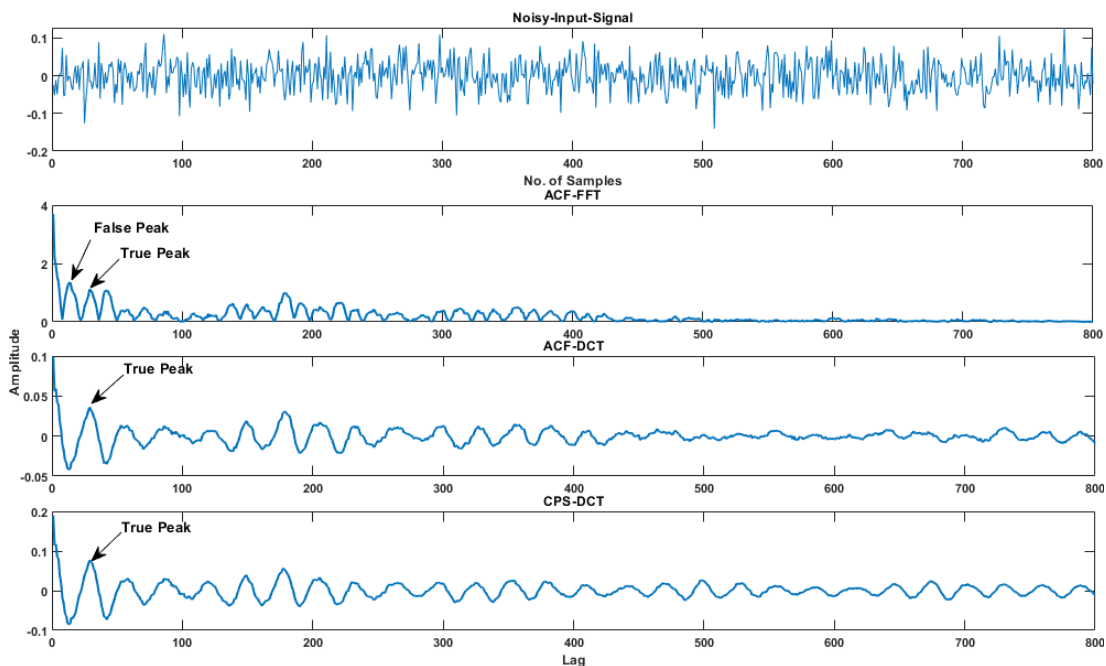


Figure 2. Validation of CPS-DCT using output waveform

In Figure 2, the false peak represents the vocal tract effect, while the true peak indicates the fundamental frequency. The conventional ACF output waveform is notably impacted by the vocal tract effect, resulting in a false peak close to the true peak. The adoption of DCT in place of FFT within ACF helps alleviate the vocal tract effect. Whereas our proposed method plays a crucial role in achieving a smoother signal than the DCT-based ACF. It not only significantly reduces the vocal tract effect but also provides a more seamless waveform compared to other methods. The results from the autocorrelation method applied to a voiced frame are illustrated in Figure 2. The waveform in the Figure 2 represents the effect of FFT and DCT in ACF of the speech signal. These figures depict the outcome of speech delivered by a male speaker in the presence of white noise. We have already explored that in the cross-correlation of noisy and clean speech, this component becomes zero. Hence, clean speech is significantly emphasized, and the ACF proves to be very effective in the case of a noisy signal. However, ACF is considerably influenced by the vocal tract effect, leading to some unsmooth occurrences in the signal due to noise. The use of DCT-based ACF can mitigate the vocal tract effect, yet some residual noise occurrences are still observable in the signal. Also when we used DCT in ACF, the detection accuracy went down. In order to further diminish the impact of noise characteristics and achieve better accuracy, we have introduced our proposed method that combines the feature of CPS with DCT. On the other hand, Figure 3 represents the validation of our proposed idea by utilizing the harmonic characteristics. From Figure 3, we have observed that DCT based CPS (proposed) is more effective against noise characteristics than that of FFT and DCT based power spectrum. In the case of FFT based power spectrum, we have investigated that harmonics are highly affected by noise which is marked by circle.

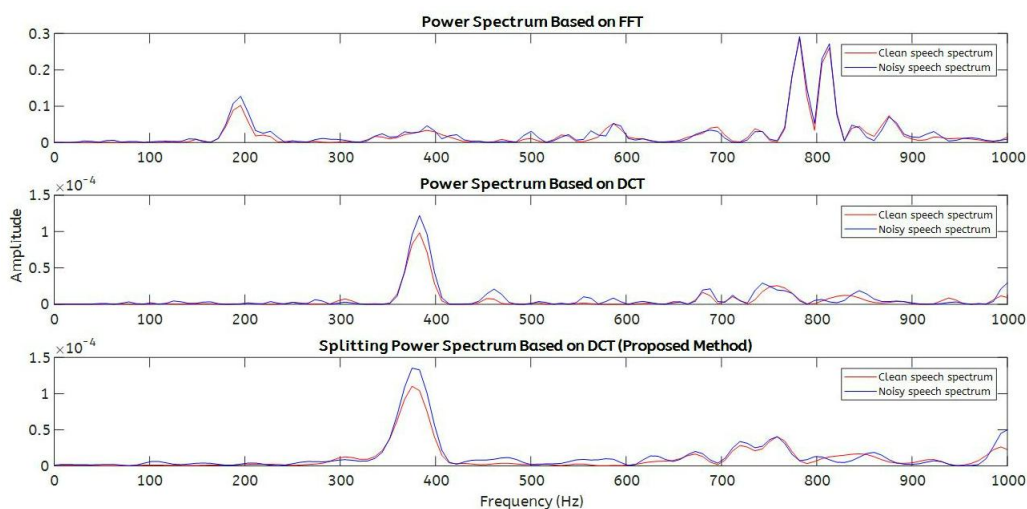


Figure 3. Validation of CPS-DCT using harmonic characteristics

3. RESULTS AND DISCUSSION

In this section, we assess the effectiveness of the CPS in identifying the fundamental frequency in the presence of noisy speech. Our assessment involves conducting experiments on speech signals to examine the performance of the cumulation-based approach. Ultimately, we present a comparative analysis of the outcomes achieved with our proposed method against those obtained from conventional pitch detection methods.

3.1. Experimental conditions

The proposed pitch detection method is implemented using speech signals obtained from the KEELE database [29] and the NTT database [30]. This database contains speech recordings from ten speakers, evenly divided between five males and five females. The collective duration of speech signals extracted from the KEELE database, encompassing the speeches of all ten speakers, amounts to around 5.5 [m]. These speech signals were sampled at a frequency of 16 [kHz]. Eight utterances by Japanese speakers, each lasting ten seconds and with a 3.4 [kHz] band limitation and 10 [kHz] sampling rate, are available in the NTT database. This research introduces a novel idea that proves to be more suitable for speech processing applications, particularly in the accurate retrieval of pitch from speech signals under noisy conditions. To simulate noisy speech sam-

ples, we blend clean speech recordings with noise collected from environments with high levels of background sound. To create the appropriate noisy voice samples, our method combines several forms of noise with the original speech signals. Four distinct noise categories, each with different SNR levels, were introduced into the initial signals to evaluate the algorithms' robustness to noise. These noise categories include white noise, babble noise, train noise, high frequency (HF)-channel noise, all obtained from the NOISEX-92 [31], sampled at a frequency of 20 [kHz]. The noises were adjusted to a 16 [kHz] sample frequency in order to match the KEELE database's signal properties and 10 [kHz] sample frequency in order to match the NTT database's signal properties. The SNR, or signal-to-noise ratio was systematically varied at levels of (0, 5, 10, 15 and 20 [dB]) for the assessment. The remaining experimental parameters for extracting the fundamental frequency were as follows:

- Frame length without PEFAC and BaNa, the frame length is 50 [ms].
- The frame shift is 10 [ms].
- Window type: rectangular, with the exception of BaNa and PEFAC.
- DCT (IDCT) points: 2048 points (KEELE) and 1024 points (NTT) when BaNa and PEFAC are not present.

3.2. Evaluation criteria

Pitch estimation error is determined by measuring the difference between the reference and estimated fundamental frequencies. The accuracy of basic frequency detection is assessed, following Rabiner's rule [31], utilizes the fundamental frequency detection error $e(l)$.

$$e(l) = F_{est}(l) - F_{true}(l) \quad (7)$$

Where l is frame number, $F_{est}(l)$ is estimated fundamental frequency at the l -th frame from a noisy spoken signal, and $F_{true}(l)$ is true fundamental frequency at the l -th frame.

If the absolute value of $e(i)$ exceeds 10% , (i.e. $|e(i)| > 10\%$) of $F_{true}(i)$, it falls under the category of gross pitch error (GPE), and the overall proportion of this error is computed for each uttered frame in the speech data. The error was designated as the fine pitch error (FPE) if $|e(i)| \leq 10\%$ from the ground truth first harmonic frequency. We specifically identified and evaluated the voiced portions in sentences concerning the fundamental frequency. Our analysis utilized a search range from $f_{min} = 50[Hz]$ to $f_{max} = 400[Hz]$, corresponding to the fundamental frequency range commonly observed in most people.

3.3. Results and performance comparison

In this section, we conduct a comparative analysis between our proposed method and conventional approaches, such as PEFAC, BaNa, and YIN, using distinct utterances from the KEELE and NTT databases. We evaluate performance under four types of noise: white noise, babble noise, HF channel noise, and train noise. Parameters like frame length, window function, and the number of DFT (IDFT) points specific to PEFAC and BaNa were adjusted, while other parameters remained consistent across methods. The Hamming window function was applied uniformly in PEFAC and BaNa. For BaNa, the frame duration was set to 60 [ms], and 2^{16} points were used for DFT (IDFT) points. The source code of BaNa, tailored for this environment, was implemented (as described in [32]). PEFAC utilized a Hamming window function with a duration of 90 [ms] for both the window function and frame length. The source code used 2^{13} as the value for the DFT (IDFT) points. The implementation of PEFAC in this environment is well-suited for BaNa (as indicated in [17], [33]). Performance evaluation was conducted using the GPE and the FPE. The average GPE and FPE results obtained from the experimental outcomes of the proposed method, PEFAC, BaNa, YIN, were considered for utterances from both female and male speakers at various SNRs (0, 5, 10, 15 and 20 [dB]).

Tables 1-8 present a comparison of GPE for the KEELE database and NTT database, respectively under various noise conditions, including white noise, babble noise, HF channel noise, and train noise. On the other hand, Tables 9-16 present a comparison of FPE for the KEELE database and NTT database, respectively under the above noise conditions. The GPE and FPE values of our proposed method are contrasted with those of PEFAC, BaNa, and YIN.

Table 1. Average GPE rate (%) for KEELE database for white noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	20.58	37.15	22.61	31.37
5	15.96	34.38	19.58	21.59
10	13.86	33.01	17.80	16.57
15	13.12	32.50	16.97	14.29
20	12.90	31.98	16.59	12.87

Table 2. Average GPE rate (%) for KEELE database for babble noise

SNR[dB]	Proposed	PEFAC	BaNa	YIN
0	35.18	49.01	40.54	36.89
5	22.88	41.86	29.48	23.68
10	16.57	37.41	22.84	16.64
15	13.09	34.98	19.69	13.16
20	11.87	33.39	17.70	12.14

Table 3. Average GPE rate (%) for KEELE database for train noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	33.44	43.17	29.08	34.38
5	22.81	38.99	23.11	22.76
10	16.98	35.59	20.04	16.36
15	14.50	33.40	18.31	13.42
20	13.49	32.25	17.36	12.16

Table 4. Average GPE rate (%) for KEELE database for HF-channel noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	24.70	40.13	22.64	31.55
5	17.64	36.86	19.82	21.01
10	14.79	34.37	17.90	16.06
15	13.45	32.98	17.31	13.76
20	13.04	32.11	16.57	12.79

Table 5. Average GPE rate (%) for NTT database for white noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	4.71	17.47	8.00	14.20
5	1.90	12.89	5.52	4.70
10	1.38	11.34	3.98	2.08
15	1.36	11.93	3.26	1.55
20	1.38	13.21	3.30	1.46

Table 6. Average GPE rate (%) for NTT database for babble noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	28.26	39.86	27.71	31.75
5	10.01	24.75	12.60	12.31
10	2.80	16.11	5.20	3.20
15	1.58	12.45	4.08	1.52
20	1.44	11.69	4.02	1.41

Table 7. Average GPE rate (%) for NTT database for train noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	14.98	25.28	10.91	20.32
5	4.66	16.3657	5.72	6.76
10	1.92	12.28	4.28	2.34
15	1.38	10.21	3.44	1.61
20	1.36	9.29	3.47	1.33

Table 8. Average GPE rate (%) for NTT database for HF-channel noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	5.73	18.91	5.97	14.32
5	2.34	13.13	4.52	4.84
10	1.62	11.00	4.41	2.02
15	1.49	10.72	4.29	1.48
20	1.45	10.06	4.13	1.39

Table 9. Average FPE rate (Hz) for KEELE database for white noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	4.42	5.45	5.23	4.54
5	4.14	5.36	5.22	3.97
10	4.03	5.32	5.19	3.60
15	3.99	5.26	5.14	3.46
20	3.97	5.25	5.08	3.44

Table 10. Average FPE rate (Hz) for KEELE database for babble noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	4.54	5.62	5.29	4.12
5	4.28	5.49	5.18	3.79
10	4.10	5.38	5.11	3.59
15	4.01	5.30	5.09	3.50
20	3.98	5.24	5.08	3.50

Table 11. Average FPE rate (Hz) for KEELE database for train noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	4.48	5.51	5.30	3.96
5	4.24	5.40	5.15	3.68
10	4.06	5.33	5.11	3.53
15	3.98	5.31	5.05	3.45
20	3.95	5.27	5.03	3.44

Table 12. Average FPE rate (Hz) for KEELE database for HF channel noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	4.62	5.51	5.24	4.21
5	4.30	5.38	5.21	3.80
10	4.10	5.33	5.21	3.56
15	3.99	5.30	5.14	3.48
20	3.97	5.29	5.11	3.43

Table 13. Average FPE rate (Hz) for NTT database for white noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	3.01	3.42	2.39	3.82
5	2.69	3.34	2.20	2.59
10	2.53	3.25	2.09	2.16
15	2.49	3.20	2.00	2.03
20	2.49	3.15	1.95	1.99

Table 14. Average FPE rate (Hz) for NTT database for babble noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	2.26	3.88	2.69	3.09
5	2.40	3.52	2.25	2.42
10	2.50	3.31	2.03	2.15
15	2.50	3.21	1.93	2.02
20	2.48	3.16	1.84	1.99

Table 15. Average FPE rate (Hz) for NTT database for train noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	2.84	3.61	2.51	3.25
5	2.70	3.44	2.19	2.44
10	2.56	3.25	2.05	2.44
15	2.51	3.15	1.94	2.02
20	2.49	3.13	1.87	1.99

Table 16. Average FPE rate (Hz) for NTT database for HF channel noise

SNR [dB]	Proposed	PEFAC	BaNa	YIN
0	3.13	3.55	2.34	3.85
5	2.69	3.40	2.19	2.70
10	2.54	3.28	2.09	2.18
15	2.50	3.17	2.01	2.02
20	2.48	3.11	1.93	1.99

In the case of KEELE database, the proposed approach consistently exhibits the lowest average GPE rate compared to other techniques across almost all SNRs in all noise cases except low SNR (0 [dB]) at train and HF channel noise cases. At SNR (0 [dB]) in train and HF channel noise cases, BaNa provides the slightly lower gross pitch error rate due to processing strategy according to the noise characteristics. On the other hand, in the case of NTT database, the proposed method shows the almost similar properties with the KEELE database.

In the case of FPE of Tables 9-12 in KEELE database, the proposed method provides the lower FPE (Hz) than that of the PEFAC and BaNa at almost all SNRs in all noise cases except the YIN method. The proposed method is highly competitive with the YIN method except white noise case. In the case of NTT database, the FPE (Hz) of the proposed method is lower than that of PEFAC and YIN method and highly competitive with BaNa except babble noise. In babble noise, proposed method shows the superior performance compared with the other methods.

4. CONCLUSION

Accurately estimating perfect pitch poses a challenge in speech analysis, especially in noisy environments. In this study, we introduce an improved method that excels in isolating noise from the waveform, particularly in babble noise scenarios, outperforming other techniques. This method exhibits a lower average GPE rate compared to alternative approaches, and it achieves this without any complicated post-processing. Additionally, it efficiently mitigates the impact of vocal tract effects by equalizing unnecessary ripples in the waveform. According to their noise type and SNRs, our research so demonstrates that it is more robust than other traditional methods without requiring any complex post-processing. In the future, research might focus on creating a new pitch extraction technique that is more effective in speech processing applications and incredibly resilient to extremely low SNR instances across a range of real-world noise scenarios.

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Humaira Sunzida	✓	✓							✓		✓			
Nargis Parvin	✓					✓		✓	✓					
Jafrin Akter Jeba					✓			✓	✓					
Sulin Chi				✓		✓			✓					
Md. Shiplu Ali						✓			✓					
Moinur Rahman					✓					✓	✓			
Md. Saifur Rahman		✓	✓		✓		✓		✓			✓	✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

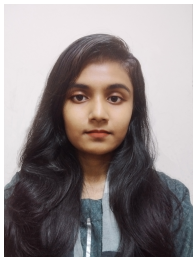
The authors confirm that the data supporting the findings of this study are available within the article.





REFERENCES

- [1] S. S. Upadhyaya, "Pitch detection in time and frequency domain," *Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012*, 2012, doi: 10.1109/ICCICT.2012.6398150.
- [2] M. S. Rahman, "Pitch extraction for speech signals in noisy environments," *Ph.D. Dissertation*, Department of Mathematics, Electronics, and Informatics, Saitama University, Saitama, Japan, 2020. [Online]. Available: <https://sucra.repo.nii.ac.jp/record/19377/files/GD0001258.pdf>.
- [3] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 6, pp. 1066–1078, 2016, doi: 10.1109/TASLP.2016.2540805.
- [4] D. Wang, C. Yu, and J. H. L. 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, doi: 10.1109/TASLP.2017.2667879.
- [5] D. Gerhard, "Pitch extraction and fundamental frequency: history and current techniques theory of pitch," *Technical Report TR-CS*, 2003. [Online]. Available: [https://www.cs.bu.edu/fac/snyder/cs583/Literature and Resources/PitchExtractionMastersThesis.pdf](https://www.cs.bu.edu/fac/snyder/cs583/Literature%20and%20Resources/PitchExtractionMastersThesis.pdf).
- [6] N. S. B. Ruslan, M. Mamat, R. R. Porle, and N. Parimon, "A comparative study of pitch detection algorithms for microcontroller based voice pitch detector," *Advanced Science Letters*, vol. 23, no. 11, pp. 11521–11524, 2017, doi: 10.1166/asl.2017.10320.
- [7] L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of Voice*, vol. 29, no. 4, pp. 410–417, 2015, doi: 10.1016/j.jvoice.2014.09.016.
- [8] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977, doi: 10.1109/TASSP.1977.1162905.
- [9] A. Cohen et al., "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-22, no. 5, pp. 353–362, 1974, doi: 10.1109/TASSP.1974.1162598.
- [10] R. Chakraborty, D. Sengupta, and S. Sinha, "Pitch tracking of acoustic signals based on average squared mean difference function," *Signal, Image and Video Processing*, vol. 3, no. 4, pp. 319–327, 2009, doi: 10.1007/s11760-008-0072-5.
- [11] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, 2001, doi: 10.1109/89.952490.
- [12] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002, doi: 10.1121/1.1458024.
- [13] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 322–335, 2012, doi: 10.1109/TASLP.2011.2161579.
- [14] H. Hajimolahoseini, R. Amirfattahi, S. Gazor, and H. Soltanian-Zadeh, "Robust estimation and tracking of pitch period using an efficient bayesian filter," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 7, pp. 1219–1229, 2016, doi: 10.1109/TASLP.2016.2551041.
- [15] W. Hu, X. Wang, and P. Gómez, "Robust pitch extraction in pathological voice based on wavelet and cepstrum," *European Signal Processing Conference*, pp. 297–300, 2015. [Online]. Available: <https://new.eurasip.org/Proceedings/Eusipco/Eusipco2004/defevent/papers/cr1417.pdf>.
- [16] M. S. Rahman, Y. Sugiura, and T. Shimamura, "Utilization of windowing effect and accumulated autocorrelation function and power





- spectrum for pitch detection in noisy environments,” *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 11, pp. 1680–1689, 2020, doi: 10.1002/tee.23238.
- [17] S. Gonzalez, “Pefac-a pitch estimation algorithm robust to high levels of noise,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014, doi: 10.1109/TASLP.2013.2295918.
- [18] B. Li and X. Zhang, “A pitch estimation algorithm for speech in complex noise environments based on the Radon transform,” *IEEE Access*, vol. 11, pp. 9876–9889, 2023, doi: 10.1109/ACCESS.2023.3240181.
- [19] Z. Mnasri, S. Rovetta, and F. Masulli, “A novel pitch detection algorithm based on instantaneous frequency for clean and noisy speech,” *Circuits, Systems, and Signal Processing*, vol. 41, no. 11, pp. 6266–6294, 2022, doi: 10.1007/s00034-022-02082-8.
- [20] F. Huang and T. Lee, “Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 99–109, 2013, doi: 10.1109/TASL.2012.2215589.
- [21] 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, 2012, doi: 10.1109/TASL.2011.2168518.
- [22] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, “SPICE: self-supervised pitch estimation,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 1118–1128, 2020, doi: 10.1109/TASLP.2020.2982285.
- [23] S. Singh, R. Wang, and Y. Qiu, “DeepF0: end-to-end fundamental frequency estimation for music and speech signals,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. June, pp. 61–65, 2021, doi: 10.1109/ICASSP39728.2021.9414050.
- [24] N. Yang, H. Ba, W. Cai, I. Demirkol, and W. Heintzman, “BaNa: a noise resilient fundamental frequency detection algorithm for speech and music,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 12, pp. 1833–1848, 2014, doi: 10.1109/TASLP.2014.2352453.
- [25] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, Jan. 1974, doi: 10.1109/T-C.1974.223784.
- [26] P. Duhamel and M. Vetterli, “Fast Fourier transforms: a tutorial review and a state of the art,” *Signal Processing*, vol. 19, no. 4, pp. 259–299, Apr. 1990, doi: 10.1016/0165-1684(90)90158-U.
- [27] F. J. Harris, “Time domain signal processing with the DFT,” *Handbook of Digital Signal Processing*, pp. 633–699, 1987, doi: 10.1016/b978-0-08-050780-4.50013-8.
- [28] F. Plante, G. Meyer, and W. Ainsworth, “A pitch extraction reference database,” *4th European Conference on Speech Communication and Technology*, pp. 837–840, 1995, doi: 10.21437/Eurospeech.1995-191.
- [29] Y. Meng, “Speech recognition on DSP: algorithm optimization and performance analysis,” *Master Thesis, Department of Electronic Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong*, 2004. [Online]. Available: <http://www.ee.cuhk.edu.hk/myuan/Thesis.pdf>.
- [30] NTT Advanced Technology Corp, *20 countries language database*, NTT Advanced Technology Corp, 1988.
- [31] A. Varga and H. 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*, vol. 12, no. 3, pp. 247–251, 1993, doi: 10.1016/0167-6393(93)90095-3.
- [32] University of Rochester, “Wireless communication and networking group,” *hajim.rochester.edu*. Accessed: Mar. 02, 2024. [Online]. Available: <https://hajim.rochester.edu/ece/sites/wcng/code.html>.
- [33] M. Brookes, “VOICEBOX: speech processing toolbox for MATLAB,” *ee.ic.ac.uk*. Accessed: Mar. 02, 2024. [Online]. Available: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

BIOGRAPHIES OF AUTHORS






Humaira Sunzida     obtained her B.Sc. (Engineering) degree in Information and Communication Technology from Comilla University, Cumilla, Bangladesh, in 2024. She started her undergraduate studies in the Department of Information and Communication Technology at Comilla University in 2019. Her current research interests encompass speech analysis and digital signal processing. She can be contacted at email: humairasunzida.311@stud.cou.ac.bd.






Nargis Parvin     received her B.Sc. (Honours) and M.Sc. degrees in Information and Communication Engineering from the University of Rajshahi, Rajshahi, Bangladesh, in 2006 and 2007, respectively. In 2013, she joined as a Lecturer in the Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla Cantonment, Cumilla, Bangladesh, where she is currently serving as Assistant Professor. She pursued her Ph.D. degree in the field of wireless sensor network (WSN) at the Graduate School of Science and Engineering at Saitama University, Japan. Her research interests include wireless sensor network, speech analysis and digital signal processing. She can be contacted at email: nargis.cse@baiust.ac.bd.






Jafrin Akter Jeba    obtained her B.Sc. (Engineering) degree in Information and Communication Technology from Comilla University, Cumilla, Bangladesh, in 2024. She started her undergraduate studies in the Department of Information and Communication Technology at Comilla University in 2019. Her current research interests encompass speech analysis and digital signal processing. She can be contacted at email: jafrinjeba10@gmail.com.






Sulin Chi    received her B.S. degree from the University of Jinan, Shandong, China in 2017, and her M.S. degree from Saitama University, Japan in 2020. She is currently pursuing her Ph.D. in the field of communication signal processing at the Graduate School of Science and Engineering, Saitama University, Japan. In 2024, she joined Otemon Gakuin University, Osaka, Japan, where she is currently an Assistant Professor. Her research interests include signal processing, blind equalizer, wireless sensor network, and their applications. She can be contacted at email: s-chi@haruka.otemon.ac.jp.






Md. Shiplu Ali    received his B.Sc. (Engineering) in Information and Communication Technology from Comilla University, Cumilla, Bangladesh, in 2023. He started his undergraduate studies in the Department of Information and Communication Technology at Comilla University in 2018. His current research interests include speech analysis and digital signal processing. He can be contacted at email: shipluahmedcou11809042@gmail.com.



Moinur Rahman    received his B.Sc. (Engineering) and M.Sc. (Engineering) degrees in Information and Communication Technology from Comilla University, Cumilla, Bangladesh, in 2018 and 2019, respectively. In 2022, he joined as a Lecturer in the Department of Computer Science and Engineering, The People's University of Bangladesh, 3/2 Asad Avenue, Dhaka, Bangladesh. In March, 2023 he joined as a lecturer in the Department of Information Technology, University of Information Technology and Sciences, Baridhara, Dhaka, Bangladesh. Now he is currently serving as a lecturer in the Department of Information and Communication Technology, Comilla University from January 2025. His current research interests include speech analysis and digital signal processing. He can be contacted at email: moinur.rahman@cou.ac.bd.



Md. Saifur Rahman    received his B.Sc. (Honours) and M.Sc. degrees in Information and Communication Engineering from the University of Rajshahi, Rajshahi, Bangladesh, in 2006 and 2007, respectively. In 2012, he joined as a Lecturer in the Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh, where he is currently serving as an Associate Professor. He pursued his Ph.D. degree in the field of speech processing at the Graduate School of Science and Engineering at Saitama University, Japan. His research interests include speech analysis and digital signal processing. He can be contacted at email: saifurice@cou.ac.bd.