Scientific Research Publishing

Journal of Signal and Information Processing, 2024, 15, 1-17
https://www.scirp.org/journal/jsip
ISSN Online: 2159-4481
ISSN Print: 2159-4465

Frame Length Dependency for Fundamental Frequency Extraction in Noisy Speech

Md. Saifur Rahman[1], Any Chowdury[1], Nargis Parvin[2], Arpita Saha[1], Moinur Rahman[3]

[1]Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh
[2]Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh
[3]Department of Information Technology, University of Information Technology and Sciences (UITS), Dhaka, Bangladesh
Email: saifurice@cou.ac.bd, anychowdhury1998@stud.cou.ac.bd, arpitasaha@stud.cou.ac.bd, nargis.cse@baiust.ac.bd, pranta7907@gmail.com

How to cite this paper: Rahman, M.S., Chowdury, A., Parvin, N., Saha, A. and Rahman, M. (2024) Frame Length Dependency for Fundamental Frequency Extraction in Noisy Speech. *Journal of Signal and Information Processing*, **15**, 1-17.
https://doi.org/10.4236/jsip.2024.151001

Received: December 18, 2023
Accepted: February 17, 2024
Published: February 20, 2024

Abstract

The fundamental frequency plays a significant part in understanding and perceiving the pitch of a sound. The pitch is a fundamental attribute employed in numerous speech-related works. For fundamental frequency extraction, several algorithms have been developed which one to use relies on the signal's characteristics and the surrounding noise. Thus, the algorithm's noise resistance becomes more critical than ever for precise fundamental frequency estimation. Nonetheless, numerous state-of-the-art algorithms face struggles in achieving satisfying outcomes when confronted with speech recordings that are noisy with low signal-to-noise ratio (SNR) values. Also, most of the recent techniques utilize different frame lengths for pitch extraction. From this point of view, This research considers different frame lengths on male and female speech signals for fundamental frequency extraction. Also, analyze the frame length dependency on the speech signal analytically to understand which frame length is more suitable and effective for male and female speech signals specifically. For the validation of our idea, we have utilized the conventional autocorrelation function (ACF), and state-of-the-art method BaNa. This study puts out a potent idea that will work better for speech processing applications in noisy speech. From experimental results, the proposed idea represents which frame length is more appropriate for male and female speech signals in noisy environments.

Keywords

Pitch Estimation, Fundamental Frequency, BaNa, ACF, Frame Length

1. Introduction

Speech is the expression of thoughts in spoken words, which is the main purpose of communication. Speech involves the production of sounds by the vocal tract, including the lungs, vocal cords, pharynx, mouth, and lips. Speech can be represented as silence, unvoiced, and voiced depending on whether the vocal cords are vibrating or not [1].

One of the most significant prosodic characteristics of speech is its fundamental frequency. In speech, the fundamental frequency (F_0) is the lowest frequency component of a complex sound wave produced during the vocal cords vibrating. The approximate frequency of the quasi-periodic structure of voiced speech signals is referred to as the fundamental frequency of a speech signal, which is sometimes indicated as (F_0). Individuals have distinctive frequency ranges, which can be influenced by various factors, such as linguistic context, gender, and age. It can also be affected by intonation, stress, emotion, and illness. Everybody has a different set of fundamental frequencies depending on how their voice chords are shaped. Men typically have a fundamental frequency range of 50 [Hz] to 250 [Hz], whereas women exhibit a range of 120 [Hz] to 500 [Hz] [2].

The fundamental frequency is determined by the rate at which the vocal cords vibrate when producing voiced speech sounds, resulting in the perception of pitch in speech. The fundamental frequency in the voiced speech sounds is also known as pitch, which is defined as our perception of a fundamental frequency. In speech, "pitch" refers to the perceptual characteristics that enable us to distinguish between the highness or lowness of a sound or tone as perceived by the ear. Accurate pitch (F_0) detection of a speech signal is essential for speech processing applications such as speech synthesis [3] [4], speech enhancement [5], speech recognition [6] [7] [8] [9], emotion identification [10]. [11] enhances the clarity of speech over noise by implementing pitch enhancement techniques within the frequency domain. [5] utilizes the pitch period to construct enduring models for both background noise and speech in the context of enhancing speech quality. One study employs prosodic events, specifically pitch accents, to enhance the performance of a baseline automatic speech recognition (ASR) system [6]. Another different research constructs a speech recognition system that is child-friendly, achieved by lowering its sensitivity to pitch variations [7].

Extraction of precise pitch data from the speech is essential to intensify the feasibility of the aforementioned applications. Nevertheless, numerous challenges persist in retrieving pitch information from speech. Detecting the precise F_0 becomes a difficult task in the presence of noise-contaminated speech signals [12]. Additionally, since a clean speech waveform [13], which undergoes significant changes in structure during vocal tract passage, pitch extraction has been challenging, even in a noise-free environment.

Prior to now, techniques for determining pitch have leaned on the distinct attributes of speech signals, whether it be the periodic pattern in the time domain [14] or the harmonic structure in the spectral domain [15].

Within the time domain, a diverse array of algorithms comes into play for extracting pitch from speech signals. These encompass methodologies like the Autocorrelation function (ACF) [16], Average magnitude difference function (AMDF) [17], Average squared mean difference Function [18], Weighted autocorrelation function (WAF) [19], Praat [20] and YIN [21].

Among the array of methods for pitch detection, the autocorrelation function reigns supreme. The autocorrelation function (ACF) [14] assesses the resemblance between two sections of a speech signal and pinpoints the period that yields the closest separation. The AMDF [17] is a simplified version of ACF that computes a signal by taking the average of the magnitudes of the difference between a signal and a delayed version of itself. WAF [19] is calculated through the multiplication of the signal's autocorrelation function with a designated range of weights. This approach weighs an autocorrelation function using AMDF's inverse. The WAF [19] can be used to eliminate noise and other unwanted signals from a signal during signal filtering. Praat [20] selects the best F_0 candidate for each short segment of the sound by looking at the maxima of the autocorrelation of the segment and applying the Viterbi algorithm to determine the least expensive path through all the segments. YIN [21] directs its attention to the interplay between the customary ACF and the difference function, utilizing a cumulative average approach to the difference function. This methodology aims to curtail inaccuracies in pitch extraction.

Pitch extraction methods rooted in ACF exhibit robustness in the presence of white noise and remain unaffected by phase irregularities in the waveform. Conversely, the efficacy of ACF-driven pitch extraction tends to diminish when clean speech encounters noise-induced influences, resulting in reduced performance. The behavior of the autocorrelation function is also susceptible to variations in the attributes of the vocal tract.

Within the frequency domain, a multitude of techniques for pitch extraction are devised to mitigate the impact of vocal tract characteristics. In this context, the quest for F_0 entails identifying harmonic peaks within the power spectrum.

Among the most often used methods is the cepstrum method (CEP) [22]. It derives the cepstrum by inverting the Fourier transform of the logarithmic magnitude of the Fourier spectrum. This encapsulates the period within speech harmonics, leading to an evident peak aligning with the frequency period. The logarithmic function in the CEP aids in segregating periodic attributes from the speech signal's vocal tract characteristics. The CEP yields precise outcomes within a noiseless environment, yet its effectiveness experiences notable degradation when confronted with the complexities of noisy conditions. The modified CEP (MCEP), as presented in [23], introduces additional steps involving liftering and clipping onto the logarithmic spectrum. This process serves a dual purpose: it eradicates vocal tract attributes while also erasing undesired spectral notches linked to noise in the log spectrum. Furthermore, the MCEP eliminates high-frequency components, thereby heightening the precision of pitch extrac-

tion. In the WLACF-CEP [24], the impact of noise on a noisy speech signal is mitigated, enabling its fusion with the CEP method to improve pitch extraction accuracy. The WLACF-CEP exhibits a notable capacity to withstand the challenges posed by an array of noise types. The technique known as pitch estimation filter with amplitude compression (PEFAC) [25] operates within the frequency domain for pitch detection. It employs sub-harmonic summations [26] within the logarithmic frequency domain. Additionally, PEFAC integrates a unique amplitude compression strategy aimed at enhancing its resilience against noise interference.

In the frequency domain, the harmonics do not adhere precisely to integer multiples of the fundamental frequency (F_0). Furthermore, the extent of drift is more pronounced among the higher-order harmonics compared to their lower-order harmonics. Consequently, it becomes necessary to establish a tolerance range to accommodate these variances when computing the ratios of harmonic frequencies.

In recent years, numerous strategies have been developed to enhance the effectiveness of overcoming and mitigating the impacts of background noise. [27] employs the Radon transform and proposes an innovative approach of pitch estimation for speech in challenging noise environments, integrating the Viterbi algorithm to smooth pitch patterns and mitigating the impact of formants using both logarithmic and power functions. [28] relies on introducing a practical connection between the fundamental frequency (F_0) and the instantaneous frequency (F_i). It approximates the F_0 contour as a smoothed envelope of remaining F_i values, identifying voiced or unvoiced speech regions and extracting the F_0 contour. The TAPS algorithm, as detailed in [29], involves the training of a collection of peak spectrum exemplars to estimate pitch by comparing clean and noisy speech data temporal accumulations. Chu and Alwan's SAFE [30] model aims to comprehend how noise impacts the positions and amplitudes in the clean speech spectrum. SPICE as described in [31] enhances pitch estimation by training trains constant Q transform of signals and calibrates the learned data for improved results. Deep F_0 [32] broadens the network's receptive range to encompass pitches across diverse noise levels. Harmo F_0 [33] has been shown to perform better than Deep F_0 by utilizing a variety of dilated convolutions, including multiple rates of dilated causal convolutions, for pitch estimation.

The BaNa [34] introduces a novel hybrid approach to pitch detection that amalgamates the concept of utilizing harmonic frequency ratios within predefined tolerance thresholds with the Cepstrum methodology. This fusion of techniques allows BaNa to effectively extract F_0 from noisy signals. The utilization of harmonic frequency ratios alongside meticulously adjusted tolerance ranges imparts robustness to the algorithm against the influence of additive noise.

From the above observation, we have investigated that most of the methods use different frame lengths for extracting the fundamental frequency in noisy environments. Like BaNa-60 [ms], RAPT-30 [ms], ACF-51.2 [ms], YIN-33 [ms], HMM-80 [ms], PEFAC-90 [ms] etc. However, a considerably longer frame

length is actively used in above state-of-the-art methods. Therefore, Speech harmonics have narrowing peaks when the frame length is extended. But nobody can ensure the accurate frame length where we can easily extract the fundamental frequency with high extraction accuracy for both male and female speech in noisy environments. In this research, we stress the usage of the frame lengths for male and female speech signals individually in order to increase the accuracy of pitch extraction in noisy situations, specially at low SNRs.

2. Methodology

In the case of fundamental frequency extraction from speech signal, all researchers have used different frame lengths with windowing technique for the purpose of segmentation. But, they didn't mention, which frame length is appropriate for segmentation to extract the accurate pitch peak in male and female speech signals. But there is also matter of consideration that there are differences between the male and female speech signal. Let's assume that the clean speech signal and the noise are, respectively, $s_{clean}(k)$ and $v_{noise}(k)$. The noisy speech signal, $y_{noisy}(k)$, can therefore be expressed as follows:

$$y_{noisy}(k) = s_{clean}(k) + v_{noise}(k) \tag{1}$$

Therefore, windowing, which is the process of separating a voice signal into periodic segments of a frame length, is a critical component of fundamental frequency extraction methods. To reduce the impact of splitting on the statistical properties of the signal, windowing functions can be applied to the temporal segments. Smoothing functions that go to zero at the borders are what windowing functions are. When a window function is applied to the input signal, the resulting segment of a frame length drops to zero at the edge, making the irregularity there undetectable. The following are some regularly used window functions for segmentation of a frame length, which is also represented as shown in Figure 1. The most basic type of window is rectangular, which achieves a similar effect by replacing all values in an input stream excluding N with zeros, creating the illusion that the signal is rapidly turning on and off. When the voice signal passes through this simple window, the values across the window are altered to zero. It is presented as follows: Hanning is one of the most commonly used random signal window function. The Hanning window eliminates ripple, allowing for a more accurate depiction of the frequency spectrum of the original signal. It is presented as follows:

$$win_{han}(k) = \begin{cases} 0.5 - 0.5\cos\left(2\pi k/(K-1)\right) & 0 \le k \le K-1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Utilizing the Hamming window, the structure is improved by lowering the largest (nearest) side lobe, giving it about one-fifth the height of the Hanning window. It is presented as follows:

$$win_{ham}(k) = \begin{cases} 0.54 - 0.46\cos\left(2\pi k/(K-1)\right) & 0 \le k \le K-1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

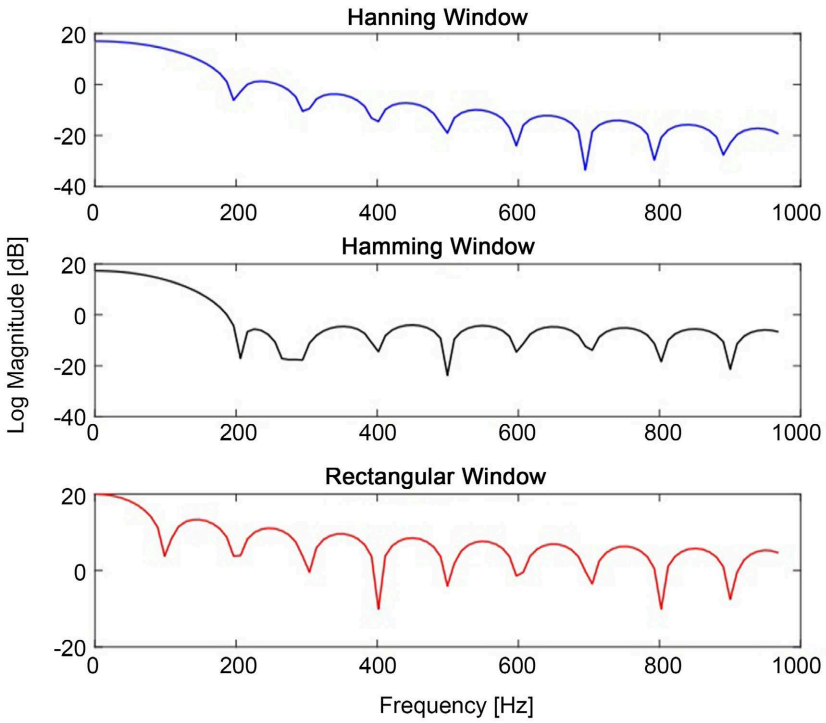

Figure 1. Magnitude spectrum of different window functions.

By utilizing the window function properties, most researchers employ different frame lengths to derive the pitch when getting the fundamental frequency from speech signals. In this research, we use four frame length for the situations of male and female speakers individually necessary to boost the accuracy of the detected pitch, as opposed to BaNa approaches, which calculate all of the frame length. Based on the foregoing discovery, we investigated 30 [ms], 50 [ms], 60 [ms] and 90 [ms] for pitch extraction in both male and female speakers in noisy situations, and we discovered more appropriate frame length for the male and female speakers separately.

Figure 2 shows that more noise has less of an influence on clean speech in the case of 50 [ms] for male speakers. We can see that the clean speech spectrum and the noisy speech spectrum are almost same, and the number of harmonics is more accurate than the other frame length. As a result, we receive more relevant pitch information. Almost all harmonics are lacking to discern the pitch peak in the case of 30 [m], 60 [ms], 90 [ms].So we can say that 50 [ms] is the accurate frame length for male speaker in pitch extraction.

On the other hand, **Figure 3** shows that more noise has less of an influence on clean speech in the case of 30 [ms] for female speakers. We can see that the clean speech spectrum and the noisy speech spectrum are almost same, and the number of harmonics is more accurate than the other frame length. As a result, we receive more relevant pitch information. Almost all harmonics are lacking to discern the pitch peak in the case of 30 [ms], 60 [ms], 90 [ms]. So we can say that 30 [ms] is the accurate frame length for female speaker in pitch extraction.

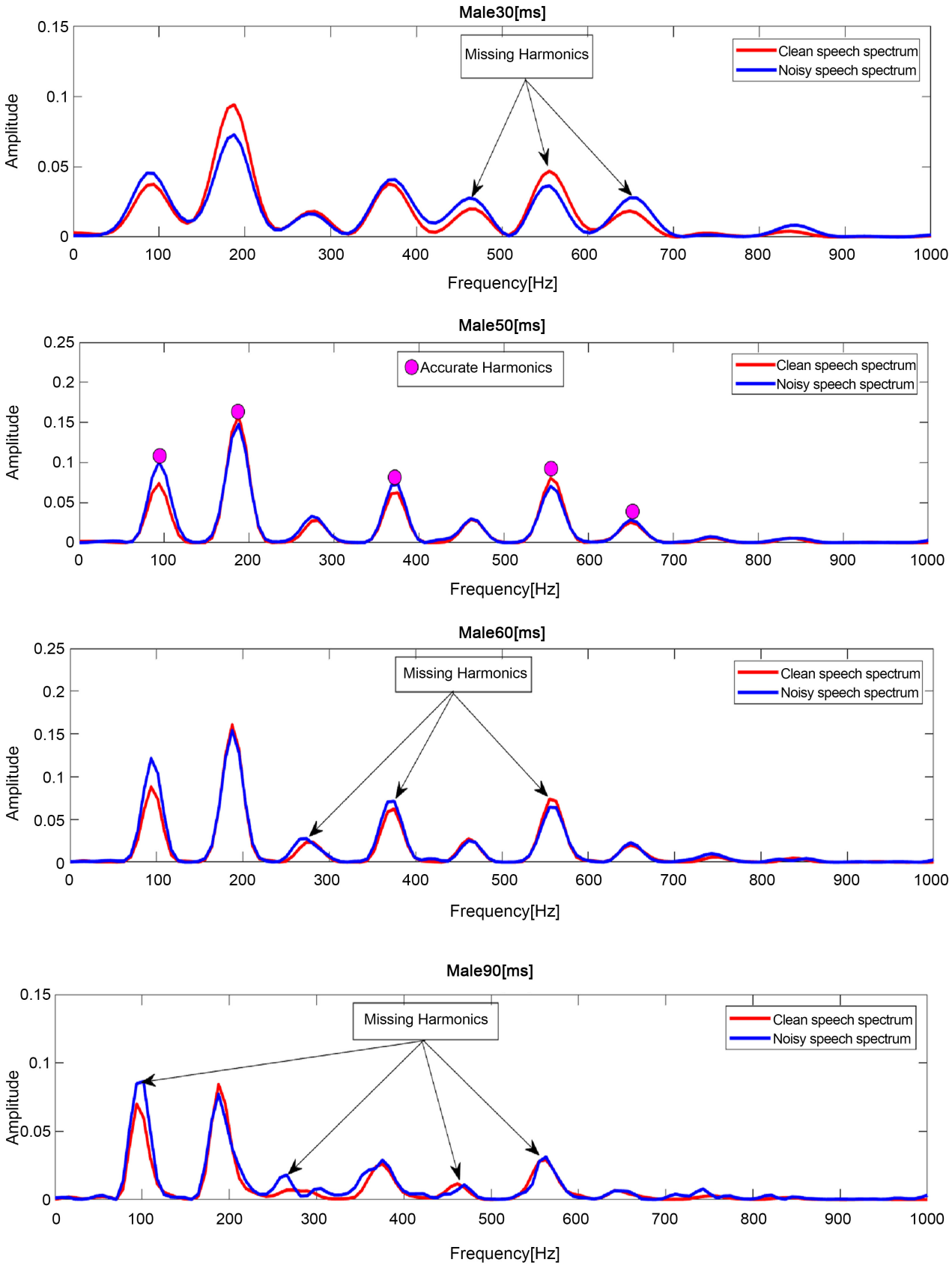

Figure 2. Harmonic characteristics for male speaker.

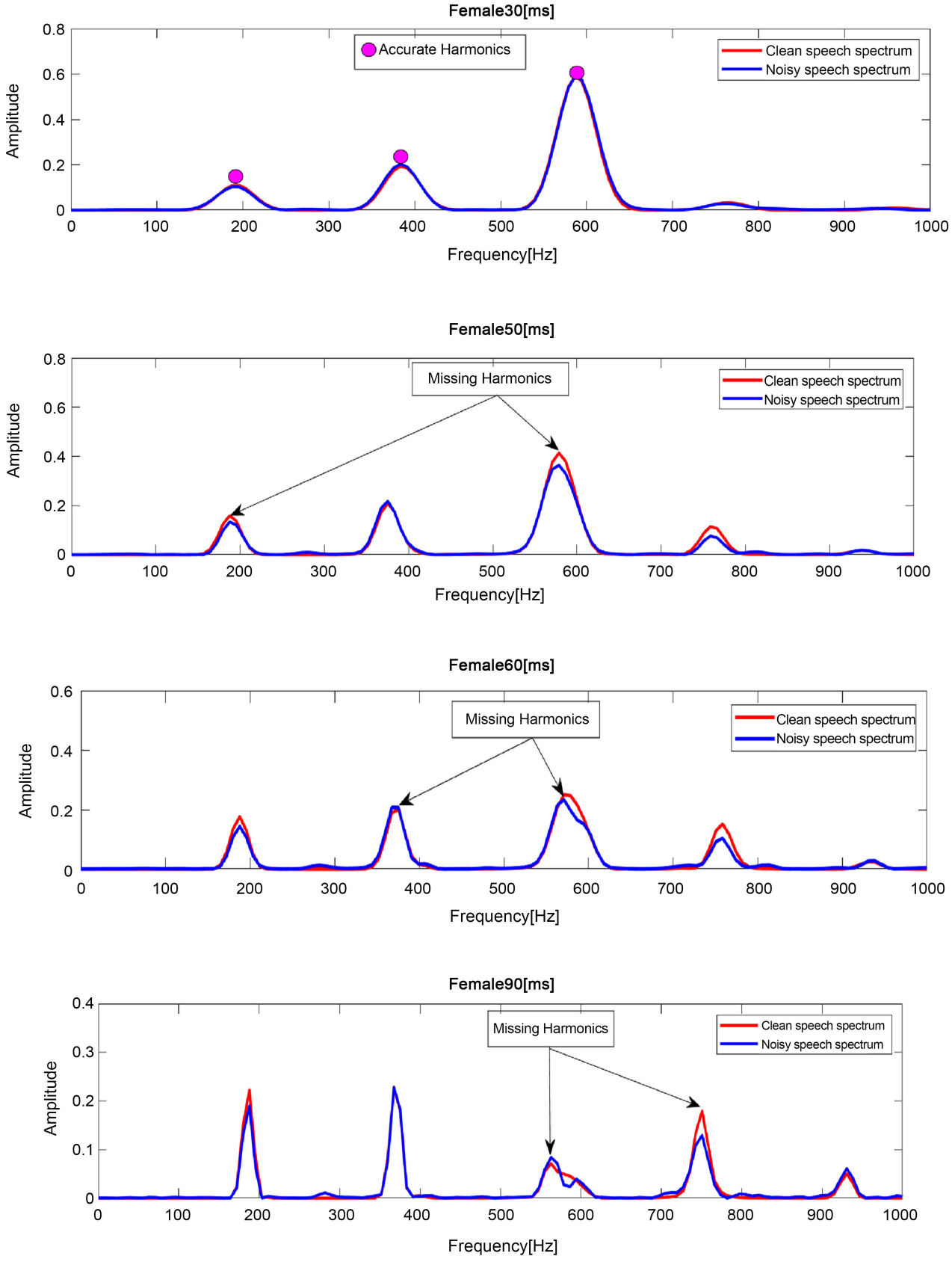

Figure 3. Harmonic characteristics for female speaker.

3. Experimental Results and Discussion

To assess the performance of the proposed method for detecting fundamental frequency in a noisy speech environment for selecting the more accurate frame length in male and female speech signals, separately, where experiments were conducted by utilizing speech signals.

3.1. Experimental Conditions

Speech samples captured from the KEELE database [35] are used to develop the proposed pitch detection method. Speech signals from the KEELE database were uttered by five male and five female speakers. The total length of the ten speakers' speeches is around 6 [m]. They were sampled at a rate of 16 [kHz]. We added several sorts of noise to the speech signals to make noisy speech signals. A computer generated white noise with a zero mean and unit variance, which was then added to the voice signals with amplitude modification. Pink noise, babble noise, and train noise, which were all noise extracted from the NOISEX-92 database [36] at a sampling rate of 20 [kHz]. When these noises were introduced to the speech data in the KEELE database, they were resampled using a sampling frequency of 16 [kHz]. The SNR was set to −10, −5, 0, 5, 10, 20 [dB] and the other experimental conditions for fundamental frequency extraction were:

Frame length: 30 [ms] to 90 [ms] for both BaNa and ACF;

Frame shift: 10 ms;

Window function: Hanning for both BaNa and ACF;

DFT (IDFT) length: 2048 points.

3.2. Evaluation Standards

If the detected F_0 deviates more than 10% from the ground truth value in noisy speech data, it is regarded as a gross pitch error. Otherwise, it is regarded as a fine-pitch error [16]. The fraction of wrongly identified F_0 values in spoken speech segments is referred to as the gross pitch error (GPE) rate. The GPE rate has frequently been used as an error measurement parameter for F_0 detection.Because the F_0 of human speech is typically higher than 50 [Hz] and can reach 600 [Hz] for children or female voices, we set the lower limit and the upper limit for F_0 of human speech to be F_0min = 50 [Hz] and F_0max = 600 [Hz] respectively. We employ widely used noise database with four types of common wide band noise to investigate the parameter sensitivity of the BaNa and ACF methods.

3.3. Comparing Results and Performance

For the validation of our proposed idea, the effectiveness of pitch extraction in noisy situations was assessed between the ACF and BaNa [34]. In [12], nine methods were tested, and BaNa was shown to be the top pitch extractor in noisy situations. Here, we consider four forms of noise, namely White, Pink, Babble, and Train noises. According to [34], the frame duration for BaNa was configured

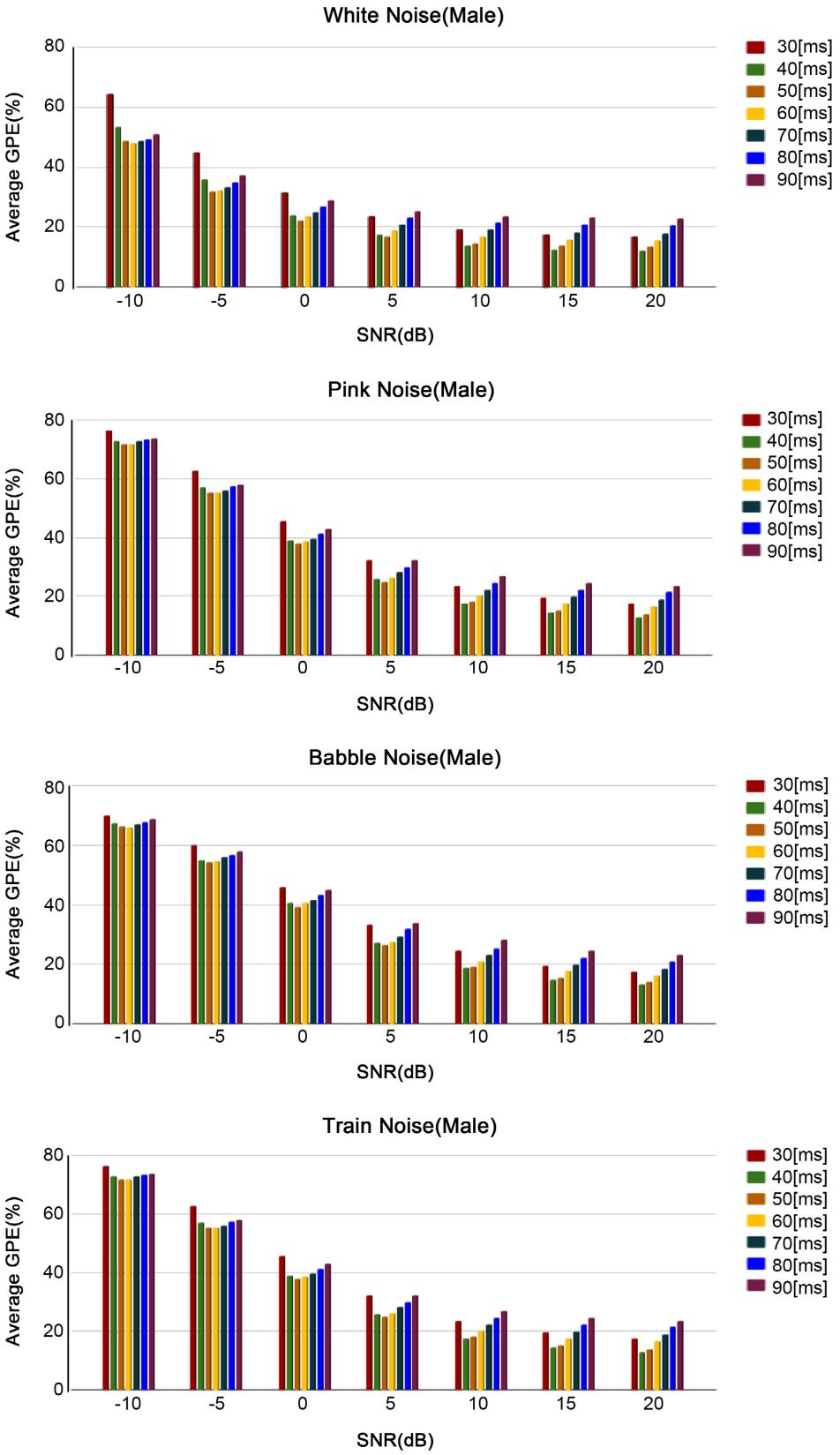

Figure 4. Average GPE comparison of different frame length for male speakers using ACF method in KEELE database.

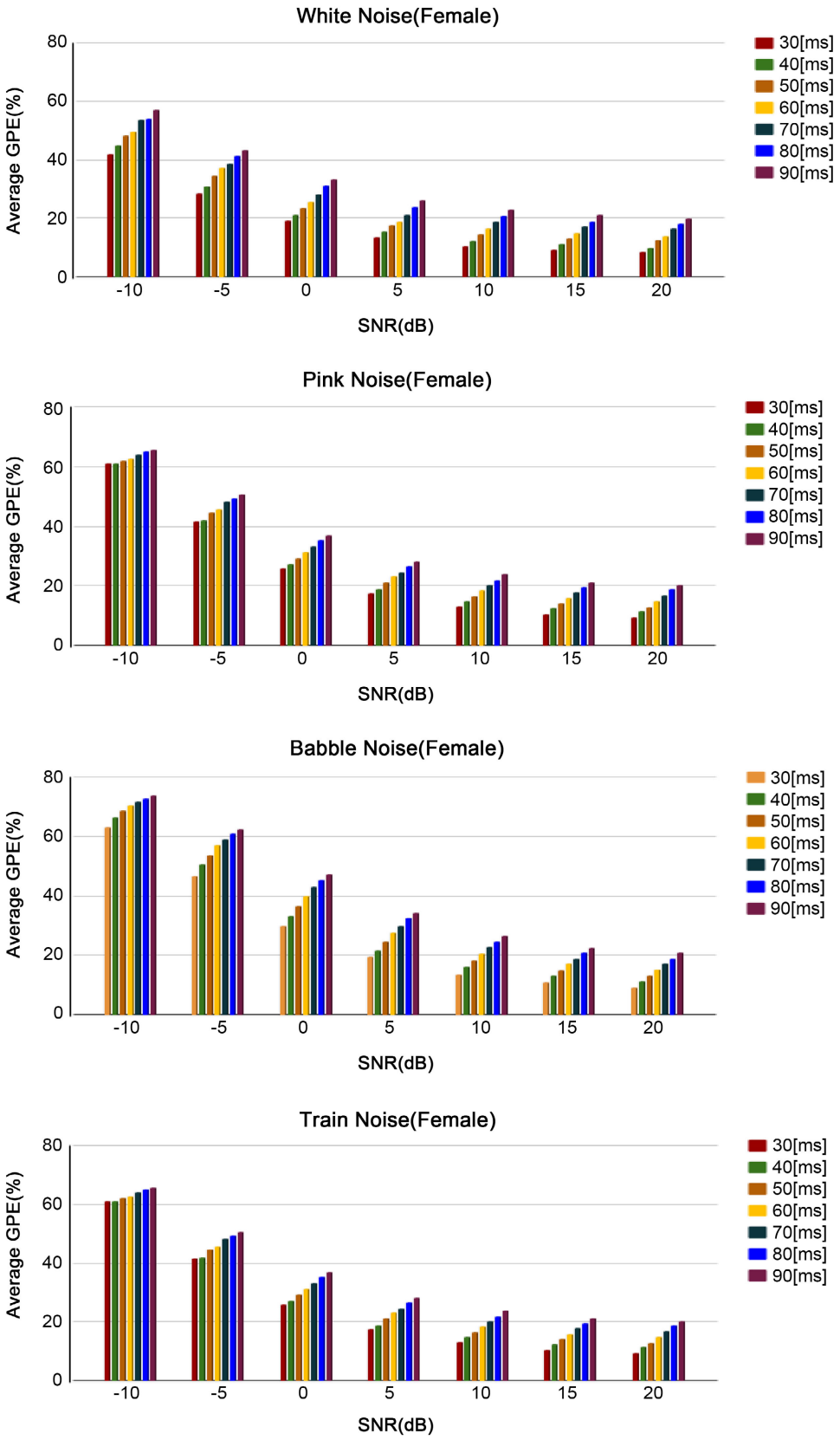

Figure 5. Average GPE comparison of different frame length for female speakers using ACF method in KEELE database.

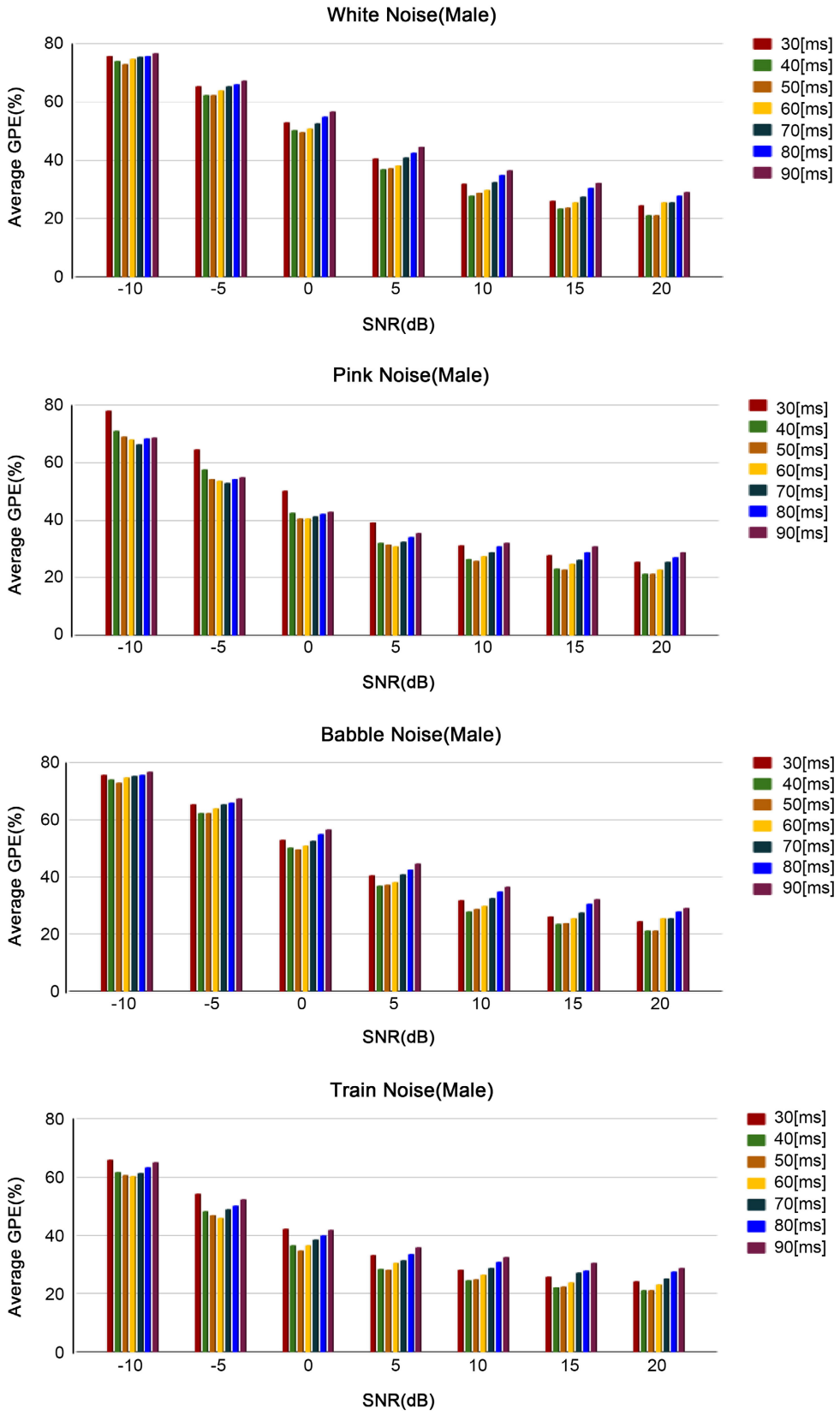

Figure 6. Average GPE comparison of different frame length for male speakers using BaNa method in KEELE database.

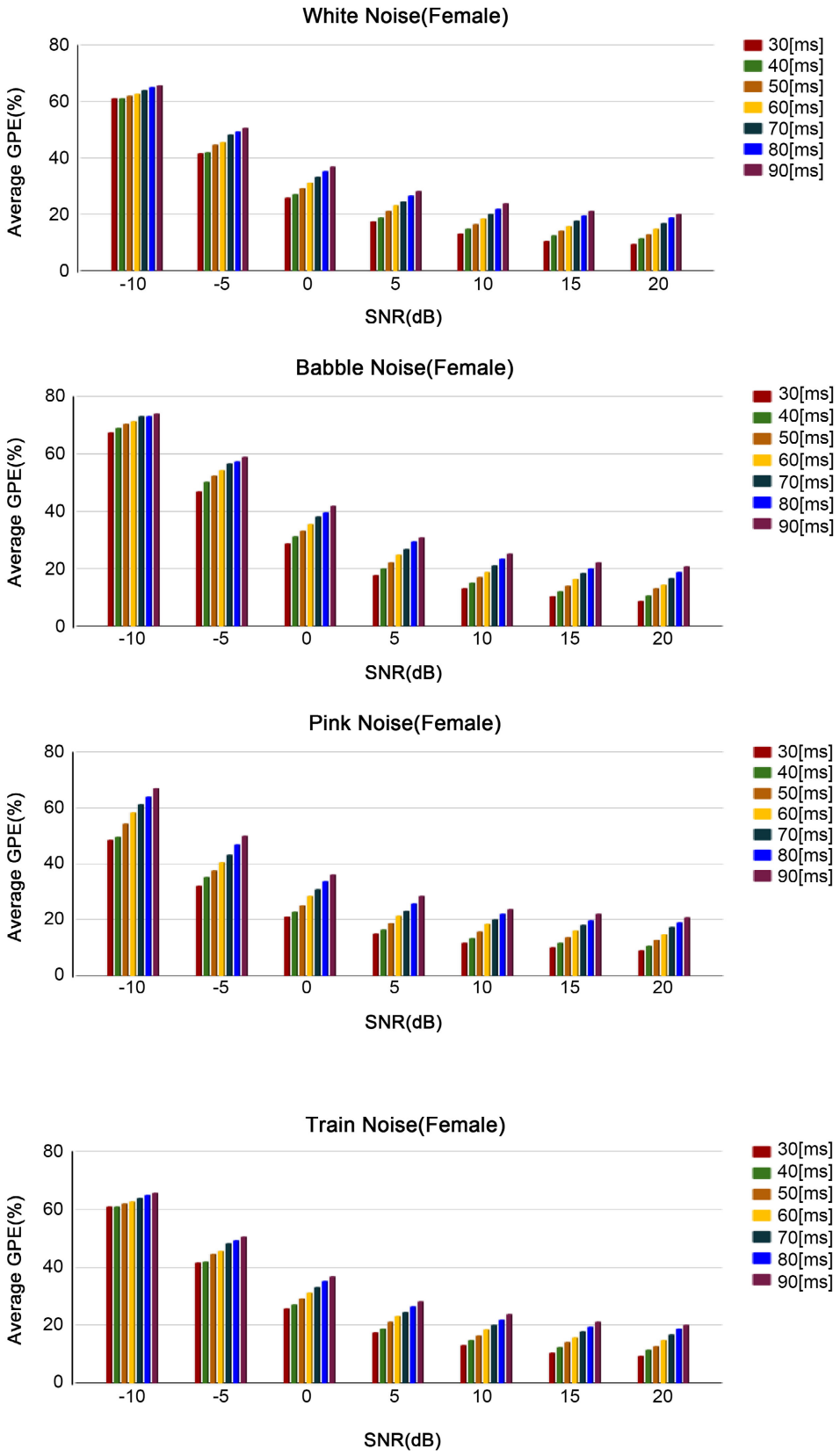

Figure 7. Average GPE comparison of different frame length for female speakers using BaNa method in KEELE database.

to 60 [ms], and 2^{16} points were used for the DFT (IDFT) points. This environment is ideal for BaNa. BaNa was implemented using source code from [37]. For the validation of our proposed idea, we have also utilized the source code of BaNa [37].

Figures 4-7 display the average GPE rate for different frame length at male and female speakers, respectively in KEELE database with various forms of noise. We have utilized the ACF and BaNa for the investigation of our proposed idea.

In the case of male speaker for ACF and BaNa methods in Figure 4 and Figure 6, we have investigated that frame length 50 [ms] provides the lower GPE rate than that of other frame lengths at low SNRs (−10 [dB] to 5 [dB]) almost all noise cases. On the contrary, frame length 50 [ms] is highly competitive with frame length 40 [ms] at high SNRs (10 [dB] to 20 [dB]).

On the other hand, In the case of female speaker for ACF and BaNa methods in Figure 5 and Figure 7, we have investigated that frame length 30 [ms] provides the lower GPE rate than that of other frame lengths at almost all SNRs in every noise cases.

4. Conclusion

The accuracy of fundamental frequency extraction methods varies for different speaker types due to the inherent differences in the characteristics of male and female speech signals. As a result, the impacts of different frame length vary across different speakers. In this paper, different frame length applied to male and female speakers have been analytically and experimentally examined. The observations above indicate that our proposed idea offers a superior way to extract more accurate pitch information from speech that has been affected by noise. We conducted experimental evaluations for validation of our proposed idea at ACF and BaNa methods in KEELE database. The experimental results show that the proposed idea acquires the lowest gross pitch error (GPE) rate while taking frame length 50 [ms] for male speakers and frame length 30 [ms] for female speakers for all kinds of noise and SNR levels investigated.

Acknowledgements

Sincere thanks to the members of JAMP for their professional performance, and special thanks to managing editor *Hellen XU* for a rare attitude of high quality.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Vary, P. and Martin, R. (2006) Digital Speech Transmission: Enhancement, Coding and Error Concealment. John Wiley & Sons, New York.

https://doi.org/10.1002/0470031743

[2] Shahnaz, C. (2002) Pitch Extraction of Noisy Speech Using Dominant Frequency of the Harmonic Speech Model. Department of Electrical and Electronic Engineering.

[3] Ling, Z.H., Wang, Z.G. and Dai, L.R. (2010) Statistical Modeling of Syllable-Level F0 Features for HMM-Based Unit Selection Speech Synthesis. 7*th International Symposium on Chinese Spoken Language Processing*, Tainan, 29 November-3 December 2010, 144-147. https://doi.org/10.1109/ISCSLP.2010.5684833

[4] Sakai, S. and Glass, J. (2003) Fundamental Frequency Modeling for Corpus-Based Speech Synthesis Based on a Statistical Learning Technique. 2003 *IEEE Workshop on Automatic Speech Recognition and Understanding*, St Thomas, 30 November-4 December 2003, 712-717. https://doi.org/10.1109/ASRU.2003.1318527

[5] Buera, L., Droppo, J. and Acero, A. (2008) Speech Enhancement Using a Pitch Predictive Model. 2008 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, 31 March-4 April 2008, 4885-4888. https://doi.org/10.1109/ICASSP.2008.4518752

[6] Ananthakrishnan, S. and Narayanan, S. (2007) Improved Speech Recognition Using Acoustic and Lexical Correlates of Pitch Accent in a N-Best Rescoring Framework. 2007 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 15-20 April 2007, IV-873-IV-876. https://doi.org/10.1109/ICASSP.2007.367209

[7] Sinha, R. and Shahnawazuddin, S. (2018) Assessment of Pitch-Adaptive Front-End Signal Processing for Children's Speech Recognition. *Computer Speech & Language*, **48**, 103-121. https://doi.org/10.1016/j.csl.2017.10.007

[8] Wang, C. (2001) Prosodic Modeling for Improved Speech Recognition and Understanding. Master's Thesis, Massachusetts Institute of Technology, Cambridge.

[9] Furui, S. (1986) Research of Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques. *Speech Communication*, **5**, 183-197. https://doi.org/10.1016/0167-6393(86)90007-5

[10] Kwon, O.W., Chan, K., Hao, J.C. and Lee, T.W. (2003) Emotion Recognition by Speech Signals. 8*th European Conference on Speech Communication and Technology* (*Eurospeech* 2003), Geneva, 1-4 September 2003, 125-128. https://doi.org/10.21437/Eurospeech.2003-80

[11] Park, H., Yoon, J.Y., Kim, J.H. and Oh, E. (2001) Improving Perceptual Quality of Speech in a Noisy Environment by Enhancing Temporal Envelope and Pitch. *IEEE Signal Processing Letters*, **17**, 489-492. https://doi.org/10.1109/LSP.2010.2044937

[12] Sukhostat, L. and Imamverdiyev, Y. (2015) Partial Regularity of Suitable Weak Solutions of the Navier-Stokes Equations. *Journal of Voice*, **29**, 410-417. https://doi.org/10.1016/j.jvoice.2014.09.016

[13] Cardozo, B. and Ritsma, R. (1968) On the Perception of Imperfect Periodicity. *IEEE Transactions on Audio and Electroacoustics*, **16**, 159-164. https://doi.org/10.1109/TAU.1968.1161978

[14] Rabiner, L., Cheng, M., Rosenberg, A. and McGonegal, C. (1976) A Comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**, 399-418. https://doi.org/10.1109/TASSP.1976.1162846

[15] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H. (2008) Tandem-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-Free Spectrum, F0, and Aperiodicity Estimation. 2008 *IEEE International Conference on Acoustics, Speech and*

Signal Processing, Las Vegas, 31 March-4 April 2008, 3933-3936. https://doi.org/10.1109/ICASSP.2008.4518514

[16] Rabiner, L. (1977) On the Use of Autocorrelation Analysis for Pitch Detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **25**, 23-33. https://doi.org/10.1109/TASSP.1977.1162905

[17] Ross, M., Shaffer, H., Cohen, A., Freudberg, R. and Manley, H. (1974) Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **22**, 353-362. https://doi.org/10.1109/TASSP.1974.1162598

[18] Chakraborty, R., Sengupta, D. and Sinha, S. (2009) Pitch Tracking of Acoustic Signals Based on Average Squared Mean Difference Function. *Signal, Image and Video Processing*, **3**, 319-387. https://doi.org/10.1007/s11760-008-0072-5

[19] Shimamura, T. and Kobayashi, H. (2001) Weighted Autocorrelation for Pitch Extraction of Noisy Speech. *IEEE Transactions on Speech and Audio Processing*, **9**, 727-730. https://doi.org/10.1109/89.952490

[20] Boersma, P., *et al.* (1993) Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. *IFA Proceedings*, **17**, 97-110.

[21] De Cheveign, A. and Kawahara, H. (2002) YIN, a Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, **111**, 1917-1930. https://doi.org/10.1121/1.1458024

[22] Noll, A.M. (1967) Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, **41**, 293-309. https://doi.org/10.1121/1.1910339

[23] Kobayashi, H. and Shimamura, T. (1998) A Modified Cepstrum Method for Pitch Extraction. IEEE. APCCAS 1998. 1998 *IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings* (*Cat. No.*98 *EX*242), Chiang Mai, 24-27 November 1998, 293-302.

[24] Rashidul Hasan, M.A.F.M., Rahman, M.S. and Shimamura, T. (2012) Windowless-Autocorrelation-Based Cepstrum Method for Pitch Extraction of Noisy Speech. *Journal of Signal Processing*, **16**, 231-239. https://doi.org/10.2299/jsp.16.231

[25] Gonzalez, S. and Brookes, M. (2014) PEFAC—A Pitch Estimation Algorithm Robust to High Levels of Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**, 518-530. https://doi.org/10.1109/TASLP.2013.2295918

[26] Hermes, D.J. (1988) Measurement of Pitch by Subharmonic Summation. *The Journal of the Acoustical Society of America*, **83**, 257-264. https://doi.org/10.1121/1.396427

[27] Li, B. and Zhang, X.W. (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

[28] Mnasri, Z., Rovetta, S. and Masulli, F. (2022) A Novel Pitch Detection Algorithm Based on Instantaneous Frequency for Clean and Noisy Speech. *Circuits, Systems, and Signal Processing*, **41**, 6226-6294. https://doi.org/10.1007/s00034-022-02082-8

[29] Huang, F. and Lee, T. (2012) Pitch Estimation in Noisy Speech Using Accumulated Peak Spectrum and Sparse Estimation Technique. *IEEE Transactions on Audio, Speech, and Language Processing*, **11**, 99-109. https://doi.org/10.1109/TASL.2012.2215589

[30] Chu, W. and Alwan, A. (2011) SAFE: A Statistical Approach to F0 Estimation under Clean and Noisy Conditions. *IEEE Transactions on Audio, Speech, and Language*

Processing, **20**, 993-944. https://doi.org/10.1109/TASL.2011.2168518

[31] Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M. and Velimirovic, M. (2020) SPICE: Self-Supervised Pitch Estimation. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **28**, 1118-1128. https://doi.org/10.1109/TASLP.2020.2982285

[32] Singh, S., Wang, R. and Qiu, Y. (2021) DeepF0: End-to-End Fundamental Frequency Estimation for Music and Speech Signals. *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP)*, Toronto, 6-11 June 2021, 61-65. https://doi.org/10.1109/ICASSP39728.2021.9414050

[33] Wei, W., Li, P., Yu, Y. and Li, W. (2022) HarmoF0: Logarithmic Scale Dilated Convolution for Pitch Estimation. *IEEE International Conference on Multimedia and Expo*, Taipei City, 18-22 July 2022, 1-6. https://doi.org/10.1109/ICME52920.2022.9858935

[34] Yang, N., Ba, H., Cai, W., Demirkol, I. and Heinzelman, W. (2014) BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, **22**, 1833-1848. https://doi.org/10.1109/TASLP.2014.2352453

[35] Plante, F., Meyer, G. and Ainsworth, W. (1995) A Pitch Extraction Reference Database. 4*th European Conference on Speech Communication and Technology*, Madrid, 18-21 September 1995, 837-840. https://doi.org/10.21437/Eurospeech.1995-191

[36] Varga, A. and Steeneken, H.J.M. (1993) Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Journal of Computer and Communications*, **12**, 247-251. https://doi.org/10.1016/0167-6393(93)90095-3

[37] WCNG, Wireless Communication Networking Group. https://hajim.rochester.edu/ece/sites/wcng/code.html