FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 730-734 | e-ISSN: 2347-470X

Fundamental Frequency Extraction by Utilizing the Combination of Spectrum in Noisy Speech

Foujia Islam[1], Nargis Parvin[2], Moinur Rahman[3], Md. Tofael Ahmed[4], Dulal Chakraborty[5], and Md. Saifur Rahman[6*]

[1]*Department of ICT, Comilla University, Bangladesh; Email: foujiaislam4567@gmail.com*
[2]*Assistant Professor, Department of CSE, Bangladesh Army International University of Science and Technology, Bangladesh; Email: nargis.cse@baiust.ac.bd*
[3]*Lecturer, Department of ICT, Comilla University, Bangladesh; Email: moinur.rahman@cou.ac.bd*
[4]*Professor, Department of ICT, Comilla University, Bangladesh; Email: tofael@cou.ac.bd*
[5]*Associate Professor, Department of ICT, Comilla University, Bangladesh; Email: dulal.ict.cou@gmail.com*
[6]*Professor, Department of ICT, Comilla University, Bangladesh; Email: saifurice@cou.ac.bd*

***Correspondence:** *Md. Saifur Rahman; Email: saifurice@cou.ac.bd*

ABSTRACT- Speech is the audible acoustic signal generated by the articulatory system (lungs, vocal folds, vocal tract, tongue, lips) to communicate language. The fundamental frequency (F_0) is the lowest frequency component of a speech waveform and corresponds to the vibration rate of the vocal folds during voiced speech. It also determines the pitch of the speaker's voice. In speech signal processing, this acoustic waveform is captured and analyzed to extract information about what is being said, how it is being said and who is saying it. In various speech processing applications such as voice synthesis, speaker recognition and emotion analysis accurate extraction of the fundamental frequency (F_0) is a vital task. However, the vocal tract's formants can sometimes significantly alter the glottal waveform's shape, making it difficult to identify the true pitch. Additionally, in the presence of background noise, traditional pitch detection techniques often experience a considerable decline in performance. This work proposes a robust method for extracting the fundamental frequency by utilizing the complementary advantages of the power spectrum and logarithmic spectrum in noisy speech environments. The power spectrum mitigates noise effects, while the logarithmic operation effectively separates vocal tract characteristics from the source excitation. The proposed approach integrates autocorrelation-based power spectral analysis with cepstral techniques derived from the log spectrum to improve pitch estimation accuracy under adverse conditions. Experimental results on noisy speech datasets show that the proposed hybrid method achieves lower gross pitch error and greater robustness than traditional methods such as BaNa, autocorrelation and cepstral techniques.

Keywords: Speech Enhancement, Power Spectrum, Logarithm Spectrum, Hybrid Method.

ARTICLE INFORMATION

Author(s): Foujia Islam, Nargis Parvin, Moinur Rahman, Md. Tofael Ahmed, Dulal Chakraborty, and Md. Saifur Rahman;
Received: 30/05/2025; **Accepted:** 22/09/2025; **Published:** 10/12/2025;
E- ISSN: 2347-470X;
Paper Id: IJEER 3005-15/19;
Citation: 10.37391/ijeer.130413
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130413.html

1. INTRODUCTION

Human conversation, an innate form of communication, depends on coordinated activity of respiratory and articulatory organs to generate speech sounds. Air from the lungs passes through the vocal folds, glottis and oral tract, emerging as acoustic waves. Voiced sounds arise from vocal fold vibration, while unvoiced sounds result from constrictions without vibration. Linguistically, speech is built from phonemes vowels and consonants that are shaped by the vocal tract. The fundamental frequency (F_0), perceived as pitch, reflects the rate of vocal fold vibration and varies across speakers and contexts [1, 2]. Pitch also fluctuates with emotional and intonational cues. Accurate F_0 estimation is essential for applications in cochlear implants, hearing aids, speech recognition and human-computer interaction (HCI). Robust extraction in noisy conditions remains a challenge, as most existing approaches target clean speech. However, with the effect of vocal tract and external noise, speech signal gets distorted and it becomes complicated to uphold the accuracy and dependability of pitch extraction procedures. This work proposes an efficient, frequency-domain–based pitch extraction technique designed for real-world noisy environments. The method avoids complex post-processing, mitigates vocal tract effects and offers time efficiency for practical applications, including IoT-based speech systems.

2. LITERATURE REVIEW

Pitch detection methods are broadly categorized into time-domain, frequency-domain and hybrid approaches [3].

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 730-734 | e-ISSN: 2347-470X

Time-domain methods analyze periodicity in the waveform using zero-crossings, autocorrelation and related measures [4–7]. The Autocorrelation Function (ACF) [8] is effective in clean speech but degrades under colored or non-stationary noise. The Average Magnitude Difference Function (AMDF) [9] offers lower complexity but suffers from pitch doubling and poor resolution. Weighted ACF (WAF) [10] reduces side-lobes using AMDF weighting, improving clarity. YIN [11] enhances the difference function with cumulative mean normalization, achieving higher accuracy but at higher computational cost and limited noise robustness. Frequency-domain methods exploit periodicity in the spectrum [12-14]. Speech is divided into frames, windowed and transformed via STFT. The Cepstrum (CEP) [15] identifies pitch from the inverse Fourier transform of the log spectrum but is noise-sensitive. Modified Cepstrum (MCEP) [16] applies liftering and clipping to reduce vocal tract and noise effects. Windowless ACF Cepstrum (WLACF-CEP) [17] avoids windowing artifacts, though still challenged by highly dynamic noise. Hybrid methods combine both domains for improved robustness. BaNa [18] estimates pitch from harmonic peaks but struggles in low-pitch or noisy speech. PEFAC [19] uses a harmonic comb filter in the log-frequency spectrum and performs well at low SNR, but is computationally heavy. YAAPT [20] blends time and frequency analyses with dynamic programming to smooth contours, requiring careful tuning. MBSC [21] aggregates subband correlograms for better noise robustness but at higher complexity.

While these methods achieve varying trade-offs, many rely on multi-band analysis, complex filtering, or heavy post-processing. Our proposed method is a purely frequency-domain approach, avoiding parameter tuning and computationally expensive steps, while retaining robustness under noisy conditions.

3. PROPOSED METHOD

A clean speech waveform by nature is extremely nonstationary and quasi-periodic rather than really periodic. Let us assume, an additive noise signal, u[m], contaminates the clean speech signal, w[m], resulting in the noisy speech signal, z[m]. The step-by-step process of the proposed algorithm for fundamental frequency extraction is outlined in the pseudocode below.

Algorithm 1: Proposed Hybrid Spectrum Method for F_0 Extraction

Input: A frame of the noisy speech signal, z[m].
Output: Estimated fundamental frequency, F_0.
BEGIN
Step 1: Apply Window Function
Let $R_{rec}[m]$ be the Rectangular window function.
$z_r[m] \leftarrow z[m] \cdot R_{rec}[m]$ // Apply window to the signal frame
Step 2: Transform to Frequency Domain
$Z[k] \leftarrow FFT(z_r[m])$ // Compute the Fast Fourier Transform
Step 3: Compute Power and Logarithmic Spectra in parallel
Branch A: Power Spectrum to mitigate noise effects
$P[k] \leftarrow |Z[k]|^2$
Branch B: Logarithmic Spectrum to separate vocal tract effects $L[k] \leftarrow \log(|Z[k]|)$
Step 4: Combine the Spectra to emphasize key spectral components
$C[k] \leftarrow P[k] \cdot L[k]$ // Element-wise multiplication to create the combined spectrum
Step 5: Transform back to Time-like Domain
$c[n] \leftarrow IFFT(C[k])$ // Compute the Inverse FFT
Step 6: Find the Fundamental Frequency
$n_{peak} \leftarrow argmax(c[n])$ // Find the index of the most prominent peak in the valid pitch range
$F0 \leftarrow n_{peak}/Sampling\ Rate$ // Convert the peak's index to frequency
RETURN F_0
END

Windowing divides the speech signal into number of periodic segments, essentially uses smoothing functions that taper to zero at the edges and reduce spectral leakage. When a speech signal is multiplied by a window function, the segment naturally decays at the edges, making the boundary irregularities less perceptible. Although windowing modifies the original signal, the transformation is carefully designed to preserve its spectral characteristics as much as possible. In our method, rectangular window function is applied on the noisy speech signal. The Rectangular window can be mathematically expressed as follows:

$$R_{rec}[m] = \begin{cases} 1, & for\ 0 \le m \le M-1 \\ 0, & otherwise \end{cases} \quad (1)$$

Rectangular window provides notable benefits in noisy conditions. Its narrower main-lobe bandwidth enhances the precision of pitch estimation by minimizing spectral leakage and concentrating energy on the primary frequency components. This makes it particularly effective when dealing with signals affected by noise. In figure 1, the block diagram of our proposed approach for extracting the (F_0) of speech signals is shown which is a pitch extraction method by combining the power spectrum and logarithm spectrum.

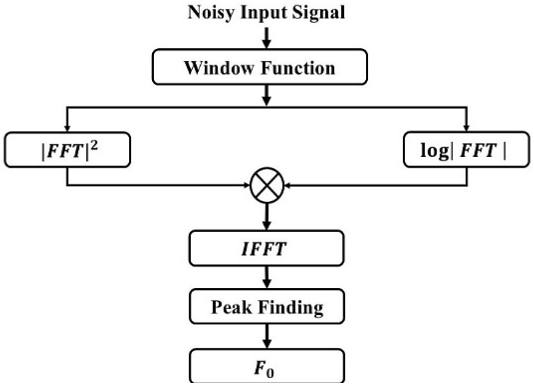

Figure 1. Block diagram of Proposed Method

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 730-734 | e-ISSN: 2347-470X

In *figure 2* we see that, in power spectrum, clean speech is affected by noise. As shown in *figure 3*, the logarithmic spectrum of clean speech becomes highly distorted when noise is introduced. We don't get proper pitch information as a consequence.

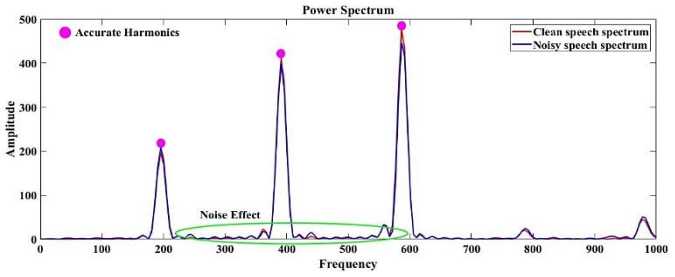

Figure 2. The Spectral Magnitude (Amplitude) vs Frequency (Hz) representation of the Power spectrum of a clean speech frame (female speaker, KEELE database) and the same frame corrupted by White Noise at an SNR of 5 dB (*figure 4*, we can find the accurate harmonics by reducing the both noise and effect of vocal tract)

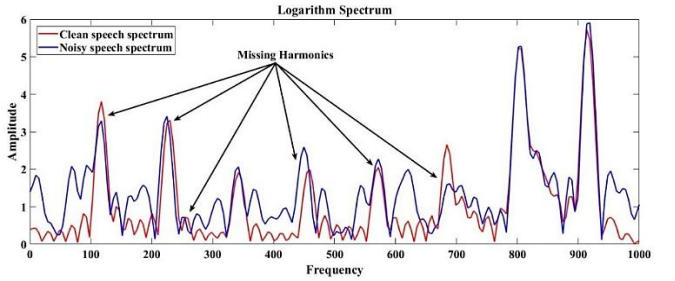

Figure 3. The Spectral Magnitude (Amplitude) vs Frequency (Hz) representation of the Log-magnitude spectrum of a clean speech frame from a female speaker (KEELE database) and the same frame corrupted by Babble Noise at 5 dB SNR

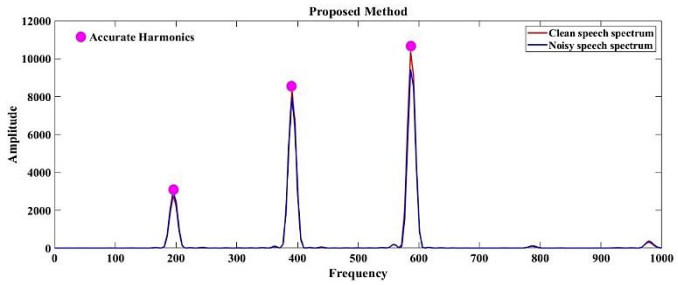

Figure 4. The Spectral Magnitude (Amplitude) vs Frequency (Hz) representation of the output spectrum of the Proposed Method for the same clean and noisy speech frames used in *figures 1 and 2* (female speaker, Babble Noise at 5 dB SNR)

4. RESULTS AND DISCUSSION

4.1. Experimental Conditions

The proposed pitch extraction technique is implemented using voice signals that were taken from the NTT and KEELE databases. The NTT database [22], developed by NTT Advanced Technology Corporation, includes recordings of four male and four female Japanese speakers. Each speech sample in the database has a duration of 11 seconds. A 10 kHz sampling rate was used for the voice signals. The KEELE database's [23] voice signals were produced by total of ten speakers, comprising five males and five females with a total duration of approximately 6 minutes across all speakers. The sampling rate for these voice sounds was 16 [kHz]. To validate our proposed method, we used five real-world noise types that simulate realistic acoustic environments: White Noise, Babble Noise, Train Noise, HF Channel Noise and Car Interior Noise [24]. Each type of noise was added to the clean speech signals according to their respective sampling rates. The experiments were conducted at varying SNR levels: 0, 5, 10, 15 and 20 dB under the following trial conditions:

- Frame length: 50ms except for BaNa;
- Frame shift: 10ms;
- Window functions: Rectangular;
- DFT (IDFT) points: 2048 for KEELE and 1024 for NTT.

4.2. Evaluation Criteria

Based on Rabiner's rule[8], the accuracy of the fundamental frequency extraction was assessed using the following equation:

$$e(l) = F_{est}(l) - F_{true}(l) \qquad (2)$$

Here $e(l)$ is the extraction error, $F_{est}(l)$ and $F_{true}(l)$ denote the estimated and true fundamental frequencies at the l-th frame, respectively, where l representing frame number. A gross pitch error (GPE) is identified if $|e(l)|$ exceeds 10% of the true fundamental frequencies. The GPE rate, expressed as a percentage, is then calculated over all voiced segments in the speeched data. *Figures 5* to *14* present the experimental results of GPE across different SNR levels for both KEELE and NTT databases under various noise conditions. From the figures, we observe that the proposed method consistently provides lower average GPE rates than ACF, CEP and BaNa across almost all SNR levels and noise types in the both KEELE and NTT databases. Overall, the proposed method demonstrates higher robustness to noise, as indicated by its consistently lower Gross Pitch Error (GPE). Our proposed method reduces the effect of vocal tract characteristics as well as suppresses the non-pitch peaks in the frequency domain, enhancing the pitch peak in the wide-band noise. Through experiments, we have confirmed that the proposed method is efficient and effective in extracting the pitch in a wide range of noise types.

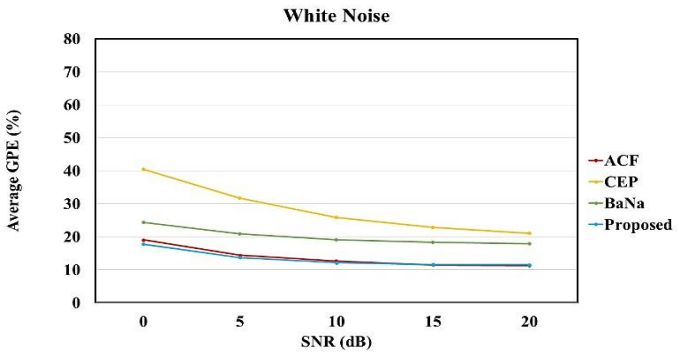

Figure 5. Average GPE Comparison for both Female Speakers and Male Speakers with White Noise across various SNR levels within KEELE database

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 730-734 | e-ISSN: 2347-470X

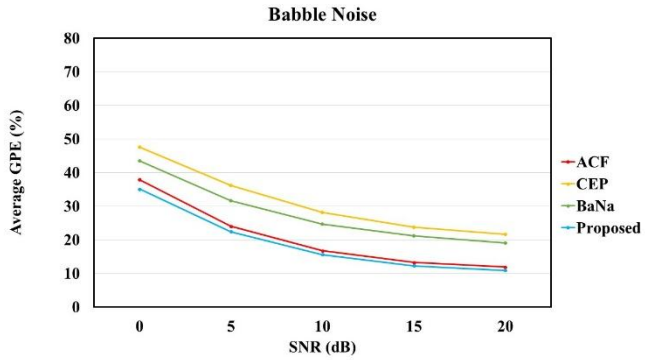

Figure 6. Average GPE Comparison for both Female Speakers and Male Speakers with Babble Noise across various SNR levels within KEELE database

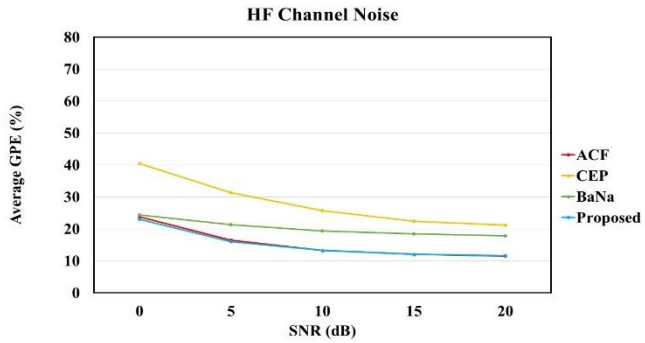

Figure 7. Average GPE Comparison for both Female Speakers and Male Speakers with HF Channel Noise across various SNR levels within KEELE database

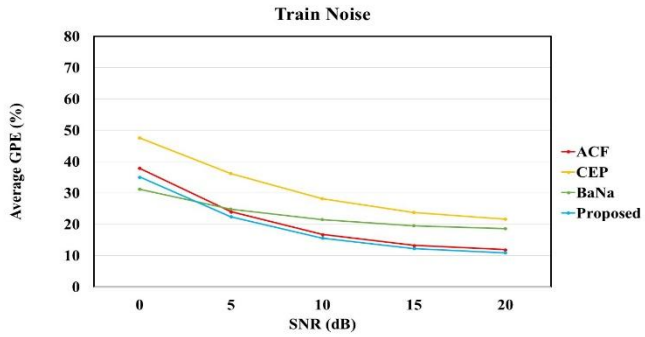

Figure 8. Average GPE Comparison for both Female Speakers and Male Speakers with Train Noise across various SNR levels within KEELE database

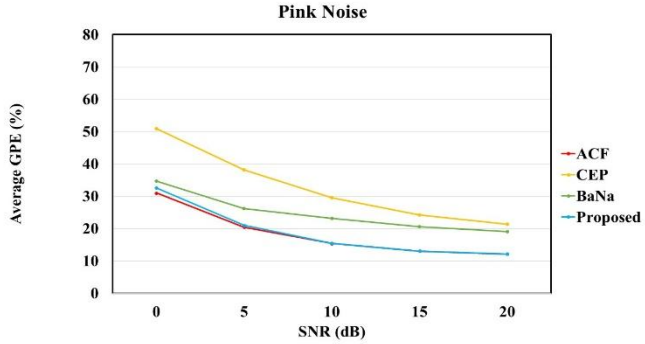

Figure 9. Average GPE Comparison for both Female Speakers and Male Speakers with Pink Noise across various SNR levels within KEELE database

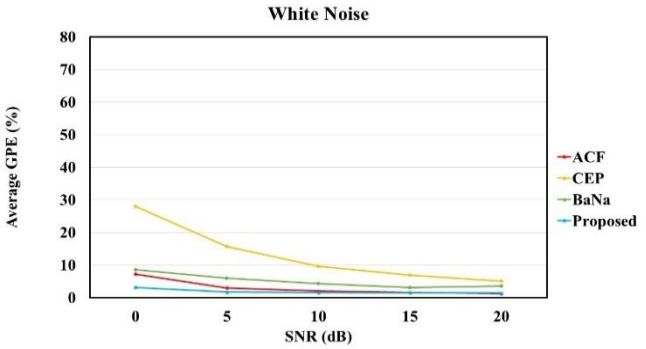

Figure 10. Average GPE Comparison for both Female Speakers and Male Speakers with White Noise across various SNR levels within NTT database

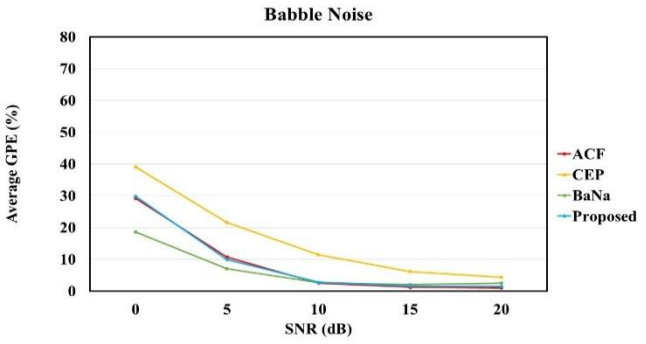

Figure 11. Average GPE Comparison for both Female Speakers and Male Speakers with Babble Noise across various SNR levels within NTT database

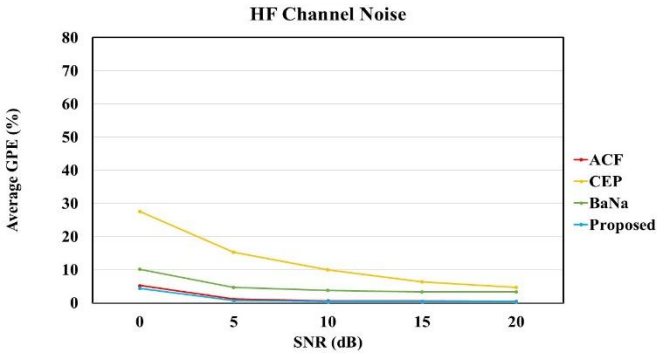

Figure 12. Average GPE Comparison for both Female Speakers and Male Speakers with HF Channel Noise across various SNR levels within NTT database

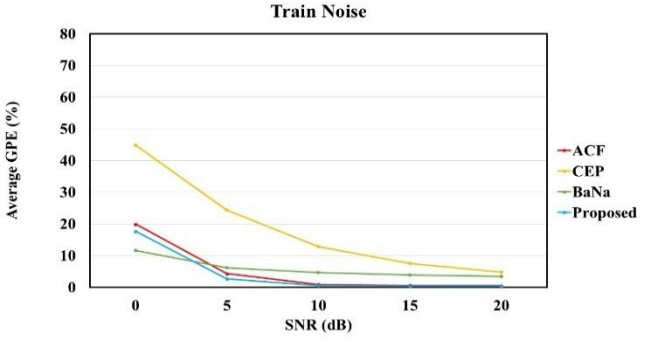

Figure 13. Average GPE Comparison for both Female Speakers and Male Speakers with Train Noise across various SNR levels within NTT database

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 730-734 | e-ISSN: 2347-470X

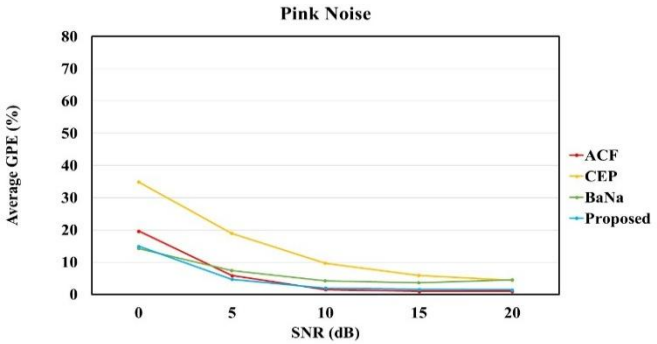

Figure 14. Average GPE Comparison for both Female Speakers and Male Speakers with Pink Noise across various SNR levels within NTT database

5. CONCLUSIONS

This paper introduced a novel method for extracting the fundamental frequency (F_0) from noisy speech by combining features from both the power and logarithmic spectra. The method captures periodicity more effectively by enhancing harmonic structures through spectral multiplication, thereby minimizing the influence of noise and vocal tract effects. The proposed algorithm was evaluated on the KEELE and NTT databases across various noise types and SNR levels. Experimental results confirmed its superiority in reducing Gross Pitch Error (GPE) compared to established techniques (ACF, Cepstrum and BaNa). The method showed particularly strong performance in challenging environments with low signal-to-noise ratios and overlapping background noise. In conclusion, the approach offers a simple yet effective solution for accurate pitch estimation in noisy speech and outperforms existing methods without relying on complex post-processing or learning-based noise compensation.

Author Contributions: Conceptualization, Foujia Islam and Dr. Md. Saifur Rahman; methodology, Foujia Islam; software, Moinur Rahman; validation, Moinur Rahman, Nargis Parvin, Dr. Md. Tofael Ahmed and Dr. Dulal Chakraborty; formal analysis, Dr. Md. Saifur Rahman and Nargis Parvin; investigation, Moinur Rahman and Dr. Md. Saifur Rahman; resources, Dr. Md. Saifur Rahman; data curation, Foujia Islam; writing—original draft preparation, Moinur Rahman; writing—review and editing, Dr. Md. Saifur Rahman and Moinur Rahman; visualization, Foujia Islam; supervision, Dr. Md. Saifur Rahman. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] H. C. Mahendru, Quick review of human speech production mechanism, International Journal of Engineering Research and Development 9 (10) (2014) 48–54.

[2] C. Shahnaz, Pitch extraction of noisy speech using dominant frequency of the harmonic speech model (2002).

[3] L. Sukhostat, Y. Imamverdiyev, A comparative analysis of pitch detection methods under the influence of different noise conditions, Journal of voice 29 (4) (2015) 410–417.

[4] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, S. Bennett, Robust prosodic features for speaker identification, in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, Vol. 3, IEEE, 1996, pp. 1800–1803.

[5] A. G. Adami, R. Mihaescu, D. A. Reynolds, J. J. Godfrey, Modeling prosodic dynamics for speaker recognition, in: 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003. Proceedings.(ICASSP'03)., Vol. 4, IEEE, 2003, pp. IV–788.

[6] J. A. Moorer, The optimum comb method of pitch period analysis of continuous digitized speech (1973).

[7] Y. Medan, E. Yair, D. Chazan, Super resolution pitch determination of speech signals, IEEE transactions on signal processing 39 (1) (1991) 40–48.

[8] L. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE transactions on acoustics, speech and signal processing 25 (1) (1977) 24–33.

[9] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, H. Manley, Average magnitude difference function pitch extractor, IEEE Transactions on Acoustics, Speech and Signal Processing 22 (5) (1974) 353–362.

[10] T. Shimamura, H. Kobayashi, Weighted autocorrelation for pitch extraction of noisy speech, IEEE transactions on speech and audio processing 9 (7) (2001) 727–730.

[11] A. De Cheveigné, H. Kawahara, Yin, a fundamental frequency estimator for speech and music, The Journal of the Acoustical Society of America 111 (4) (2002) 1917–1930.

[12] S. Seneff, Real-time harmonic pitch detector, IEEE Transactions on Acoustics, Speech and Signal Processing 26 (4) (1978) 358–365.

[13] T. Sreenivas, P. Rao, Pitch extraction from corrupted harmonics of the power spectrum, The Journal of the Acoustical Society of America 65 (1) (1979) 223–228.

[14] M. Lahat, R. Niederjohn, D. Krubsack, A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech, IEEE transactions on acoustics, speech and signal processing 35 (6) (1987) 741–750.

[15] A. M. Noll, Cepstrum pitch determination, The journal of the acoustical society of America 41 (2) (1967) 293–309.

[16] H. Kobayashi, T. Shimamura, A modified cepstrum method for pitch extraction, in: IEEE. APCCAS 1998. 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings (Cat. No. 98EX242), IEEE, 1998, pp. 299–302.

[17] R. H. MAFM, M. S. Rahman, T. Shimamura, Windowless-autocorrelation-based cepstrum method for pitch extraction of noisy speech, Journal of Signal Processing 16 (3) (2012) 231–239.

[18] N. Yang, H. Ba, W. Cai, I. Demirkol, W. Heinzelman, Bana: A noise resilient fundamental frequency detection algorithm for speech and music, IEEE/ACM Transactions on Audio, Speech and Language Processing 22 (12) (2014) 1833–1848.

[19] S. Gonzalez, M. Brookes, Pefac - a pitch estimation algorithm robust to high levels of noise, IEEE/ACM Transactions on Audio, Speech and Language Processing 22 (2) (2014) 518–530. doi:10.1109/TASLP.2013.2295918.

[20] K. Kasi, Yet another algorithm for pitch tracking (yaapt) (2002).

[21] L. N. Tan, A. Alwan, Multi-band summary correlogram-based pitch detection for noisy speech, Speech communication 55 (7-8) (2013) 841–856.

[22] 20 countries language database, NTT Advanced Technology Corp., Japan (1988).

[23] F. Plante, G. Meyer, W. Ainsworth, A fundamental frequency extraction reference database, in: Proc. Eurospeech, 1995, pp. 837–840.

[24] A. Varga, H. J. Steeneken, Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communication 12 (3) (1993) 247–251.