FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

Error Signal Based Method for Real Time Pitch Extraction from Noisy Speech

Mohammad Forhad Hossain[1], Nargis Parvin[2], Md. Ohidul Islam[3], Md. Saifur Rahman[4], Pintu Chandra Paul[5], and Moinur Rahman[6]*

[1]*Department of ICT, Comilla University, Bangladesh; Email: forhad-ict-cou@stud.cou.ac.bd*
[2]*Associate Professor, Department of CSE, Bangladesh Army International University of Science and Technology, Bangladesh; Email: nargis.cse@baiust.ac.bd*
[3]*Assistant Professor, Department of Computer Science & Engineering, Atish Dipankar University of Science and Technology (ADUST), Bangladesh; Email: ohidul.islam@adust.edu.bd*
[4]*Professor, Department of ICT, Comilla University, Bangladesh; Email: saifurice@cou.ac.bd*
[5]*Assistant Professor, Department of ICT, Comilla University, Bangladesh; Email: pintu@cou.ac.bd*
[6]*Lecturer, Department of ICT, Comilla University, Bangladesh; Email: moinur.rahman@cou.ac.bd*

***Correspondence:** *Moinur Rahman, moinur.rahman@cou.ac.bd*

ABSTRACT- Pitch estimation efficiency always gets affected by two major phenomena. The first one is additive noise which by nature are not correlated with the source speech signal. The second one is the vocal tract effect which creates unnecessary peaks within the signal waveform, resulting inaccurate detection of the pitch value. This paper proposes a Modified Weighted Autocorrelation Function (MWAF) executed in the Linear Predictive Coding (LPC) residual domain, instead of directly executing over the noisy source speech signal. Converting the noisy source speech signal to LP residual helps eliminating the vocal tract effects and emphasizing the necessary effects of glottal excitation. The Autocorrelation Function (ACF) and the Circular Average Magnitude Difference Function (CAMDF) are directly applied on this residual. The proposed MWAF is then calculated by multiplying the ACF with the reciprocal of the CAMDF, which enhances the periodic components as well as reduces the noise induced false peaks. Experimental results were evaluated on the KEELE and NTT databases across White, Babble, Train and HF Channel noise conditions. The result shows that our proposed algorithm reduces the Gross Pitch Error (GPE) by an average of 40.34% compared to baseline methods such as BaNa, PEFAC and WAF. Furthermore, the algorithm achieves an overall computational gain of 53.44% compared to BaNa, PEFAC and WAF, offering a stable solution for pitch tracking in degraded acoustic conditions.

Keywords: Pitch, Linear Predictive Coding, Modified Weighted Autocorrelation, LPC Residual.

ARTICLE INFORMATION

Author(s): Mohammad Forhad Hossain, Nargis Parvin, Md. Ohidul Islam, Md. Saifur Rahman, Pintu Chandra Paul, and Moinur Rahman;
Received: 24/04/26; **Accepted**: 05/09/26; **Published**: 30/09/26;
E- ISSN: 2347-470X;
Paper Id: IJEER 2704B11;
Citation: 10.37391/ijeer.140318
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-14/ijeer-140318.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Human speech is broadly divided into voiced, unvoiced, and silent segments. Voiced sounds are generated by vocal cord vibration, but unvoiced sounds don't have this vibration [1]. The fundamental frequency of this vibration is perceived as pitch (F_0) [2]. When the speech starts being generated it faces an adverse effect of vocal tract filters. This effect distorts the harmonic peak of the waveform. Besides, additive natural noise mixes with the speech and distorts its periodicity. This means the vocal tract filter and additive noise together affect the shape and periodicity of the speech signal. That's why extracting the pitch directly from a raw noisy speech signal is very much challenging. Systems like speech synthesis, speech coder, telecommunication require extracting the pitch accurately in real time noisy environment. There exist a number of algorithms which analyse the noisy speech signal and try to extract the fundamental frequency. But their efficiency remains inadequate for real time applications [3, 4]. The main research gap is that no previous study has addressed both the effects of the vocal tract filter and additive noise before. This study utilizes the Linear Predictive (LP) residual signal to discard the vocal tract effects from the excitation source. Proposed method computes a Modified Weighted Autocorrelation Function (MWAF) integrating the ACF and the Circular Average Magnitude Difference Function (CAMDF) directly on the LP residual. This formulation provides a highly robust pitch estimation framework that reduces noise effects mathematically without relying on time consuming processing.

1.1. Literature Review

Pitch extraction methods are broadly categorized into time-domain and frequency-domain approaches [5]. Time-domain

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

algorithms are the Autocorrelation Function (ACF) [6], Average Magnitude Difference Function (AMDF) [7], Average Squared Mean Difference Function [8], Weighted Autocorrelation Function (WAF) [9] and YIN [10]. They perform well on clean speech but degrade significantly in noisy environments. ACF measures similarity between a signal and its delayed version. This produces peaks at pitch-period multiples. AMDF computes differences, yielding deep nulls at those same pitch-period multiples. ACF is vulnerable to vocal tract interference and AMDF generates false pitch candidates at harmonic multiples. WAF addresses these issues by weighting ACF with the inverse of AMDF. But its stability depends heavily on optimal weight selection. The YIN method enhances ACF accuracy but requires computationally expensive post-processing.

Frequency-domain methods attempt to mitigate vocal tract effects. The cepstrum (CEP) method [11] separates vocal cord excitation from vocal-tract filtering via the inverse Fourier transform of the logarithmic spectrum. While effective in noiseless conditions, CEP falters in noise. Further improvements like the ACF of the log spectrum (ACLOS) [12] and the windowless ACF-based CEP (WLACF-CEP) [13] demonstrate better noise resilience.

Recent advanced techniques include PEFAC [14], which is a frequency domain pitch extraction approach that uses the three highest pitch candidates in the amplitude spectrum. It calculates the ratio of the frequencies to extract the accurate pitch from the noisy speech signal. But it concentrates only on low SNRs. Another approach BaNa [15], takes the first five pitch candidates to separate the signal into frequency bands and measures how periodic or noisy each band is. Then it uses this information to make a more robust and noise-resistant pitch estimation. The defect of BaNa is, it still requires high processing time.

Deep learning-based approaches have recently shown strong performance in pitch estimation under noisy conditions. CREPE [16], for instance, uses deep convolutional neural networks to learn frequency representations directly from raw waveforms. DC-CRN [17] and LACOPE [18] jointly tackle speech enhancement and pitch tracking. Signal-processing methods have advanced as well, including Radon transform-based techniques that address formant interference [19] and RNN-based models that capture temporal pitch dynamics [20]. More recently, hybrid frameworks combining signal processing with neural networks have aimed to balance accuracy and efficiency [21]. DSP assisted neural models now achieve competitive performance at reduced complexity and latency [22]. Lightweight architectures extend this further to edge devices [23]. However, these approaches typically rely on large annotated datasets, incur substantial computational overhead. Also, they introduce latency, limiting their deployment in real-time and resource-constrained environments. These limitations motivate the development of efficient signal processing–based alternatives. *Table 1* summarizes how these methods compare.

Table 1. Critical Comparison Between Existing Pitch Extraction Algorithm

Method	Domain	Key Strengths	Key Weaknesses / Research Gap
ACF [6]	Time	High accuracy in clean speech; simple implementation.	Highly vulnerable to vocal tract formants; produces false peaks at harmonic multiples.
AMDF [7]	Time	Computationally inexpensive (uses differences instead of products).	Generates deep nulls at sub-harmonics; struggles heavily in low SNR environments.
WAF [9]	Time	Combines ACF and AMDF to suppress false candidates.	Stability is highly dependent on optimal weight selection, which fails in dynamic noise.
YIN [10]	Time	High accuracy on clean-to-moderate noise.	Best-candidate search and interpolation add cost; degrades under strong additive noise.
Log Spectrum [11–13]	Frequency	Explicit source/filter separation.	Effective mainly at low noise; window/lifter choices are themselves noise-sensitive.
PEFAC [14]	Frequency	Extracts accurate pitch from amplitude spectrum ratios; robust in low SNR.	Performance concentrated mostly on specific low SNR bounds; struggles with non-stationary noise.
BaNa [15]	Frequency	High noise-resistance using multi-band periodicity measurements.	Requires complex post-processing and high processing time, limiting real-time application.
Deep Learning [16-23]	Hybrid/NN	State-of-the-art accuracy; robust feature extraction directly from waveforms.	Requires large annotated datasets; introduces significant computational overhead and latency.

Linear Predictive Coding (LPC) [24] is a technique in digital speech processing based on the source-filter model. LPC predicts speech samples using a linear combination of its past speech samples which effectively models the resonances of the

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

vocal tract. This predicted sample actually represents the effect of vocal tract filter on the actual speech sample [25,26]. The mathematical difference between the actual speech sample and the LPC predicted speech sample is called the error signal or the residual signal. As the LPC successfully model the vocal tract, subtracting the predicted signal removes the vocal tract's influence from the original speech. The remaining value which is called the residual, actually represents the original necessary excitation source.

In this study, we utilize the LP residual signal to completely decouple vocal tract characteristics from the speech signal. By computing the ACF and the Circular Average Magnitude Difference Function (CAMDF) directly on this residual, the proposed framework seamlessly integrates their complementary strengths. Operating in the LPC-derived residual domain significantly suppresses vocal tract artifacts and noise, yielding a highly robust and accurate pitch estimation method for degraded speech environments. The novelty of the proposed algorithm in *section 2.4*.

2. MATERIALS AND METHODS

2.1. Materials

To implement and evaluate the proposed pitch extraction algorithm, speech signals from the KEELE Database [27] and NTT Database [28] were utilized. The KEELE database subset comprised speeches from 10 speakers (five male, five female), totaling approximately 5.5 minutes of audio, sampled at 16 kHz. This 5.5-minute audio was divided into 33,111 frames out of which 16,960 frames were active voiced frames. The NTT database subset consisted of approximately 10 seconds of Japanese phrases spoken by four male and four female speakers, sampled at 10 kHz with a 3.4 kHz band restriction. Total 80 seconds of audio under this database was divided into 8,820 frames out of which 4,637 frames were active voiced frames. No machine learning training protocol was required as the method is deterministic.

To simulate real-world degraded acoustic environments, various noise profiles were added to the clean speech signals. Computer-generated additive white Gaussian noise (AWGN) was injected by scaling a normal distribution interference signal. Additionally, babble and HF channel noises (originally sampled at 20 kHz [29]) and train noise (originally sampled at 8 kHz [30]) were utilized. The sampling rates of these environmental noises were resampled to 10 kHz and 16 kHz to match the NTT and KEELE databases, respectively. The algorithms were tested across Signal-to-Noise Ratio (SNR) levels ranging from -5 dB to 20 dB. The experimental parameters were configured as follows:

Frame Length: 50 ms (except for PEFAC and BaNa, which used 60 ms).

Frame Shift: 10 ms (except for PEFAC).

Window Function: Rectangular (except for PEFAC and BaNa, which used Hamming windows).

FFT Length: 2048 points for KEELE and 1024 points for NTT.

Pitch Search Range: f_{min} = 50 Hz to f_{max} = 400 Hz, representing the standard human fundamental frequency range.

2.2. Problem Formulation

The fundamental objective of pitch extraction in a noisy environment is to isolate the true glottal excitation frequency from a corrupted observation. Let the discrete-time clean speech signal be denoted as $s[n]$, which is contaminated by an uncorrelated additive environmental noise $v[n]$. The observed noisy speech signal $x[n]$ is modeled as:

$$x[n] = s[n] + v[n] \tag{1}$$

For a voiced speech frame, $s[n]$ exhibits quasi-periodicity with a fundamental period T_0, such that $s[n] = s[n + T_0]$. Additive noise $v[n]$ and vocal tract resonances severely mask this periodicity in the time domain. The problem formulation thus reduces to accurately estimating T_0 (and consequently the fundamental frequency $F_0 = \frac{F_s}{T_0}$, where F_s is the sampling rate) from $x[n]$. For optimization we will:
Systematically decouple the vocal tract filter parameters using LP residual and

Minimize the influence of $v[n]$ by maximizing the MWAF metric over the valid pitch lag search range $[\tau_{min}, \tau_{max}]$.

These optimizations will help us gain overall at least 35-40% more efficiency in pitch extraction as well as an average of 40-50% more computational efficiency then other conventional methods.

2.3. Proposed Methodology

The proposed methodology aims to estimate the pitch of the noisy speech signal $x[n]$ accurately. At first, the signal is divided into a number of short frames using window function as follows:

$$y_i[n] = w_{rec}\left[x[n + (i-1)S_{len}]\right] \tag{2}$$

Here, $y_i[n]$ is the i^{th} frame of the $x[n]$ after applying a rectangle window function. S_{len} is the frame shift; n represents the n^{th} sample of the frame. The rectangle window function w_{rec} maintains a value of 1 across its length and 0 elsewhere, effectively turns the signal "on" and "off" within a given interval. If any sample z is given in a rectangle window function, then it will give response as:

$$w_{rec}[z] = \begin{cases} z, & 1 \le z \le N \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

In equation 2 the value of z of is $x[n + (i-1)S_{len}]$. After that, we compute the autocorrelation of the frame as:

$$r[\tau] = \frac{1}{N}\sum_{n=1}^{N-\tau} y_i[n]y_i[n+\tau] \tag{4}$$

Where $r[\tau]$ denotes the autocorrelation function of the i^{th} frame at lag τ. Linear Predictive (LP) analysis is then applied from

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

equation (5) to *equation (10)* to model the vocal tract and the inverse filter generates the LP residual signal representing the excitation source.

At first, it constructs a Toeplitz matrix R from the first M values of $r[\tau]$:

$$R = \begin{bmatrix} r(1) & r(2) & \cdots & r(M) \\ r(2) & r(1) & \cdots & r(M-1) \\ \vdots & \vdots & \ddots & \vdots \\ r(M) & r(M-1) & \cdots & r(1) \end{bmatrix} \quad (5)$$

Then a shifted autocorrelation vector A of the $r[\tau]$ is defined as:

$$A = \begin{bmatrix} r(2) \\ r(3) \\ \vdots \\ r(M+1) \end{bmatrix} \quad (6)$$

Then, the Auto-Regressive (AR) coefficients are estimated using the Yule-Walker equations, which are derived as:

$$a = -R^{-1}A \quad (7)$$

In summary, it constructs a Toeplitz matrix R from the first M autocorrelation values and uses left division to solve the linear system against the transpose matrix A of the shifted autocorrelation vector. The negative sign relates to standard LPC convention. The estimated AR coefficients characterize the spectral envelope of the speech signal and are subsequently used to compute the linear prediction residual for further analysis. In MATLAB, the AR coefficients are computed as:

$$a = -toeplitz\big(r(1:M)\big)/r(2:M+1)' \quad (8)$$

After that, we predict from $(M+1)^{th}$ to N^{th} sample of the frame using the estimated AR coefficients. The value of each current speech sample is predicted based on a linear combination of its immediate past M samples. It takes the sum of product of the AR coefficients and the sequence of M immediate previous actual samples (extracted in reverse order). For instance, the predicted n^{th} sample will be:

$$p_i[n] = \sum_{j=1}^{M} a[j]y_i[n-j] \quad (9)$$

For $1 \le n \le M$ it's assumed that $p_i[n] = 0$. $p_i[n]$ is actually predicted for $M+1 \le n \le N$. The difference between the predicted value $p_i[n]$and the actual value $y_i[n]$ is called the LP residual or prediction error. It uses a plus sign while calculating the differences because the AR coefficients have already incorporated the negative sign earlier. So, the n^{th} sample of the residual corresponding to the i^{th} frame of the noisy speech signal is calculated as:

$$e_i[n] = p_i[n] + y_i[n] \quad (10)$$

Then we apply both the autocorrelation function and the circular average magnitude difference function (CAMDF) on e_i as:

$$ACF_i[\tau] = \frac{1}{N}\sum_{n=1}^{N-\tau} e_i[n]e_i[n+\tau] \quad (11)$$

$$CAMDF_i[\tau] = \frac{1}{N}\sum_{n=1}^{N} |e_i[n] - e_i[(n+\tau) mod N]| \quad (12)$$

Where $ACF_i[\tau]$ and $CAMDF_i[\tau]$ denotes the autocorrelation function and circular average magnitude difference function of the i^{th} residual at lag τ respectively. Finally, using the ACF and the CAMDF we calculate the modified weighted autocorrelation function (MWAF) that can be expressed as:

$$MWAF_i[\tau] = \frac{ACF_i[\tau]}{CAMDF_i[\tau]+\delta} \quad (13)$$

Here, δ is a random constant to avoid the divide by zero error. For better understanding the equations used in this article, the mathematical notation is summarized in *Table2*

Table 2. List of Symbols and Mathematical Notations

Symbol	Definition	Domain / Unit
$x[n]$	Observed noisy speech signal	Discrete time
$s[n], v[n]$	Clean speech and additive noise components	Discrete time
k	The noisy speech will be sub divided into *k* frames	Integer
N, S_{len}	Frame length and frame shift	Samples
M	Linear Predictive Coding (LPC) order	Integer (12-16)
$y_i[n]$	i^{th} framed, windowed speech signal	$i \in [1,k]$ *and* $\in [1,N]$
$r[\tau]$	Autocorrelation function of the i^{th} speech frame	Lag τ
$a[j]$	j^{th}Auto-Regressive (AR) coefficient	Dimensionless
$p_i[n]$	LP predicted signal for frame i	$n \in [M+1, N]$
$e_i[n]$	LP residual (prediction error signal)	$n \in [1, N]$
$ACF_i[\tau]$	Autocorrelation function of LP residual	Lag τ
$CAMDF_i[\tau]$	Circular Average Magnitude Difference Function	Lag τ
$MWAF_i[\tau]$	Modified Weighted Autocorrelation Function	Lag τ
δ	Regularization constant	Dimensionless
F_s, F_0	Sampling rate and fundamental frequency	Hertz (Hz)

2.4. Novelty and Contribution

The standard Weighted Autocorrelation Function (WAF) mitigates some vulnerabilities of the conventional Autocorrelation Function (ACF) by utilizing the Average Magnitude Difference Function (AMDF) as a penalty term. Its

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

traditional formulation presents two significant mathematical limitations in low SNR environments. First, standard WAF operates directly on the windowed speech signal $y_i[n]$, which inherently contains highly energetic vocal tract formants. These formants introduce periodicities that mathematically mask the true glottal excitation. Second, the traditional AMDF suffers from a tapering effect at higher lags. Since the summation length decreases, this often forces the use of computationally heavy and empirically tuned weighting coefficients just to stabilize the function. What makes this work novel is the MWAF formulation itself. It mathematically circumvents both limitations through a domain transformation and a circular bounding strategy.

2.4.1. Domain Transformation *via* Spectral Flattening

Instead of computing the correlation metrics on $y_i[n]$, the proposed MWAF shifts the functional domain entirely to the LP residual signal $e_i[n]$. By extracting the Auto-Regressive (AR) coefficients a via the Yule-Walker equations, the vocal tract filter is decoupled from the source:

$$e_i[n] = y_i[n] + \sum_{j=1}^{M} a[j] y_i[n-j] \quad (14)$$

Operating on $e_i[n]$ effectively acts as a spectral whitener. Mathematically, this removes the resonant formant frequencies that cause false harmonic peaks in standard WAF. This ensures that any subsequent correlation analysis strictly evaluates the periodicity of the glottal source.

2.4.2. Circular Penalty and Unification

To address the tapering effect and reliance on heuristic weights in existing WAF models, MWAF integrates the ACF with the Circular AMDF (CAMDF). Unlike the standard AMDF, the CAMDF evaluates the absolute difference using modulo indexing as shown in *equation 12*. This circular nature guarantees that exactly N terms are evaluated for every lag τ, maintaining constant signal energy across the entire pitch search range. The proposed MWAF unites these domain-shifted functions into a stable, unified ratio, as shown in *equation 13*.

At the true fundamental period T_0, the numerator $ACF_i[T_0]$approaches a global maximum, while the denominator $CAMDF_i[T_0]$ approaches a deep null near zero. Because both functions are applied to the whitened residual $e_i[n]$, the probability of additive noise or formant interference simultaneously creating a false peak in the ACF and a false null in the CAMDF is statistically negligible. Consequently, the inverse proportional relationship strictly penalizes noise-induced artifacts and non-linearly amplifies the true F_0 peak without requiring the adaptive fractional weighting algorithms common in existing WAF literature.

2.5. Algorithm and Complexity Analysis

The proposed pitch estimation framework is visually represented in *figure 1*.

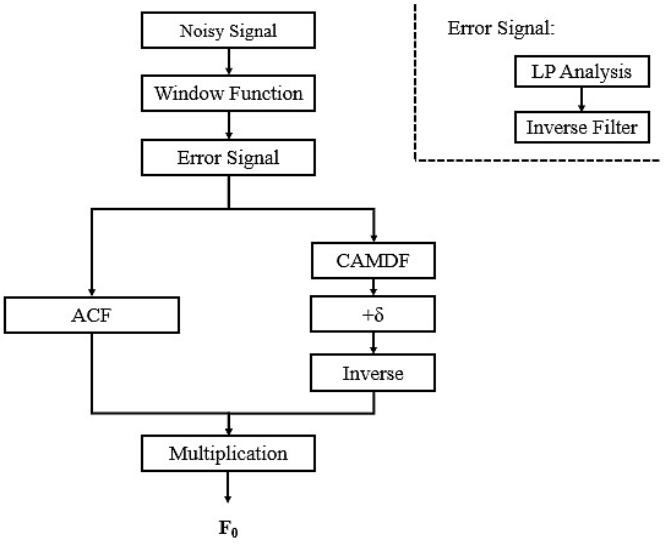


Figure 1. Block diagram of the proposed method

***Algorithm 1*: Proposed Error Signal based Pitch Extraction Algorithm**

Input: Noisy speech signal x(n), Sampling Frequency F_s

Output: Estimated Pitch F_0

BEGIN

Step 1: Frame x(n) into overlapping segments $y_i(n)$ using a rectangular window [Equation 2 & 3].

For each voiced frame $y_i(n)$:

Step 2: Compute autocorrelation $r(\tau)$ [Equation 4].

Step 3: Extract M AR coefficients from $r(\tau)$ via Yule-Walker equations [Equation 5,6 & 7].

Step 4: Estimate the predicted signal $p_i(n)$ [Equation 9].

Step 5: Calculate the LP residual: $e_i(n) = p_i(n) + y_i(n)$.

Branch A: Compute ACF of the residual: $ACF_i(\tau)$ [Equation 11].

Branch B: Compute CAMDF of the residual: $CAMDF_i(\tau)$ [Equation 12].

Step 6: Calculate MWAF: $MWAF_i(\tau) = \frac{ACF_i[\tau]}{CAMDF_i[\tau]+\delta}$.

Step 7: Identify the lag τ_{peak} corresponding to the global maximum of $MWAF_i(\tau)$.

Step 8: Calculate pitch: $F_0 = \frac{F_s}{\tau_{peak}}$

End For

END

To rigorously evaluate the computational efficiency of the proposed algorithm, a step-by-step analysis of time and space complexity is presented. Let L denote the frame length (number of samples per frame), M denote the Linear Predictive Coding (LPC) order, N denote the FFT length used for spectral analysis, and K denote the total number of active voiced frames processed. The step-by-step time complexity and space complexity of our algorithm is given in *table 3*.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

Table 3. Step-by-Step Time Complexity and Space Complexity of the Proposed Algorithm

Step	Method Name		Time Complexity	Space Complexity
1	Framing and Windowing		$O(L)$	$O(L)$
2	Autocorrelation of the Speech Frame		$O(N \log N)$	$O(N)$
3	LP Coefficient Extraction (Yule-Walker)		$O(M^3)$	$O(M^2)$
4 and 5	Predicted Signal and LP Residual Calculation		$O(LM)$	$O(L)$
6	Branch A:	ACF of the LP Residual	$O(N \log N)$	$O(N)$
	Branch B:	CAMDF of the LP Residual	$O(L^2)$	$O(L)$
7	MWAF Calculation		$O(L)$	$O(L)$
8	Peak Detection and Pitch Calculation		$O(L)$	$O(1)$

Combining these step-by-step bounds yields a per-frame time complexity of $O(L^2 + N \log N + M^3 + L.M)$. Given that the LPC order M (e.g., 16) is a small constant relative to L and N, the terms dependent on M have a negligible impact on runtime. Therefore, for an entire speech signal containing K voiced frames, the overall asymptotic time complexity of the algorithm simplifies to: $O(K.[L^2 + N \log N])$. During runtime, the peak memory footprint is determined by the allocation of the temporary arrays for the spectral operations (the N -point FFT buffers) and the Toeplitz matrix, alongside the global array storing the final estimated pitch values for the whole signal (N_{total}). Thus, the overall space complexity is constrained to: $O(N_{total} + N + M^2)$.

All experimental analyses and statistical validations were implemented in MATLAB R2021a. The analysis parameters were standardized to a frame size of 51.2 ms, a step size of 10 ms, and an LPC order of 16. To ensure full reproducibility of the reported results, the source code, parameter configurations, and evaluation scripts will be uploaded online after the publishing of this article.

3. RESULTS

3.1. Evaluation Criteria

The effectiveness of the suggested approach was evaluated by determining the error rate in accordance with Rabiner's rule [31], defined as:

$$E_r(z) = F_{cal}(z) - F_{true}(z) \quad (15)$$

Here z is the frame index, $F_{cal}(z)$ is the estimated fundamental frequency, and $F_{true}(z)$ is the ground-truth fundamental frequency for the z-th frame. If $|E_r(z)| > 10\%$ of $F_{true}(z)$, the estimation was classified as a Gross Pitch Error (GPE). The GPE percentage was calculated exclusively over the active voiced frames of the speech data. In particular, we recognized and assessed the voiced parts of words in terms of the basic frequency.

To assess the generalized performance and overall robustness of the proposed pitch extraction algorithm, a comprehensive statistical analysis was also conducted. The error rates across all evaluated conditions, encompassing White, Train, HF Channel, and Babble noise at all tested SNR levels, were pooled into a single dataset. One-tailed paired t-tests were then performed comparing the proposed method against WAF, PEFAC, and BaNa. To correct for multiple comparisons and strictly control the Type I error rate, a Bonferroni correction was applied, setting the threshold for statistical significance at $p < 0.0167$ $\left(\alpha = \frac{0.05}{3}\right)$.

3.2. Experimental Results

To demonstrate the robustness of the suggested method, particularly in high-noise environments where conventional algorithms degrade, we compared our results against the BaNa [15], PEFAC [14], and WAF [9] techniques. *Figure 2* to *figure 5* present the performance comparison for the KEELE database under white, babble, train, and HF channel noise conditions. *Figure 6* to *figure 9* present the corresponding performance comparison for the NTT database across the same noise conditions. Average GPEs were recorded at -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB SNRs for both male and female speakers combined.

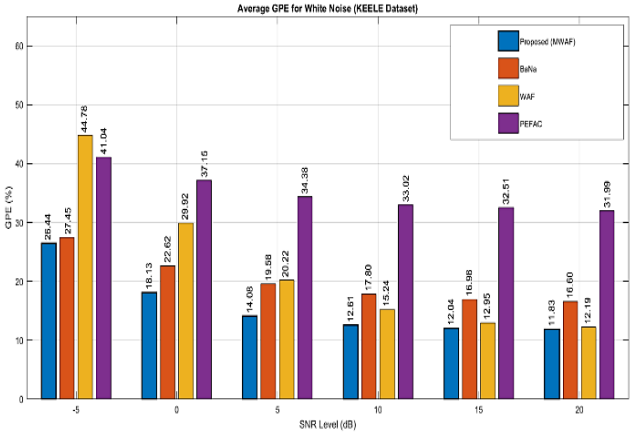


Figure 2. Average Gross Pitch Error (GPE) Rate for White Noise in KEELE Database

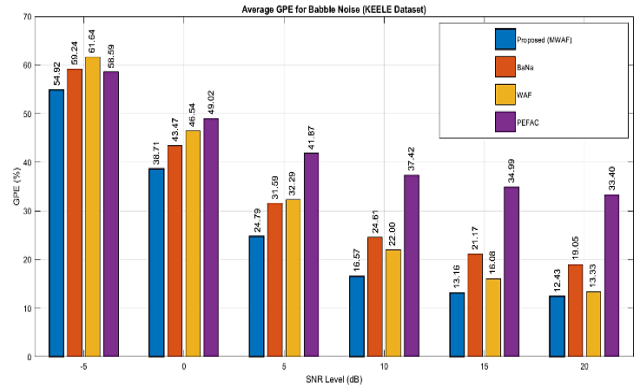


Figure 3. Average Gross Pitch Error (GPE) Rate for Babble Noise in KEELE Database

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

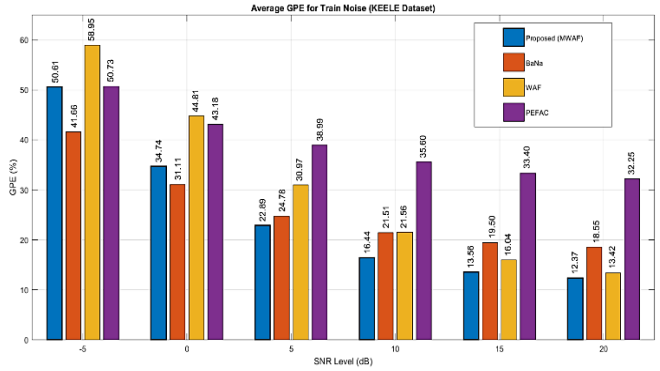


Figure 4. Average Gross Pitch Error (GPE) Rate for Train Noise in KEELE Database

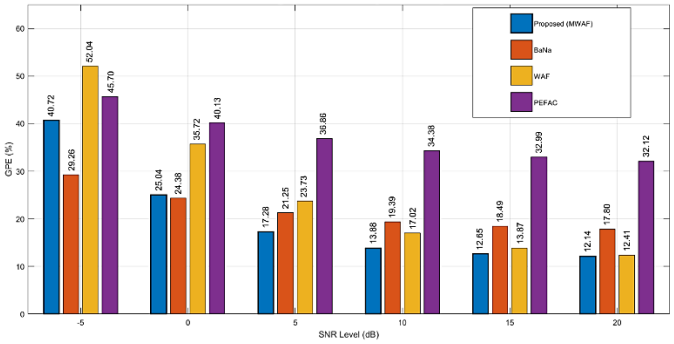


Figure 5. Average Gross Pitch Error (GPE) Rate for HF Channel Noise in KEELE Database

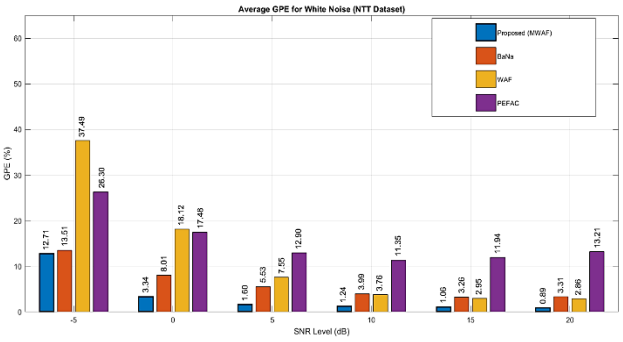


Figure 6. Average Gross Pitch Error (GPE) Rate for White Noise in NTT Database

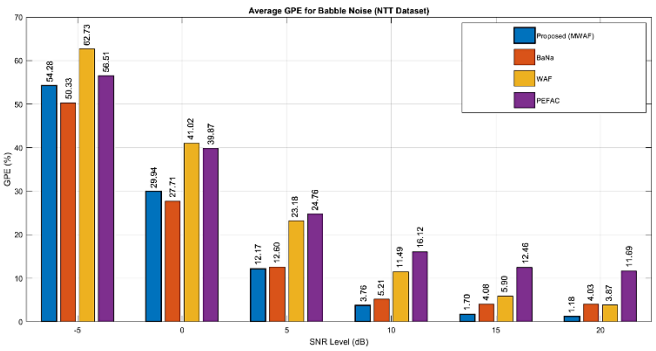


Figure 7. Average Gross Pitch Error (GPE) Rate for Babble Noise in NTT Database

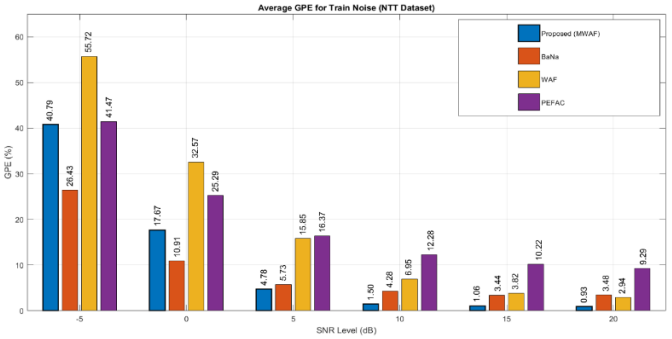


Figure 8. Average Gross Pitch Error (GPE) Rate for Train Noise in NTT Database

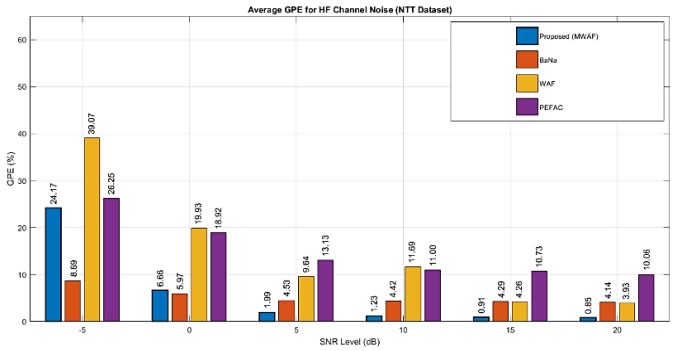


Figure 9. Average Gross Pitch Error (GPE) Rate for HF Channel Noise in NTT Database

3.3. Ablation Study

To isolate the individual contributions of LP residual extraction, CAMDF penalization and their unified integration in MWAF, an ablation experiment was performed. Three variants were evaluated, LP Residual Only, CAMDF Inverse Only and Proposed MWAF. The experiment was done on KEELE dataset in White and Babble noise at SNR level -5 dB. The results are shown in *table 4*.

Table 4. Average Gross Pitch Error (GPE) Rate for KEELE Dataset

Method	White Noise			Babble Noise		
	Female	Male	Average	Female	Male	Average
LPC_Residual	89.51992	80.95006	85.23499	89.69342	81.8463	85.76989
CAMDF	36.65336	33.4238	35.03858	59.9125	55.6489	57.78074
MWAF	26.36996	26.4339	26.40193	55.61714	54.16374	54.89044

4. DISCUSSION

4.1. Performance Comparison

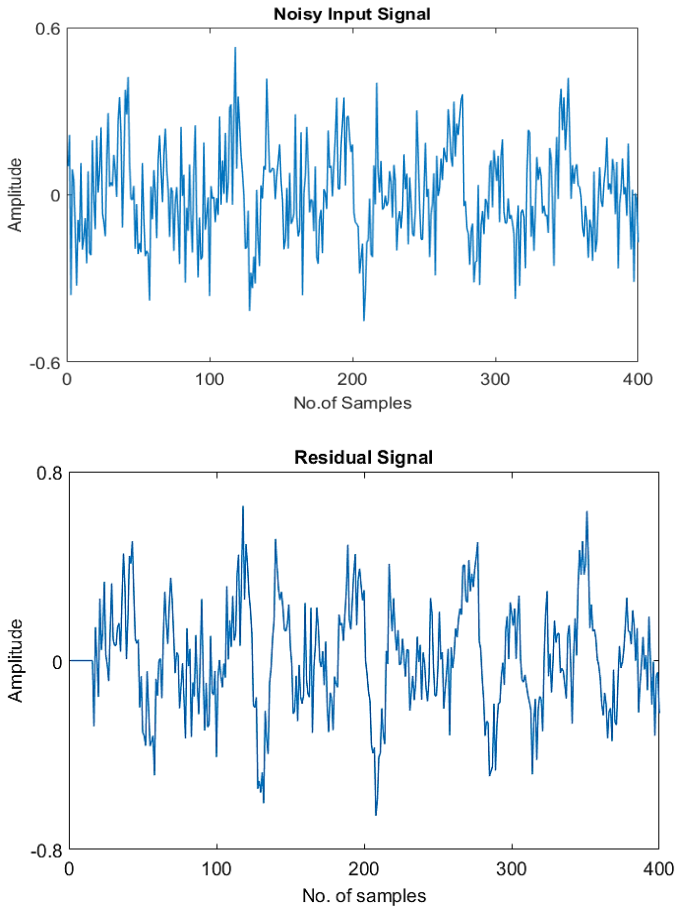
Considering the results of *figure 2 to figure 9*, our proposed method consistently exhibits the lowest average Gross Pitch Error (GPE) rate among all compared approaches in the presence of White noise and Babble noise in the KEELE Database across all SNRs. In the case of Train noise and HF Channel noise, BaNa exhibits a slightly lower GPE rate at SNR levels -5 dB and 0 dB. But at SNR levels 5 dB to 20 dB, our proposed method shows the best result compared with

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

others methods. In the context of White Noise in NTT Database, our proposed method exhibits the lowest average GPE rate among all compared. In the case of Babble noise, Train noise and HF Channel noise BaNa showed lower error rates particularly for -5 dB and 0 dB. But our proposed method showed lower error rates at 5 dB to 20 dB compared to other approaches. Crucially, the average GPE produced by the proposed MWAF method is optimally lower than the baseline methods in nearly all scenarios. The only exception occurs at extremely low SNR levels against BaNa in some noisy environments. However, this is because BaNa relies on highly complex post-processing specifically tuned for low SNR environments, making it significantly more computationally expensive and time-consuming than the proposed approach.

4.2. Technical Rationale for LP Residual Robustness

The raw noisy speech signal faces a major issue, the vocal tract effect. This introduces dominant formants masking the true fundamental frequency. LP residuals flatten the unnecessary vocal tract effect and enhance the glottal excitation effectively. The resulting error signal exhibits a much clearer, impulse-like periodicity than the raw noisy speech signal, as shown in *figure 10*. *Figure 10* also shows that, the weighted autocorrelation of this residual significantly minimizes formant-induced false peaks, leading to a higher resilience against non-stationary background noise.



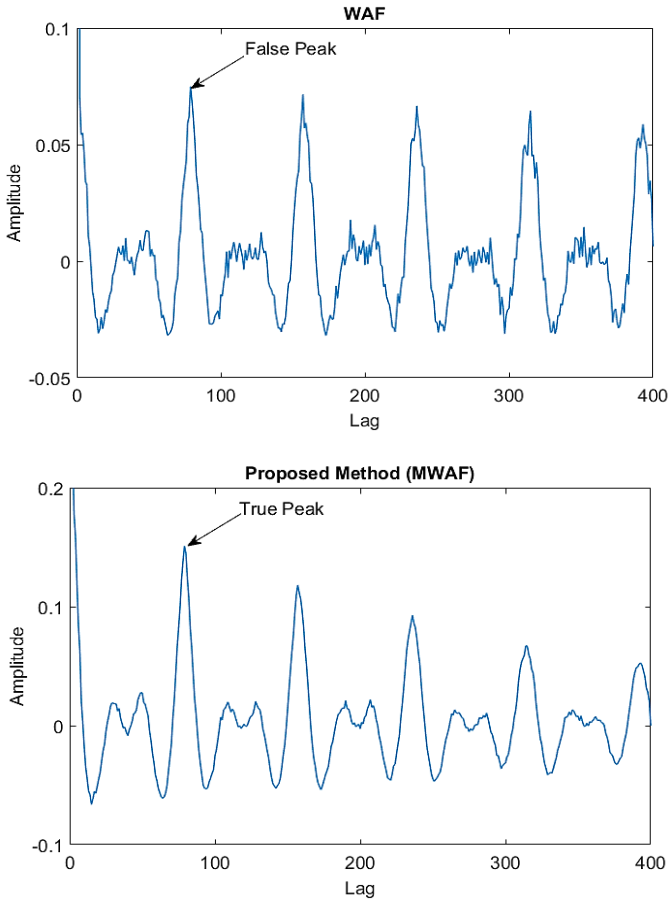


Figure 10. (a) Raw Noisy Speech Signal Having Much Uncorrelated Distortion; (b) LP Residual Signal Having Clearer Impulse Like Periodicity; (c) Response of conventional WAF Yielding False Peak; (d) Response of Our Proposed Method Yielding True Peak

4.3. Statistical Significance of Overall Performance

The overall statistical evaluation confirmed that the proposed method yields a generalized, robust improvement in pitch extraction across diverse noisy environments. The result is shown in *Table 5.*

Table 5. Statistical Significance of Overall Performance (Proposed MWAF vs. Baselines)

Comparing Method	Degrees of Freedom (df)	Statistic (t)	p-value	Bonferroni Threshold (α)	Statistically Significant?
WAF	47	-9.18	$p < 0.001$	0.0167	Yes
PEFAC	47	-15.03	$p < 0.001$	0.0167	Yes
BaNa	47	-2.20	0.0162	0.0167	Yes

Here, the evaluation shows that the proposed MWAF method demonstrated highly significant reductions in error compared to WAF and PEFAC. Furthermore, when compared against the closest performing baseline BaNa, the proposed method achieved a *p*-value of 0.0162 (t(47) = -2.20). This falls strictly below the Bonferroni-corrected significance threshold of 0.0167. The improvement over BaNa is confirmed to be

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

statistically significant. These findings indicate that the modified weighted autocorrelation function offers a consistent, statistically robust advantage in extracting pitch across challenging noisy conditions.

4.4. Computational Time

Table 6 presents the computational time per second of speech for all methods in KEELE database. For generating the processing time, all methods have been used with 10th Gen Intel(R) Core (TM) i5-10500, where CPU clock speed is 3.10 GHz and memory is 8 GB. In order to calculate the processing time for accurate measurement, five trials for each approach have been taken into account.

Table 6. Average Processing Time in Seconds

BaNa	WAF	PEFAC	Proposed
19.5716	2.5508	1.3038	3.6354

PEFAC recorded the lowest execution time followed by WAF. The proposed MWAF achieved an 81.43% reduction in execution time compared to BaNa. As a result of using a big FFT size (2^{16} points), BaNa uses a high processing time. Additionally, he processing time needed for our proposed method is small enough to implement it in real time pitch extraction systems on low power devices.

4.5. Limitations

While the proposed MWAF method exhibits high robustness, its performance inherently depends on the accuracy of the LP analysis. The computational overhead while calculating the LP residual introduces extended runtime delay. In environments with extreme low-frequency non-stationary noise, the AR coefficients may occasionally model the noise rather than the vocal tract, leading to a degraded residual signal. Future work will investigate dynamic LPC order selection to mitigate this issue. Additionally, more optimization steps like down sampling the LP Residual can be taken to reduce the computation time. The current evaluation is constrained to single-speaker database recordings (KEELE and NTT) with standard sampling rates ($10\ to\ 16\ KHz$). Further validation is required for multi-talker overlapping speech and ultra-wideband audio ($48\ KHz$).

5. CONCLUSIONS

In this paper, we presented a highly robust pitch estimation algorithm utilizing a Modified Weighted Autocorrelation Function (MWAF) applied to the LPC residual signal. By shifting the core analysis from the raw speech signal to the LPC residual, the algorithm successfully neutralized the interfering effects of vocal tract formants. Furthermore, combining the ACF and CAMDF into a single, unified MWAF equation provided a distinct mathematical advantage, effectively emphasizing the true fundamental frequency while aggressively suppressing noise-induced artifacts. Based on the experimental data, we have verified that the proposed method surpasses existing baseline methods—such as BaNa, PEFAC, and WAF—across all tested SNR conditions under white, babble, train, and HF channel noise environments.

Author Contributions: Conceptualization, Md. Saifur Rahman and Moinur Rahman; methodology, Mohammad Forhad Hossain, Md. Saifur Rahman; software, Mohammad Forhad Hossain; validation, Nargis Parvin, Md. Saifur Rahman and Pintu Chandra Paul; formal analysis, Moinur Rahman, Md Ohidul Islam, Nargis Parvin; investigation, Nargis Parvin, Md Ohidul Islam; resources, Md. Saifur Rahman; data curation, Mohammad Forhad Hossain; writing—original draft preparation, Mohammad Forhad Hossain; writing—review and editing, Moinur Rahman, Md. Saifur Rahman; visualization, Mohammad Forhad Hossain; supervision, Moinur Rahman. All authors have read and agreed to the published version of the manuscript.

Funding Source: This research received no external funding.

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

REFERENCES

[1] Vary, P. and Martin, R., "Digital Speech Transmission: Enhancement, Coding and Error Concealment," John Wiley & Sons, New York, 2006, https://doi.org/10.1002/0470031743.

[2] Hess W. "Pitch determination of speech signals: algorithms and devices," in Springer Science & Business Media, Dec 6 2012, https://doi.org/10.1007/978-3-642-81926-1

[3] 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.

[4] D. Wang, C. Yu and J. H. L. Hansen, "Robust Harmonic Features for Classification-Based Pitch Estimation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 952-964, May 2017, doi: 10.1109/TASLP.2017.2667879.

[5] S. S. Upadhya, "Pitch detection in time and frequency domain," in International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, 2012, pp. 1-5, doi: 10.1109/ICCICT.2012.6398150.

[6] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection", IEEE Transaction on Acoustics, Speech, Signal Processing, Vol. ASSP-25, No. 1, pp. 24–33, 1977, doi: 10.1109/TASSP.1977.1162905.

[7] M. Ross, H. Shaffer, A. Cohen, R. Freudberg and H. Manley, "Average magnitude difference function pitch extractor," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 22, No. 5, pp. 353-362, 1974, doi: 10.1109/TASSP.1974.1162598.

[8] 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.

[9] 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

[10] A. De Cheveigne 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.

[11] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3, pp. 333–338, 1999, doi: 10.1109/89.759042.

[12] Kunieda N, Shimamura T, Suzuki J, "Pitch extraction by using autocorrelation function on the log spectrum," Electronics and

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 14, Issue 3 | Pages 842-851 | e-ISSN: 2347-470X

Communications in Japan, Part 3, Vol. 83, No.1, pp. 90–98, 2000, doi: 10.1002/(SICI)1520-6440(200001)83.

[13] Hasan MAFMR, Rahman MS, Shimamura T. "Windowless autocorrelation-based cepstrum method for pitch extraction of noisy speech," Journal of Signal Processing, Vol. 16, No. 3, pp. 231-239, 2012, doi: 10.2299/jsp.16.231.

[14] S. Gonzalez and M. Brookes, "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.

[15] N. Yang, H. Ba, W. Cai, I. Demirkol and W. Heinzelman, "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.

[16] J. W. Kim, J. Salamon, P. Li and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 161-165, doi: 10.1109/ICASSP.2018.8461329.

[17] Zhang, Y., Wang, H., Wang, D. "Densely-connected Convolutional Recurrent Network for Fundamental Frequency Estimation in Noisy Speech," Proc. Interspeech 2022, pp. 401-405, doi: 10.21437/Interspeech.2022-11156

[18] H. Schröter, T. Rosenkranz, A.N. Escalante-B, and A. Maier, "LACOPE: Latency-Constrained Pitch Estimation for Speech Enhancement," Proc. Interspeech 2021, pp. 656-660, 2021, doi: 10.21437/Interspeech.2021-633

[19] B. Li and X. Zhang, "A Pitch Estimation Algorithm for Speech in Complex Noise Environments Based on the Radon Transform," in IEEE Access, vol. 11, pp. 9876-9889, 2023, doi: 10.1109/ACCESS.2023.3240181.

[20] V. Këpuska and H. Elharati, "RNN-Based F0 Estimation Method with Attention Mechanism," MDPI Information, vol. 16, no. 12, 2024.

[21] T. T. Nway and T. Shimamura, "Weighted Autocorrelation with Convolutional Neural Network for Noisy Speech Pitch Estimation," 2024 IEEE 13th Global Conference on Consumer Electronics (GCCE), Kitakyushu, Japan, 2024, pp. 510-511, doi: 10.1109/GCCE62371.2024.10760946.

[22] K. Subramani, J. -M. Valin, J. Büthe, P. Smaragdis and M. Goodwin, "Noise-Robust DSP-Assisted Neural Pitch Estimation with Very Low Complexity," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 11851-11855, doi: 10.1109/ICASSP48485.2024.10447962.

[23] C. Y. Jeong, Y. Song, S. Shin, and M. Kim, "Efficient pitch-estimation network for edge devices," ETRI Journal, vol. 47, no. 1, pp. 112-122, 2025, doi: 10.4218/etrij.2023-0430.

[24] Rabiner, Lawrence R. "Theory and applications of digital speech processing." (2011).

[25] Fant, G. "Acoustic theory of speech production. the Hague, the Netherlands: Mouton & Co.", 1960, pp: 582-585.

[26] J. Makhoul, "Linear prediction: A tutorial review," in Proceedings of the IEEE, vol. 63, no. 4, pp. 561-580, April 1975, doi: 10.1109/PROC.1975.9792.

[27] Plante F, Meyer G, Ainsworth W, "A fundamental frequency extraction reference database," Proceedings of the Eurospeech, pp. 837–840, 1995, doi: 10.21437/Eurospeech.1995-191.

[28] 20 Countries Language Database, NTT Advanced Technology Corp., Jpn, (1988)

[29] Varga, A., Steeneken, HJ., 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.

[30] Itahashi, S.: Creating speech copora for speech science and technology, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences vol. 74, no. 7, pp. 1906 1910, 1991.

[31] Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C., 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.

© 2026 by the Mohammad Forhad Hossain, Nargis Parvin, Md. Ohidul Islam, Md. Saifur Rahman, Pintu Chandra Paul, and Moinur Rahman. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).