Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Spectral Subtraction Based Weighted Function for Pitch Extraction in Noisy Speech

Miss. Nargis Parvin[1], Jafrin Akter Jeba[2], Mousumi Hasan[3], Umma Sadia Tabassum EVA[4] and Md. Saifur Rahman[5] ✉

[134]Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla and Bangladesh

[25]Department of Information and Communication Technology (ICT), Comilla University, Cumilla, Bangladesh

Corresponding Author: Md. Saifur Rahman, **E-mail**: saifurice@cou.ac.bd

| ABSTRACT

Many speech-related works employ the pitch period as a crucial component. Speech signals are typically collected in challenging noisy settings for real-world projects. Therefore, it is now more important than ever for the algorithm to be noise resistant in order to estimate pitch accurately. However, when dealing with noisy speech files at a low signal-to-noise ratio (SNR) value, many state-of-the-art algorithms are unable to produce satisfactory results. In this work, a new noise-resistant pitch estimation algorithm based on spectral subtraction is presented, which uses a weighted function to lessen the impact of the vocal tract effect. Furthermore, to enhance the correlation between the pitch estimates and smoothen the pitch contours, we employ a weighted function that combines the spectral subtraction-based technique as the numerator and the circular average magnitude difference function (CAMDF) as the denominator. We have utilized two noisy speech databases using seven different kinds of recorded ambient noise, and we evaluated our system against three cutting-edge methods. The suggested method lowers the Gross Pitch Error (GPE) rate at practically all SNRs in white noise and performs best on the NTT and KEELE databases.

| KEYWORDS

Pitch Estimation, Spectral Subtraction, Weighted Function, Circular Average Magnitude Difference Function (CAMDF), Gross Pitch Error (GPE).

| ARTICLE INFORMATION

ACCEPTED: 01 October 2024 **PUBLISHED:** 26 October 2024 **DOI:** 10.32996/jcsts.2024.6.4.17

1. Introduction

When someone speaks, their vocal folds vibrate, and the duration between each vibration and the opening and closing of the folds indicate the pitch. Pitch can be defined in terms of pitch period, and its inverse is called fundamental frequency, based on the quasi-periodic nature of vocal fold vibration. Individuals with high fundamental pitches also tend to speak with high tones in their voices. The fundamental frequency of each person differs depending on their unique traits. The lowest fundamental frequency that males can attain is about 50 [Hz], whereas females and children can reach up to 500 [Hz].

We also have a wealth of references for our research in these many fundamental frequency bands. There are many speech-related works that can benefit from using the pitch period as a crucial piece of information. Pitch augmentation in the frequency domain is used by [Park et al., 2010] to improve the comprehensibility of loud speech. Buera et al. [2008] build long-term speech and noise models for improving human speaking using the pitch period. Additionally, useful data for automated speech recognition (ASR) systems is the pitch period. In order to enhance speech recognition in a baseline ASR system, one study employs prosodic events, such as pitch accents [Ananthakrishnan, 2007]. By lessening pitch variation sensitivity, a different study creates a pitch-adaptive speech recognition system for kids [Sinha, 2008]. Accurate pitch information must be extracted from the speech in order to make the aforementioned applications more useful. However, obtaining a pitch from a speech presents a number of challenges. Firstly, there is a non-perfect periodicity in the speech signals generated by the voiced sounds [Cardozo, 1968]. The structures of the speech signals are then significantly altered as a result of these signals as they travel through the vocal tract and produce formants [Sukhostat, 2015].

2. Literature Review

Traditionally, pitch extraction methods have been based on characteristics of the speech signal, such as the frequency domain's harmonic structure or the time domain's periodic pattern. Voice signals in the time domain are subjected to pitch extraction algorithms, including the Autocorrelation Function (ACF), Average Magnitude Difference Function (AMDF), Circular Average Magnitude Difference Function (CAMDF), Weighted Autocorrelation Function (WAF), and YIN [Rabiner, 1977; Ross, 1977; Gang, 2003; Shimamura, 2001; De Cheveigne, 2002]. The most widely used pitch extraction method is the Autocorrelation Function (ACF), which calculates the nearest distance between two speech signal segments by comparing their levels of similarity. ACF is used to measure the correlation between the different voice input delay times. AMDF is another tool for analyzing correlations between different speech input delays. It does this by calculating the absolute value of the space between the current and lag speech signals. In WAF, ACF is weighted using the inverse of an AMDF. As a result, WAF effectively suppresses noise. By applying a function that ascertains variations in the speech signal's cumulative mean normalized squares, the YIN technique increases the accuracy of the ACF. It also makes use of post-processing methods. Pitch extraction methods based on ACFs perform well in noisy environments and are not impacted by the vocal tract effect.

Many methods of pitch extraction are being developed in the frequency domain, which is more effective against vocal tract characteristics. The cepstrum (CEP) method is one of the most well-known approaches [Noll, 1964; Ahmadi, 1999]. An inverse Fourier transform is used to the log-amplitude spectrum in order to generate the CEP.

Recently, two sophisticated methods have been explored [Gonzalez, 2014]; Yang, 2014]. In PEFAC, Gonzalez (2014) employs an amplitude compression method to improve its robustness in noise. The BaNa approach selects the top five spectral spikes in the amplitude spectrum of the speech data by taking into account noisy speech spikes [Yang, 2014]. Additionally, there are a number of pitch extraction methods that use the Hidden Markov Model (HMM) to measure the hidden patterns from sightings in order to create pitch tracks [Wang, 2017]. Another unique pitch extraction technique based on neurons and topologies in Neural Networks (NN) [Liu, 2017]. Another pitch estimator is suggested by using an acoustic filter bank to break speech signals into subbands under the time-frequency sparsity assumption [Lin, 2019]. Additionally, an encoding model is applied to the subband signals in order to obtain distinct and reliable fundamental frequency assessments of occasionally noisy and semi-periodic human sounds [Lin, 2018; Mnasri, 2022; Li].

To develop an efficient pitch extraction technique that is robust against noises and doesn't require laborious post-processing. ACF-based spectral subtraction of a speech signal is a potent tool for mitigating noise effects; it can accurately recover the pitch period in a noisy environment while still preserving the periodicity of the spoken signal. On the other hand, the ACF based subtraction approach is highly dependent on the features of the voice tract. As a result, we have utilized the weighted function with the combination of the above method and CAMDF. Therefore, the proposed approach is effective against noise and vocal tract characteristics. As a result, its extraction accuracy is increased without relying on post processing.

3. Methodology

In the case of speech signals, almost all researchers utilized the spectral subtraction method for speech enhancement. However, they did not concentrate on pitch extraction. Therefore, we have observed that the spectral subtraction based method is a powerful tool for pitch extraction. Assume that $S_{clean}(k)$ and $V_{noise}(k))$ represent the clean speech signal and noise, respectively. As a result, the noisy speech signal, $Y_{noisy}(k)$, has the following expression:

$$Y_{noisy}(k) = S_{clean}(k) + V_{noise}(k) \tag{1}$$

Figure 1 shows a block diagram of the spectral subtraction based method where ACF is used as a preprocessor. ACF method is used to reduce the noise from noisy speech signals. Then, the spectral subtraction method is applied to the noise free speech signal where the value of α is 1 and the average noise frame is considered as the value 3 [Nadika, 2024]. The spectral subtraction method is used as the numerator part of the proposed method.

On the other hand, we have considered Figure 2, which is represented as the block diagram of the proposed method. In the numerator part of the proposed method, we utilized the modified spectral based method, which is more effective against noise but not so good against vocal tract characteristics. As a result, we utilized the circular average magnitude difference function (CAMDF) as the denominator to reduce the noise and vocal tract characteristics. Finally, we have investigated whether the proposed method is more effective in extracting the accurate pitch period by reducing the noise effect and effect of vocal tract characteristics at low SNRs.

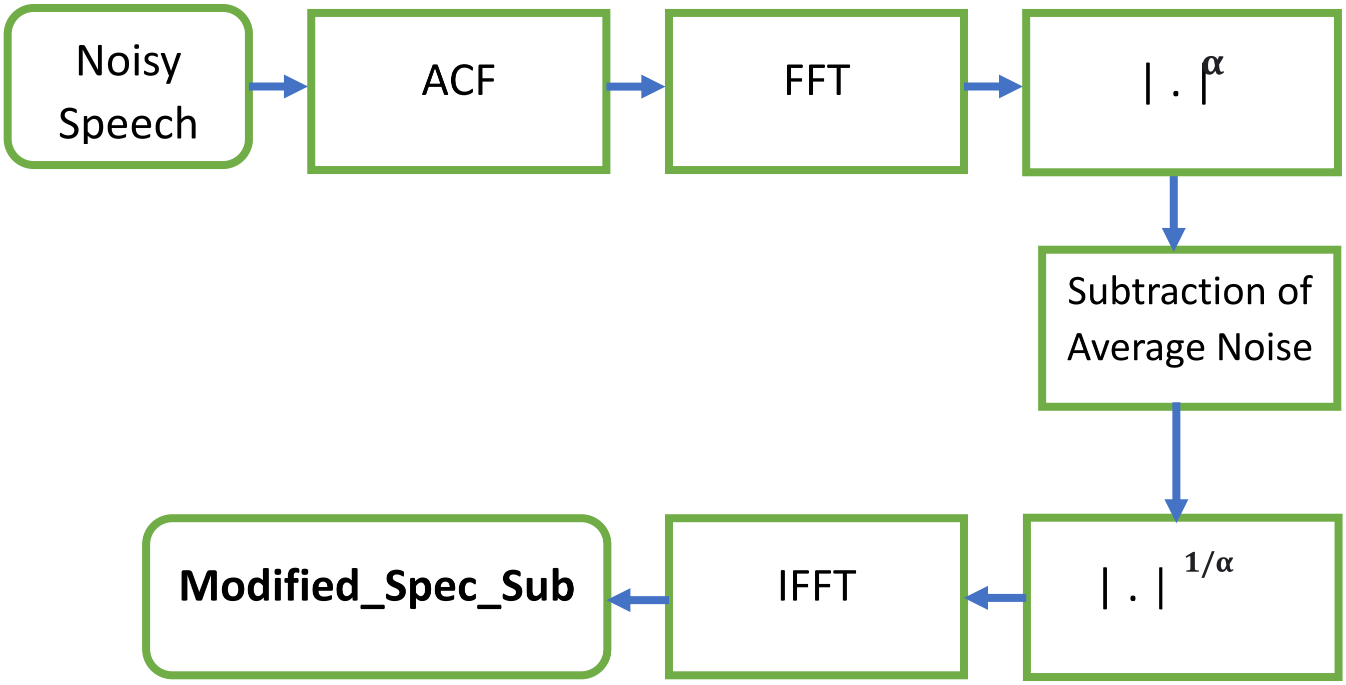

Fig. 1: Block diagram of spectral subtraction based method

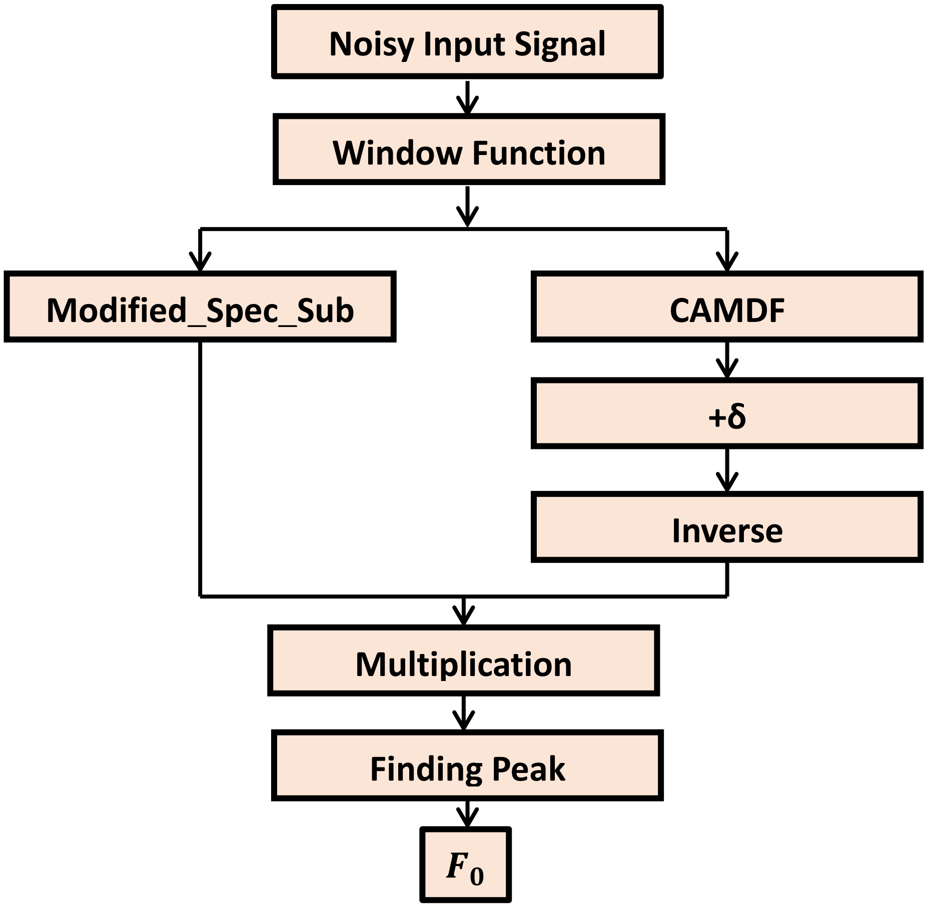

Fig. 2: Block diagram of the proposed method

4. Results and Discussion

For the validation of our proposed approach, we have utilized the speech signals collected from the NTT database [20 Countries Language Database, 1988] and the KEELE database [Plante, 1995]. The NTT database consists of four male and female speech signals, respectively. Around 10 seconds of Japanese phrases with a 10 [kHz] sampling rate and 3.4 [kHz] band limitation were included in each voice stream. On the other hand, With a sample rate of 16 [kHz], the voice signals from the 10 speakers in the KEELE database have a total length of roughly 5 [m] when voice signals are impacted by noise, resulting in noisy speech signals. The voice signals are combined with a computer-generated interference signal that has a normal distribution and is altered in magnitude. White noise is the name given to this interference signal. 8 [kHz] samples of train noise, 20 [kHz] samples of babble noise, and 20 [kHz] samples of HF channel noise. To analyze the signals from the NTT database and the KEELE database, the noises' sample rate was adjusted to 10 [kHz] and 16 [kHz], respectively. The following additional experimental parameters were used as the frame length, with SNR levels between -5 and 20 [dB] being used: The window function is rectangular, the frame shift is 10 [ms], and the time interval is 50 [ms] with the exception of PEFAC (90 [ms]) and BaNa (60 [ms]).

4.1. Evaluation Criteria:

This research can also use the Gross Pitch Error (GPE) rate, which has been used in numerous cutting-edge studies as a crucial assessment technique for figuring out how accurate pitch estimate algorithms are. The percentage of frames in the vocal segment with erroneous pitch periods is calculated to determine the GPE rate, which represents the algorithm's detection rate. Equation (2) illustrates that the accuracy of the method increases with decreasing GPE rate [7].

$$E\_r\ (z) = E_{est}(z) - E_{true}(z) \tag{2}$$

The variable z, which represents the frame number, $E_{est}(z)$ and $E_{true}(z)$ represent the extracted pitch and true pitch, respectively, of the $z-$th frame. When $|E_r(z)|$>10% of the $E_{true}(z)$, the error was determined to be Gross Pitch Error (GPE), and the proportion of this error was computed for the entire voiced frame in the speech data; for the basic frequency extraction, we only identified and assessed sentence parts that were vocal. We used the search ranges of $f_{min} = 50\ Hz$ and $f_{max} = 400\ Hz$ to extract the pitch.

4.2. Performance Comparison:

Pitch extraction in noisy environments was evaluated using the proposed approach (PROP) in comparison to the traditional methods (WAF, PEFAC, and BaNa). Here, we look into four different types of noise: HF channel, white, babble, and train noises. With the exception of the frame length, window function, and number of DFT (IDFT) points for PEFAC and BaNa, the factors of the proposed technique and the current techniques were identical. BaNa and PEFAC both made use of the hamming window feature. 2^{16} points were utilized for the DFT (IDFT) points, and the frame time for BaNa was set to 60 [ms]. This environment is perfect for BaNa, and that is where the source code of BaNa was implemented [Wcng, nd]. For PEFAC, the frame length was 90 [ms], and the window function was the Hamming window function. The DFT (IDFT) points are valued at 2^{13} in the source code. This is the perfect environment for BaNa, and that is where the PEFAC source code was implemented [Brookes, nd].

The Proposed methodology (PROP), with the exception of SNR=-5 [dB] at the HF channel noise, yields a lower GPE rate than the other standard methods (Spec-Sub, WAF, PEFAC, and BaNa) at practically all SNRs in the white and HF channel noises, as shown in Fig. 3 of the NTT database. Both the PROP approach and BaNa are quite competitive at the SNR level of -5 [dB]. However, the PROP methodology outperforms other traditional methods, notably at high SNRs of train noises and chatter (5 [dB] to 20 [dB]). The PROP methodology is competitive with current methods at low SNRs (-5 dB). Additionally, the KEELE database has been taken into consideration in order to test the effectiveness of the suggested strategy more precisely.

The average GPE rates for male and female speakers, respectively, are displayed in Figure 4. The original pitch values of the laryngograph signals can be found in the KEELE database. After examination, it was found that there are some gaps. The original pitch values are, therefore, not very accurate. The resulting GPE percentages clearly show this. The high SNR (20 dB) GPE percentage in Figure 4 is significantly higher than the high SNR (20 dB) in Figure 3. This results from the initial pitch values in the KEELE database being less precise. When it comes to performance comparison, all approaches in Fig. 4 show a tendency that is comparable to that in Fig. 3.

In the NTT and KEELE datasets, Figs. 5 and 6 show the average FPE performance characteristics, respectively. With the exception of low SNR (-5 dB), which is represented in Fig. 5 for NTT databases, the PROP approach yields a lower FPE than other conventional methods in practically all noise scenarios. The PROP approach and BaNa are very competitive at low SNR (-5 dB). However, in contrast to Fig. 5, Fig. 6 shows a contrary tendency. Specifically, BaNa's performance declines, and it performs worse than the PROP approach in every scenario. On the other hand, there is a nearly identical link between the WAF and PROP approaches' performance characteristics.

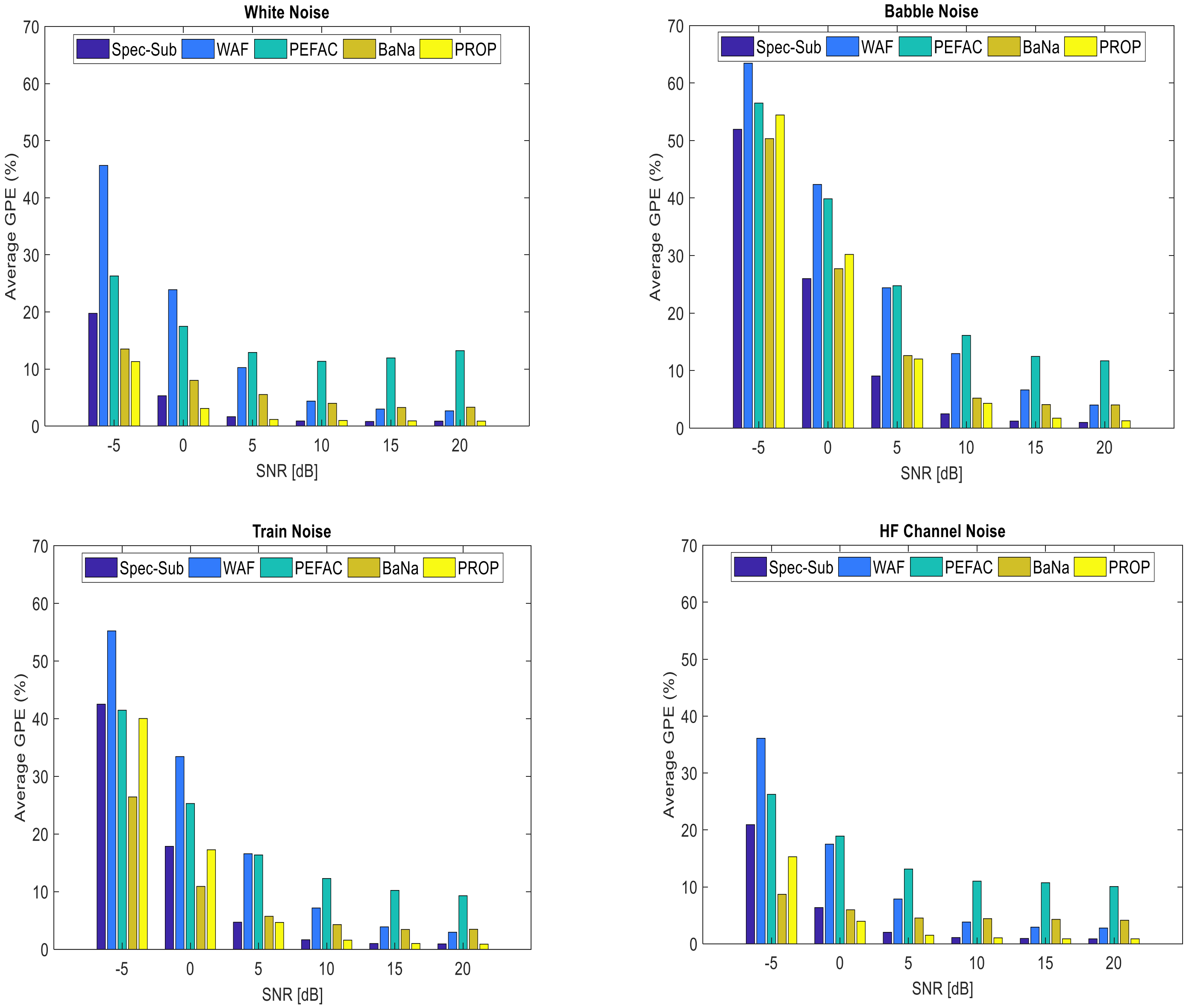

Fig. 3: Average gross pitch error (GPE)rate for different noise in NTT database

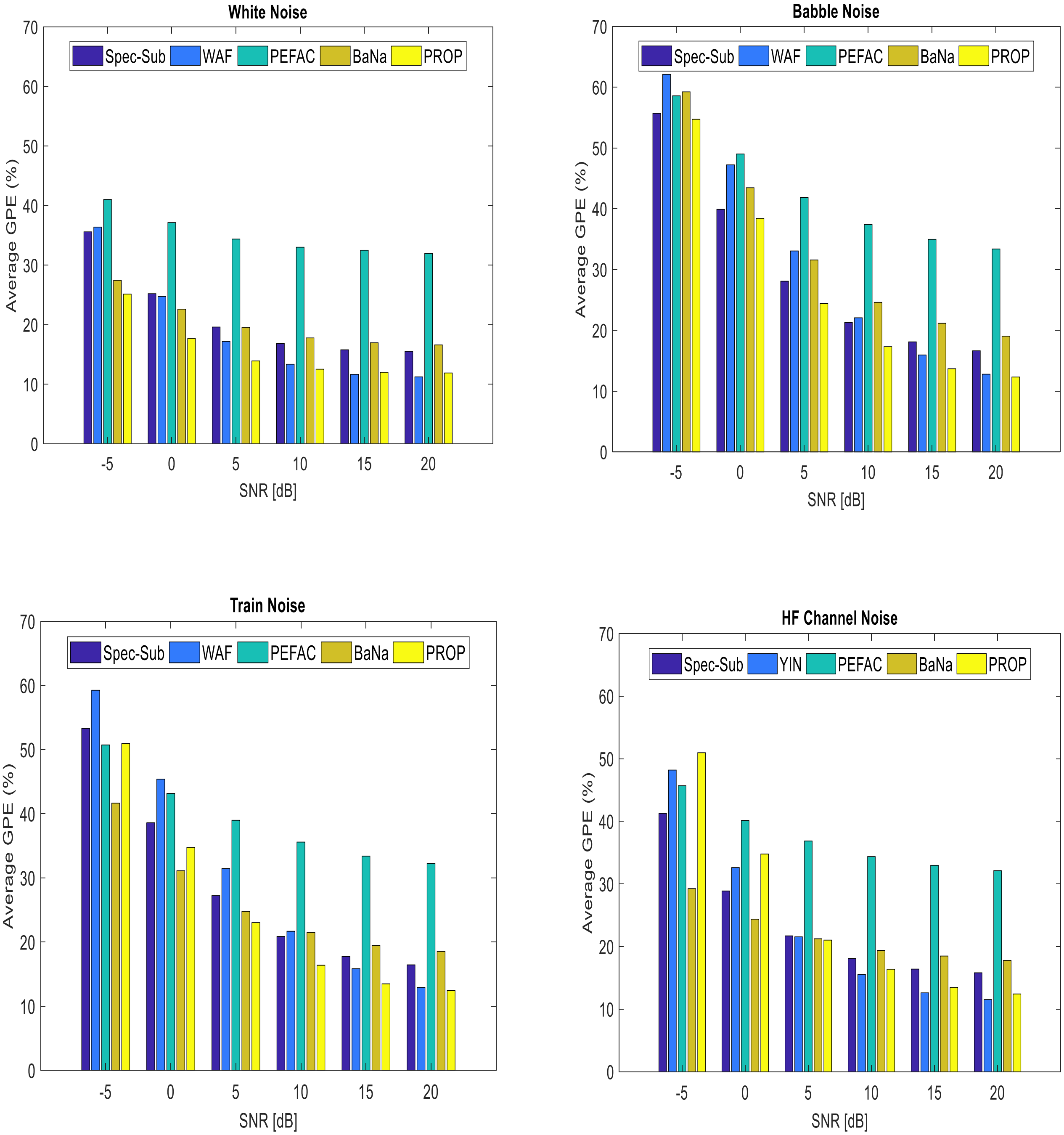

Fig. 4: Average gross pitch error (GPE)rate for different noise in KEELE database

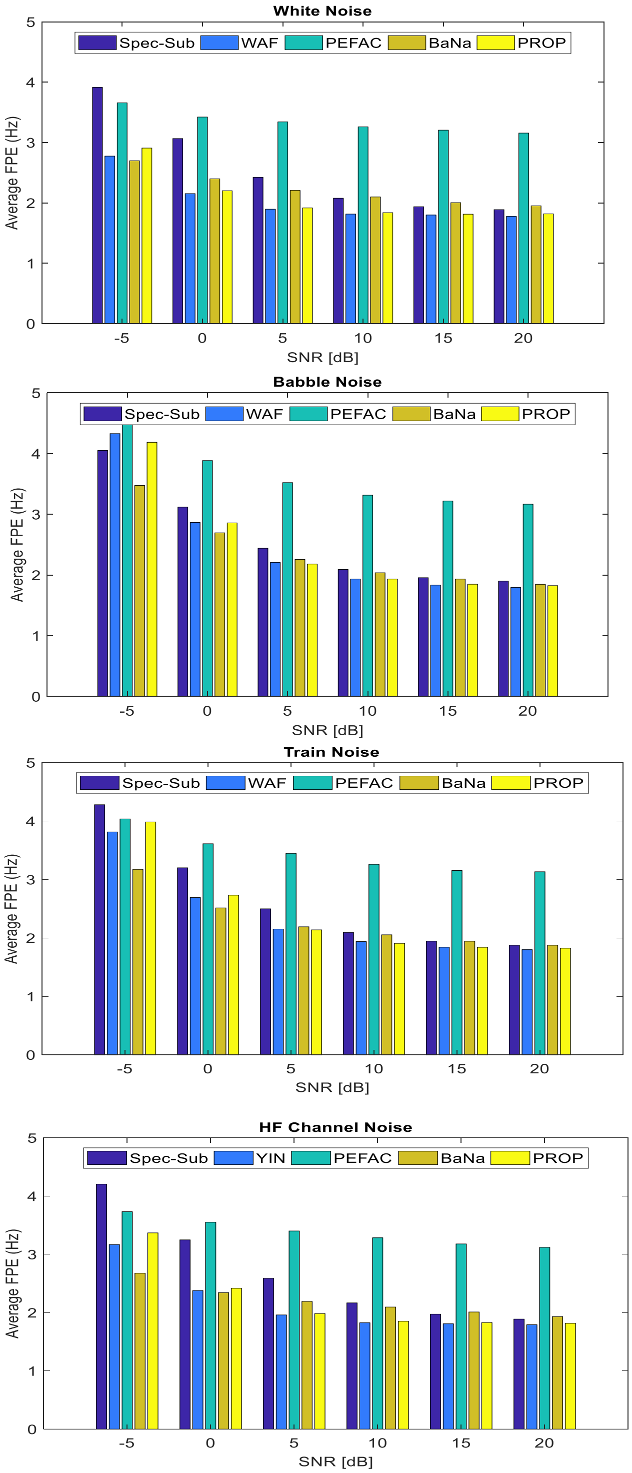

Fig. 5: Average fine pitch error (FPE) for different noise in NTT database

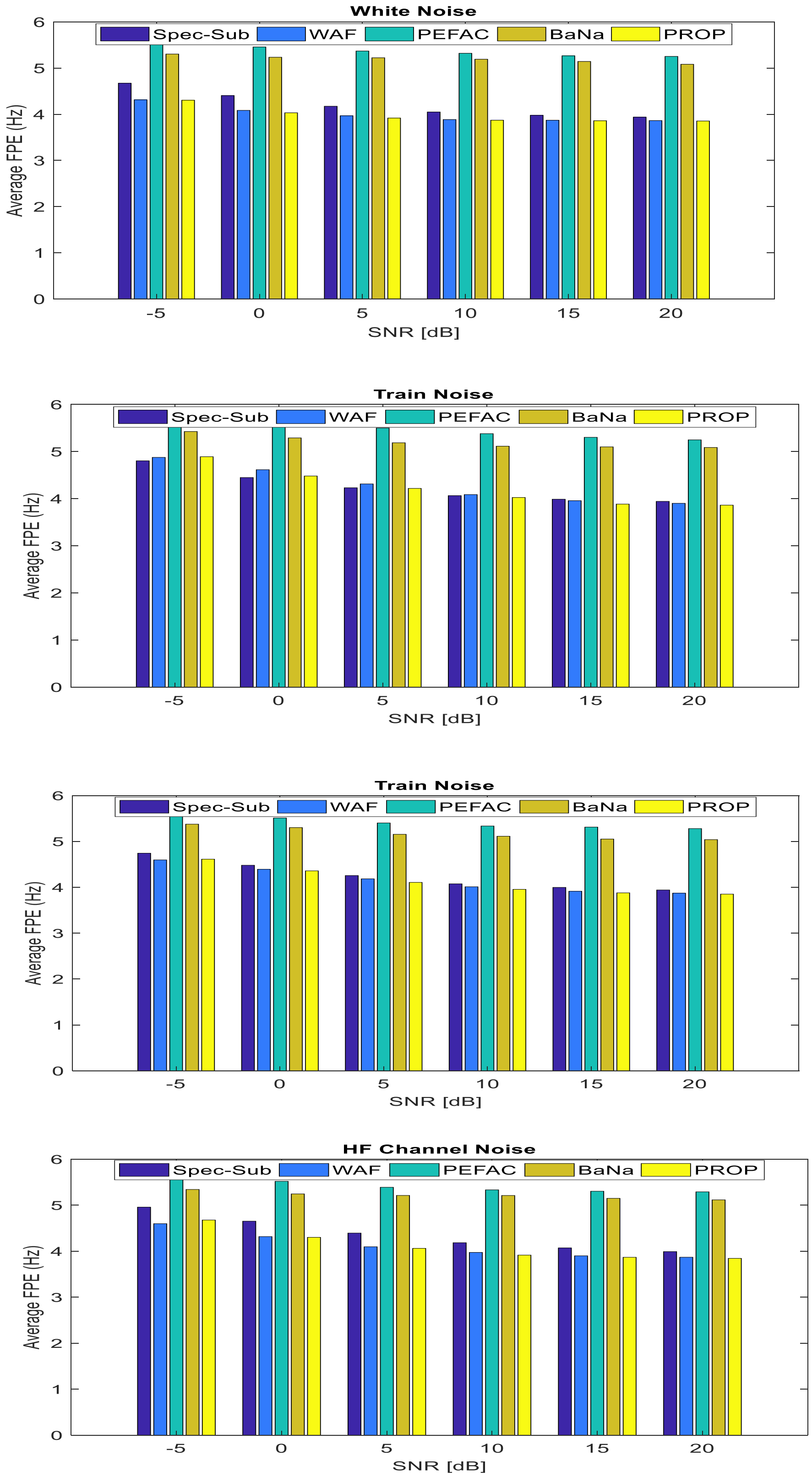

Fig. 6: Average fine pitch error (FPE) for different noise in KEELE database

5. Conclusion

This study presents the application of spectral subtraction to speech processing, along with a novel pitch estimation technique that can handle the intricate noise environments found in real-world projects. The approach that utilizes the weighted function yields a pitch that is more accurate from a speech frame that has been affected by noise compared to other simpler methods. We conduct an experimental comparison between the suggested algorithm's performance and that of WAF, PEFAC, and BaNa on two clean speech databases and the NoiseX-92 noise database. The proposed technique, according to the results, has the lowest GPE rate on the NTT and KEELE database at white noise under SNR values ranging from -5 [dB] to 20 [dB].

As a result, when the techniques we propose are appropriately applied while being aware of the noise kind and SNR degree, they will be extremely successful and efficient without requiring any laborious post-processing. Therefore, we plan to continue our study in the future to create new pitch extraction techniques that are especially resilient to extremely low signal-to-noise ratio situations in all authentic noise scenarios. The techniques that have been developed could be successfully used in a variety of speech applications.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ananthakrishnan, S., and Narayanan, S. (2007). Improved speech recognition using acoustic and lexical correlates of pitch accent in a N-Best rescoring framework. Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. IV-873–IV-876. https://doi.org/10.1109/ICASSP.2007.367209
[2] Ahmadi, S., Spanias, A. S. (1999). Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. IEEE Transactions on Speech and Audio Processing. 333–338. https://doi.org/ 10.1109/89.759042
[3] Buera, L., Droppo, J., and Acero, A. (2008). Speech enhancement using a pitch predictive model. Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 4885–4888. https://doi.org/10.1109/ICASSP.2008.4518752
[4] Brookes, M. (n.d). Voicebox toolkit, [Online]. Available, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[5] Cardozo, B., and Ritsma, R. (1968). On the perception of imperfect periodicity. IEEE Trans. Audio Electroacoustics, vol. AE-16. 159–164. https://doi.org/ 10.1109/TAU.1968.1161978
[6] De Cheveigne, A., Kawahara, H., Yin (2002). A fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*. 1917–1930. https://doi.org/ 10.1121/1.1458024
[7] Gang, X., and Liang-Rui, T. (2003). Speech pitch period estimation using circular AMDF. Proc. 14th IEEE Pers., Indoor Mobile Radio Commun. (PIMRC). 2452–2455. https://doi.org/ 10.1109/PIMRC.2003.1259159
[8] Gonzalez, S., Brookes, M., PEFAC. (2014). A Pitch Estimation Algorithm Robust to High Levels of Noise. IEEE/ACM Transactions on Audio, Speech, and Language Processing,. 518-530. https://doi.org/ 10.1109/TASLP.2013.2295918
[9] Liu, Y., and D. Wang, D. (2017). Speaker-dependent multi pitch tracking using deep neural networks. *The Journal of the Acoustical Society of America*. 710–721. https://doi.org/ 10.1121/1.4973687
[10] Lin, S. (2019). Robust Pitch Estimation and Tracking for Speakers Based on Subband Encoding and The Generalized Labeled Multi-Bernoulli Filter," IEEE/ACM Transactions on Audio, Speech, and Language Processing 827-841. https://doi.org/ 10.1109/TASLP.2019.2898818
[11] Lin, S. (2018). A new frequency coverage metric and a new subband encoding model, with an application in pitch estimation. Proceedings of Annual Conference of the International Speech Communication Association. 2147–2151. https://doi.org/ 10.21437/Interspeech.2018-2590
[12] Li, B., Zhang, X. (2023). A Pitch Estimation Algorithm for Speech in Complex Noise Environments Based on the Radon Transform. IEEE Access 11. 9876-9889. https://doi.org/ 10.1109/ACCESS.2023.3240181
[13] Mnasri, Z., Rovetta, S., Masulli, F. (2022). A Novel Pitch Detection Algorithm Based on Instantaneous Frequency for Clean and Noisy Speech. Circuits, Systems, and Signal Processing. 6266–6294. https://doi.org/10.1007/s00034-022-02082-8
[14] Noll, A. M. (1964). Short-time spectrum and cepstrum techniques for vocal-pitch detection. *The Journal of the Acoustical Society of America*. 296–302. https://doi.org/10.1121/1.1918949
[15] Nadika, R. T., Rahman, S., Parvin, N., Rahman, M., Rahman, M., and Chowdhury, N. (2024). Fundamental Frequency Extraction by Utilizing the Autocorrelation Based Spectral Subtraction Method in Noisy Speech. *International Journal on Communications Antenna and Propagation* (I.Re.C.A.P.). 42-50. https://doi.org/10.15866/irecap.v14i1.24562
[16] Park, H., Yoon, J. Y., Kim, J. H., and E. Oh, E. (2010). Improving perceptual quality of speech in a noisy environment by enhancing temporal envelope and pitch. *IEEE Signal Process*. Lett. 489–492. https://doi.org/10.1109/LSP.2010.2044937
[17] Plante F, Meyer G, Ainsworth W. (1995). A fundamental frequency extraction reference database," Proceedings of the Eurospeech, pp. 837–840. https://doi.org 10.21437/Eurospeech.1995-191
[18] Rabiner, L. (1977). On the use of autocorrelation analysis for pitch detection. IEEE Transactions on Acoustics, Speech, and Signal Processing. 24-33. https://doi.org/ 10.1109/TASSP.1977.1162905
[19] Ross, M., Shaffer, H., Cohen, A., Freudberg, R., Manley, H. (1974). Average magnitude difference function pitch extractor. IEEE Transactions on Acoustics, Speech, and Signal Processing. 353-362. https://doi.org/ 10.1109/TASSP.1974.1162598
[20] Sinha, R., and Shahnawazuddin, S. (2018). Assessment of pitch-adaptive frontend signal processing for children's speech recognition. Comput. Speech Lang. 103–121. https://doi.org/ 10.1016/j.csl.2017.10.007

[21] Sukhostat, L., and Imamverdiyev, Y. (2015). A comparative analysis of pitch detection methods under the influence of different noise conditions. *Journal of Voice*. 410–417. https://doi.org/ 10.1016/j.jvoice.2014.09.016

[22] Shimamura, T., Kobayashi, H. (2001). Weighted autocorrelation for pitch extraction of noisy speech. IEEE Transactions on Speech and Audio Processing. 727-730. https://doi.org/ 10.1109/89.952490

[23] Wang, D., Yu, C., and Hansen, J. H. (2017). Robust harmonic features for classification-based pitch estimation. IEEE/ACM Transaction on Audio, Speech, Language Processing. 952–964. https://doi.org/ 10.1109/TASLP.2017.2667879

[24] Wcng (n.d). Wireless communication networking group, [Online]. Available, http://www.ece.rochester.edu/projects/wcng/code.html.

[25] Yang, N., Ba, H., Cai, W., Demirkol, I., Heinzelman, W. (2014). BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 1833-1848. https://doi.org/ 10.1109/TASLP.2014.2352453

[26] 20 Countries Language Database, (1988). NTT Advanced Technology Corp., Jpn