

Fundamental frequency extraction by utilizing modified BaNa in noisy speech

Arpita Saha¹, Nargis Parvin², Md. Saifur Rahman¹, Moinur Rahman³, Any Chowdhury¹

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

²Department of Computer Science and Engineering, Faculty of Electrical and Computer Engineering, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh

³Department of Information Technology, Faculty of Science and Engineering, University of Information Technology and Sciences (UITS), Dhaka, Bangladesh

Article Info

Article history:

Received Dec 17, 2023

Revised Mar 31, 2024

Accepted Apr 24, 2024

Keywords:

BaNa

Fundamental frequency

Gross pitch error rate

Harmonics

Pitch

ABSTRACT

A sound's pitch can be largely understood and perceived by using its fundamental frequency. Multiple algorithms have been developed for extracting fundamental frequency, and the choice of which one to employ depends on the noise and features of the signal. Therefore, for an accurate fundamental frequency estimate, the noise resistance of the algorithm becomes even more crucial. Still, many of the most advanced algorithms fail to produce acceptable results when faced with loud speech recordings that have low signal-to-noise ratios (SNRs). In this research paper, we focus on the harmonic selection step in BaNa method, which is one of the vital parts for enhancing the extraction accuracy of fundamental frequency (F_0) in noisy situations. BaNa algorithm always emphasizes 5 harmonics on average for both male and female speakers. However, our observation reveals that relying on 5 harmonics is inadequate for male speakers in noisy conditions. Thus, we propose a new idea based on BaNa that separately utilizes the 3 harmonics for male speakers and 5 harmonics for female speakers to achieve accurate pitch extraction within noisy environments. The results demonstrate that our proposed approach attains the lowest rate of gross pitch error (GPE) across various noise types and SNR levels.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Md. Saifur Rahman

Department of Information and Communication Technology, Faculty of Engineering, Comilla University
Cumilla, 3506, Bangladesh

Email: saifurice@cou.ac.bd

1. INTRODUCTION

Speech is the primary means of communication, which is the verbal representation of ideas. During speech, sounds are produced by the vocal tract, which consists of the mouth, lips, pharynx, vocal cords, and lungs. Depending on the vibration of vocal chords, generated signal can be expressed as silence, voiced, or unvoiced [1]. Speech's fundamental frequency is one of its most important prosodic features. A speech signal's fundamental frequency, commonly symbolized as (F_0), represents the estimated frequency of the quasi-periodic pattern found in voiced speech sounds. Different factors, including age, gender, and linguistic context, might impact an individual's unique frequency range. In addition, intonation, stress, emotion, and illness can all have an impact. Depending on the structure of their vocal cords, each individual possesses a unique set of F_0 . Women generally display a F_0 range of 120 Hz to 500 Hz, while the male frequency range is usually 50 Hz to 250 Hz [2].

Pitch in speech is perceived at a F_0 that is discerned by the rate of vibration of the vocal cords during voiced speech sounds. Therefore, F_0 found in speech signal is commonly referred to as pitch. Speech processing applications such as speech synthesis [3], [4], speech enhancement [5], speech recognition [6]-[9], and emotion identification [10] need accurate pitch detection of a voice signal. By applying pitch enhancement techniques in the frequency domain, [11] improves voice clarity over noise. To improve voice quality, [5] makes use of the pitch period to build a durable paradigm for speech and background noise. In one study, a baseline automated speech recognition (ASR) system's performance is improved by including prosodic events, particularly pitch accents [6]. A different piece of research reduces the sensitivity of a voice recognition system to pitch fluctuations to create a child-friendly system [7]. To increase the viability of the aforementioned applications, accurate pitch data extraction from the voice is crucial. Nevertheless, there are still several obstacles in extracting pitch from speech. It becomes challenging to identify the exact F_0 when speech signals are tainted by noise [12]. Furthermore, since a clean speech waveform [13] goes through substantial structural alterations during vocal tract transit, extraction of pitch has been a formidable task, particularly in a noise-free environment.

Pitch determination methods up to now have relied on the unique characteristics of speech signals, such as the periodic pattern in the temporal domain [14] or the harmonic structure in the spectrum domain [15]. Within the time domain to extract pitch information from speech signals, a wide range of algorithms are applied. These includes techniques 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 variety of pitch-detecting techniques, the ACF [16] determines the time interval that produces the closest separation by evaluating the similarity between two segments of a voice signal. A simplified version of ACF, the AMDF [17], calculates a signal by averaging the magnitudes of the differences between it and its delayed version. The determinations of WAF [19] involve multiplying the autocorrelation function of the signal by a predetermined range of weights. This method uses the reciprocal of AMDF to assign importance to the autocorrelation function. During signal filtering, the WAF [19] may be used to remove noise and other undesired signals from a speech. Using the Viterbi method to find the least costly path across all of the segments, Praat [20] determines the optimal F_0 candidate for every brief sound segment by analyzing the maxima of the segment's autocorrelation. By applying a cumulative averaging method to the difference function, YIN [21] concentrates on the interaction between the standard ACF and the differentiated function. Pitch extraction errors are to be reduced by using this technology.

Techniques for deriving pitches based on ACF are strong against white noise and insensitive to waveform phase irregularities. On the contrary, when noise-induced effects are introduced into clean speech, the effectiveness of pitch extraction guided by ACF tends to decrease and performance is worse. Additionally, changes in the characteristics of the vocal tract might affect the autocorrelation function's behavior. Within the frequency domain, to mitigate the influence of vocal tract features, numerous pitch extraction methods are developed. Finding F_0 in this case involves finding harmonic peaks in the power spectrum. One commonly utilized method is the cepstrum method (CEP) [22]. By reversing the Fourier transform of the Fourier spectrum's logarithmic magnitude, the cepstrum is obtained. This captures the time interval in voice harmonics and results in a noticeable peak that coincides with the frequency interval. The CEP's logarithmic function helps to segregate periodic features from vocal tract properties in the speech signal. When faced with the complexity of noisy situations, the CEP performs accurately in quiet environments but its efficacy is significantly reduced. Kobayashi and Shimamura [23], presented the modified CEP (MCEP) incorporates liftering and clipping onto the logarithmic spectrum as additional steps. This procedure has dual purposes: it removes characteristics of the voice tract and unwanted spectral notches associated with noise in the logarithmic spectrum. In addition, high-frequency components are removed by the MCEP, which enhances pitch extraction precision. The windowless autocorrelation function based CEP (WLACF-CEP) [24] reduces the influence of noise on speech signal contaminated with noise, allowing its integration with the CEP approach, and resulting in enhanced accuracy in extracting pitch information. The WLACF-CEP demonstrates a remarkable resilience against a wide range of noise kinds. Pitch detection in the frequency domain is accomplished utilizing the pitch estimation filter with amplitude compression (PEFAC) method [25]. Sub-harmonic summations [26] are used in the logarithmic frequency spectrum. Moreover, PEFAC incorporates a special amplitude compression to bolster its resistance to noise disruption.

Not all harmonics in the frequency domain follow exact integer multiples of the F_0 . In addition, the higher-order harmonics exhibit a greater degree of drift than their lower-order harmonic counterparts. In conse-

quence, to account for these variations, a tolerance range must be set when calculating the harmonic frequency ratios. Using both logarithmic and power functions, [27] reduces the effect of formants and utilizes the Radon transform to provide a novel method for estimating pitch in noisy speech conditions. It also incorporates the Viterbi algorithm for pitch pattern refinement. [28] based on establishing a pragmatic relationship between the instantaneous frequency (F_i) and the F_0 . It determines whether speech areas are voiced or unvoiced and extracts the F_0 contour by approximating it as a smoothed envelope of remaining F_i values. To estimate pitch by comparing the temporal accumulations of clean and noisy speech samples, the temporally accumulated peak spectrum (TAPS) algorithm, as described in [29], trains a set of peak spectrum exemplars. To understand how noise affects the locations and amplitudes within the spectrum of clear speech, Chu and Alwan developed the statistical algorithm for F_0 estimation (SAFE) [30] model. Pitch estimation is enhanced using self-supervised pitch estimation (SPICE), as stated in [31], by refining the acquired data and training a constant Q transform of signals. To accommodate pitches with varying noise levels, Deep F_0 [32] expands the network's receptive range. It has been demonstrated that Harmo F_0 [33] outperforms Deep F_0 in pitch estimation by employing a range of dilated convolutions.

Almost all researchers are considered into the female and male speakers for pitch extraction. From the above observations, in the case of female speakers, we have investigated that the fewest harmonics are present in the first formant range (up to 1 kHz). Therefore, female speech signal is less affected when it is contaminated by noise. On the contrary, In the case of male speakers, we have investigated that a significant amount of harmonics are present in the first formant range (up to 1 kHz). Therefore, male speech signal is highly affected when it is contaminated by noise. So, the male speech signal could not maintain the periodicity for extracting the pitch and showed a higher error rate than that of the female speech signal. Depending on the above properties, BaNa [34] opts for the initial five amplitude spectral peaks from the speech signal's spectrum on average for both male and female speakers. Nonetheless, our observation reveals that relying on five harmonics is inadequate for male speakers under noisy conditions. Our objective is to develop a method for extracting pitches capable of maintaining its accuracy despite the presence of noise, effectively countering both white noise and various colored noises. In this research, we investigated 3, 5, and 7 harmonics for the male and female speakers separately, and we found more expedient harmonics for each speaker individually which is represented as modified BaNa (proposed). This study looked into the effects of harmonic characteristics While previous studies investigated the impact of vocal tract and noise features separately, they did not explicitly address its influence on the harmonic characteristics of male and female speech signals. In this study, we emphasize the use of individual harmonics in male and female voice signals to enhance the precision of pitch extraction, particularly in noisy environments, especially at low signal-to-noise ratios (SNRs).

2. METHOD

In the case of extracting fundamental frequency from speech signals, researchers are employing various harmonic combinations in their pursuit of acquiring precise pitch. However, it's worth noting that they did not specify exact harmonic suitable for precisely extracting the pitch peak in both male and female speech signals. Noise addition can alter the speech peaks' form, and this change depends on the noise type and its intensity. The real-world noise makes pitch recognition quite challenging due to its ability to obfuscate the periodic pattern of the speech waveform. The performance of pitch detection significantly decreases at low SNRs for all previously mentioned approaches.

In this article, we have investigated the existing state-of-the-art approach BaNa [34] for the extraction of pitch to identify the strengths and weakness, where BaNa is one of the most power full pitch extraction method up to now [12], [27]. Pitch detection with BaNa is a hybrid technique that uses the cepstrum method to extract pitch from noisy signals and harmonic frequency ratios. BaNa consists of the following steps: i) preprocessing, ii) search for harmonic peaks, iii) calculate pitch candidates, and iv) selection of the pitch from the candidates F_0 . For the proposed idea, we have also utilized the four steps which is similar to BaNa [34]. In step 2: BaNa considers the 5 harmonics for male and female speech signals, simultaneously. On the other hand, our proposed idea is to find out the most appropriate harmonics for male and female speech signals, separately which is represented as in Figures 1-6.

From the aforementioned finding, presented in Figures 1(a) and (b), we have discovered that when 5 harmonics are considered, the peaks are shifted due to the addition of noise, leading to inaccurate pitch peak detection in the case of male speakers, as depicted in Figure 1(a). Conversely, for female speakers, as illustrated

in Figure 1(b), most of the harmonics are more accurate according to the pitch peak, even with the presence of noise. In this research work, in contrast to BaNa methods, instead of calculating all the pitch harmonics, we employ three distinct harmonics for both speakers separately required to increase the detected pitch precision. Based on the preceding observation, we explored 3, 5, and 7 harmonics for extracting pitch in both speakers individually within noisy environments, and we found more appropriate harmonics for the male and female speakers separately. We propose, 3 harmonics for male speakers and 5 harmonics for female speakers are more appropriate when extracting speech from noisy speech.

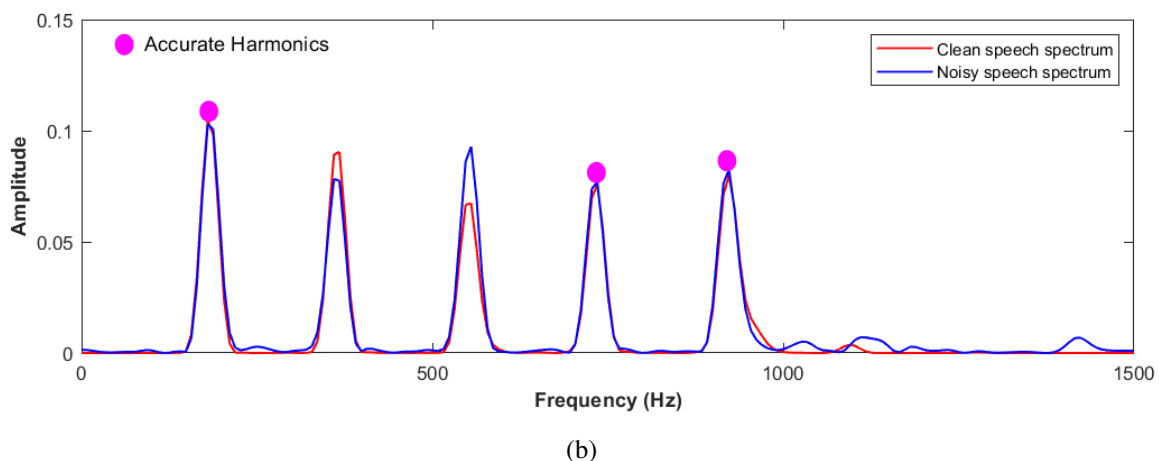
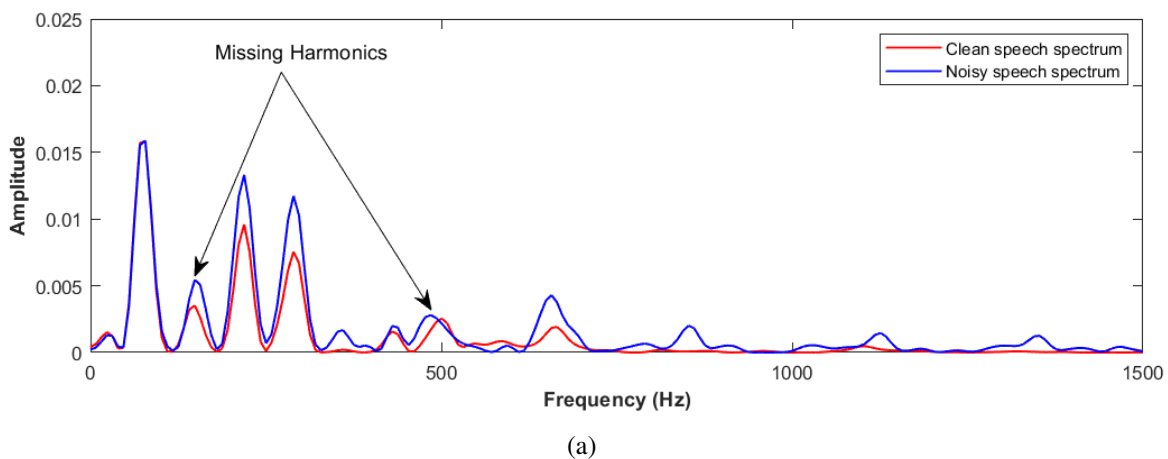


Figure 1. Speech signals' harmonic characteristics in clean, noisy environments for (a) male speakers and (b) female speakers, in case of 5 harmonics

Based on the insights gained from the above observations and investigations, we have illustrated Figures 2(a) and (b), which depict the harmonic characteristics of male speakers in case of 3 harmonics and female speakers in case of 5 harmonics, respectively, in noisy environments. In the instance of 3 harmonics, considering male speakers, from Figure 2(a), we observe that additional noise has less impact on clean speech. We see that the clean speech spectrum and the noisy speech spectrum are nearly similar, and the number of harmonics is more precise than that of the 5 harmonics as well as the 7 harmonics. As a consequence, we acquire more appropriate pitch information. In the case of 5 and 7 harmonics, almost all harmonics are missing to detect the pitch peak. On the other hand, from Figure 2(b), we have observed that in the case of female speakers, while considering 5 harmonics, clean speech is not much impacted by additional noise since female speakers have peaks with large amplitudes. As a consequence, the pitch information we receive is more relevant. The state-of-the-art method BaNa also considers the 5 harmonics, and we have also investigated that the 5 harmonics are more accurate than the 3 harmonics and 7 harmonics in the case of female speakers.

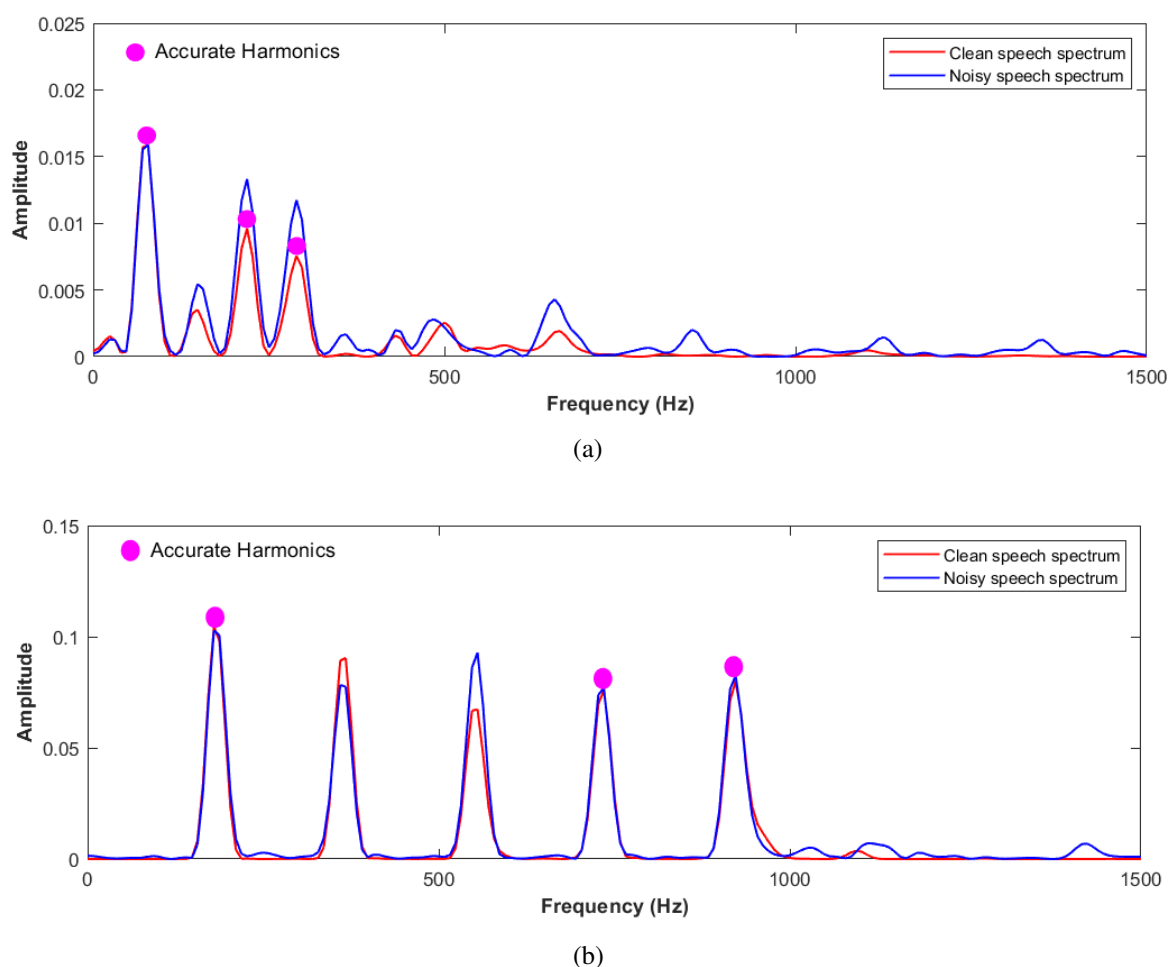


Figure 2. Speech signals' harmonic characteristics in clean, noisy environments for (a) male speakers in case of 3 harmonics and (b) female speakers in case of 5 harmonics

3. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments using speech signals were carried out to evaluate the effectiveness of the suggested approach for extracting the accurate pitch in noisy speech. The proposed method is highlighted to extract the more accurate harmonics in male and female speech signals, separately. These experiments were conducted by utilizing the two databases of speech signals.

3.1. Experimental conditions

We employed speech signals derived from the KEELE database [35] and the NTT database [36]. The KEELE database provided speech recordings delivered by a group of five male and five female speakers. The combined duration of the speech signals from these ten speakers within the KEELE database amounts to approximately 6 m. These speech recordings underwent sampling at a frequency of 16 kHz. In the NTT database, there are a total of eight Japanese speakers, split equally between males and females. The speech materials attributed to them possess a duration of 11 s. These vocal signals underwent sampling at a frequency of 10 kHz. It is possible to create noisy speech samples by combining clean speech samples with noise recorded in noisy places. We combined several kinds of noise with the voice signals to produce noisy speech signals. A total of seven distinct categories of noise were infused into the original signals, each with varying SNR levels, to assess how resilient the algorithms are to noise. These noise types encompass white noise, pink noise, babble noise, HF channel noise, car-interior noise, military vehicle noise, and train noise, sourced from the NOISEX-92 database [37] and recorded at a sampling frequency of 20 kHz. For evaluation purposes, these noises were

resampled to 16 kHz for assessing signals from the KEELE database and 10 kHz for assessing signals from the NTT database. The additional experimental setups for extracting fundamental frequency are as in Table 1.

Table 1. Optimal values of tune parameters for which BaNa algorithm is tested

Parameters	Optimal value
Frame length	60 ms
Number of chosen spectral peaks	3, 5, and 7
SNR level	-10 to 20 dB
Frame shift	10 ms
Windowing function	Hanning window
Fast fourier transform (FFT) size (points)	2^{16}
Lower limit and upper limit of human speech F_0	$F_{0\min} = 50$ Hz, $F_{0\max} = 600$ Hz
Error measurement metric	GPE rate

The evaluation of fundamental frequency extraction precision was carried out using the ensuing fundamental frequency extraction error, denoted as $e(l)$, based on Rabiner's rule [14]:

$$e(l) = F_{est}(l) - F_{true}(l) \quad (1)$$

Herein, l signifies the frame number;

$F_{est}(l)$ represents the fundamental frequency extracted from the $l - th$ frame and

$F_{true}(l)$ corresponds to the true fundamental frequency of the $l - th$ frame.

If $e(l) > 10\%$ of the $F_{true}(l)$, then the error is classified as gross pitch error (GPE). If not, it's referred to as fine pitch error (FPE).

3.2. Preliminary experiments

It is impossible to overlook the fact that a speaker's features have a significant impact on the extraction performance when it comes to pitch extraction, particularly when it comes to low or high pitches [1], [2], which correspond to female and male speech characteristics, respectively. Additionally, distinct additive noise characteristics, flat-spectral pattern or not, and whether or not they are time-invariant produce distinct outcomes for pitch extraction. This is because a complicated combination of formant properties, speech harmonics, and noise structure produced in framed voiced speech all cause nonuniform behaviors. As a result, it's critical to examine the accuracy of pitch extraction performance independently for both male and female speech as well as independently for each type of noise based on the number of harmonics.

For the proposed method, it is important to set the appropriate number of harmonics in BaNa of the female and male speaker per amount of noise and clean speech. Therefore, we conduct the preliminary experiments to investigate the best harmonics level in KEELE and NTT databases, respectively. Here, we select the value of harmonics 3, 5, and 7 which is more accurate with the amount of noise. Figures 3-6 represent the relationship between the average GPE of BaNa for male speakers and female speakers with respect to the harmonics level, respectively with white noise and color noises (pink, babble, train, high frequency (HF) channel, car interior, and military vehicle noises) in KEELE and NTT databases, respectively.

In the case of KEELE database, From the evaluation in Figure 3, we have observed that the harmonics level 3 shows a lower GPE rate than the other's harmonic level at all SNRs level of all noises (from Figures 3(a)-(g)) in the male speaker. On the other hand as in Figure 4, the harmonics level 5 shows a lower GPE rate than the other's harmonic level at all SNRs level of all noises (from Figures 4(a)-(g)) in the female speaker. In the case of the NTT database, From the experimental Figures 5 and 6, we have observed that harmonic levels 3 and 5 provide the lower GPE rate at all SNRs of all noise cases (from Figures 5(a)-(g) and from Figures 6(a)-(g)) in male and female speakers, respectively, which shows similar behavior in KEELE database. According to the experimental result, for estimating the average GPE, we used the harmonic level as 3, and 5 at the male and female speakers, respectively in the proposed idea. Both threshold values are highly effective at low SNR values as well as high SNRs of speech signals in both databases.

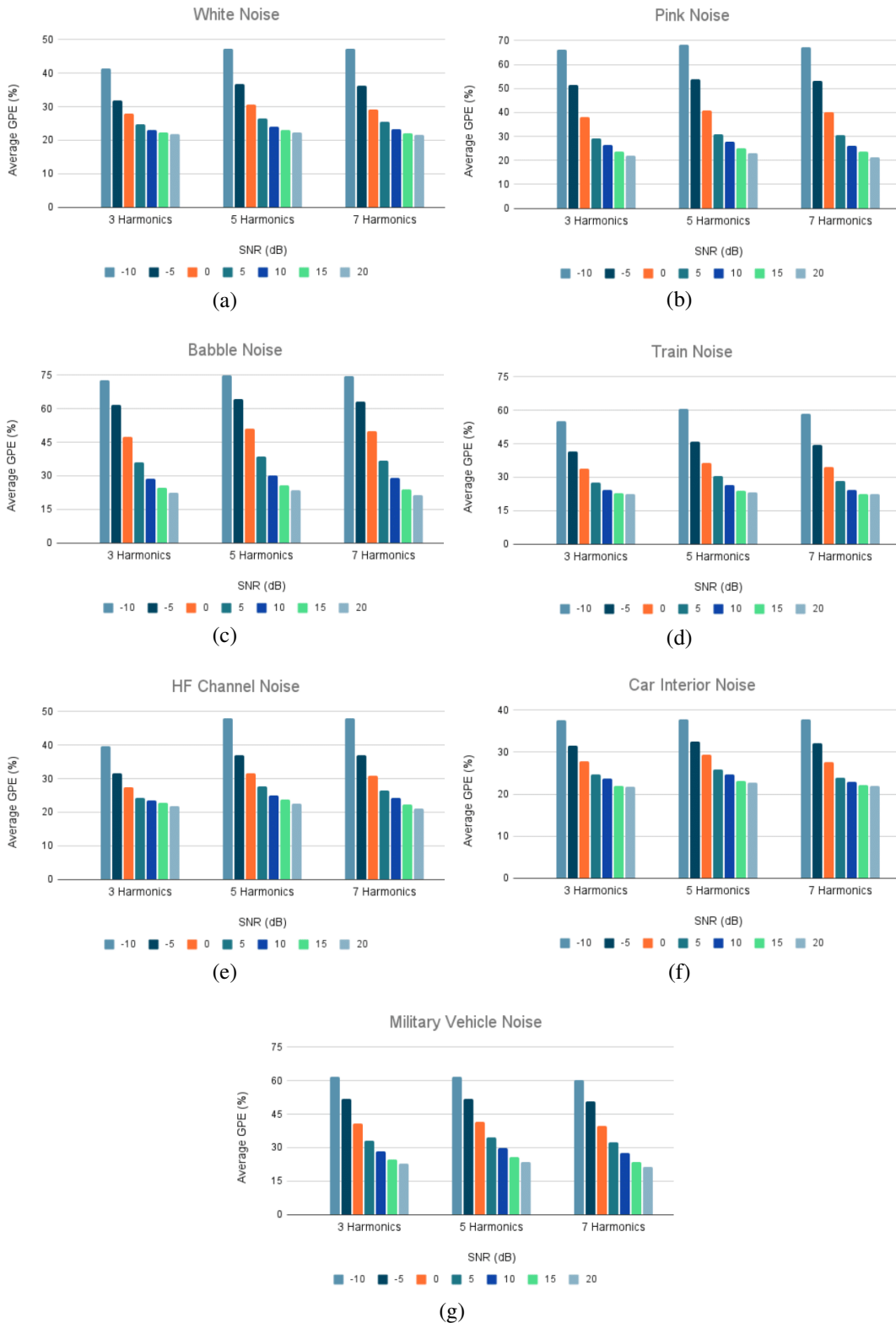


Figure 3. For male speakers in KEELE database, the average GPE rate at various harmonic levels on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

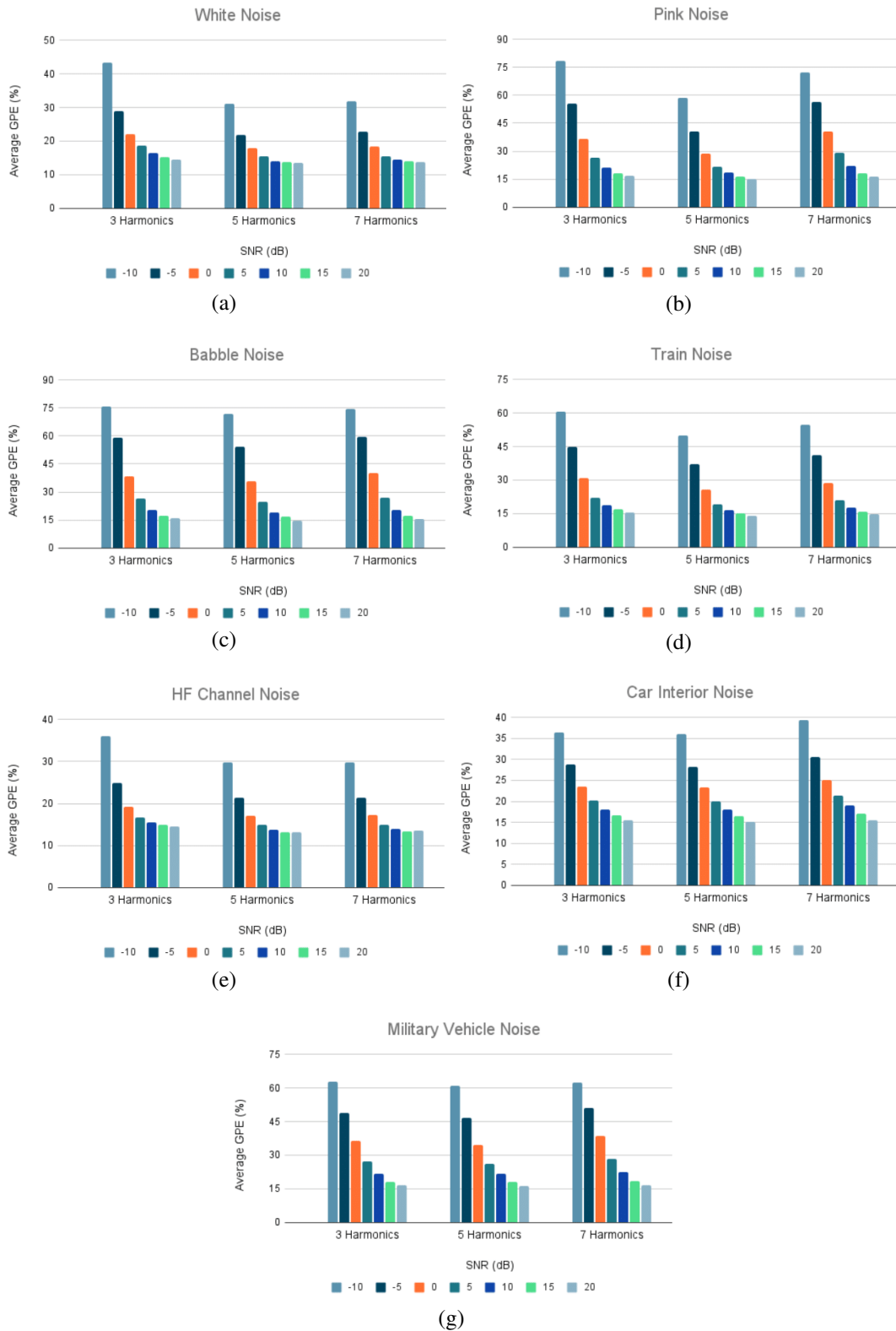


Figure 4. For female speakers in KEELE database, the average GPE rate at various harmonic levels on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

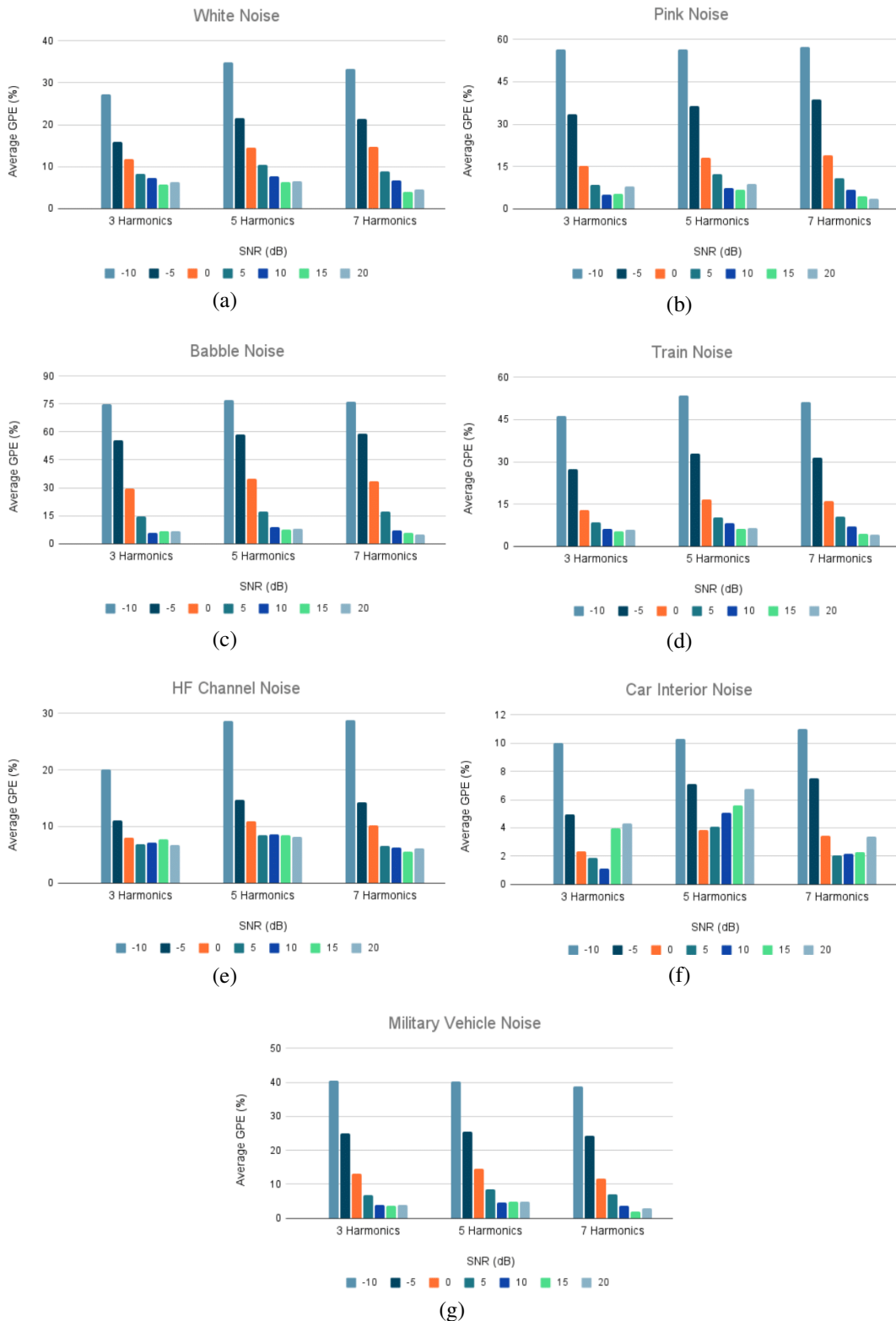


Figure 5. For male speakers in NTT database, the average GPE rate at various harmonic levels on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

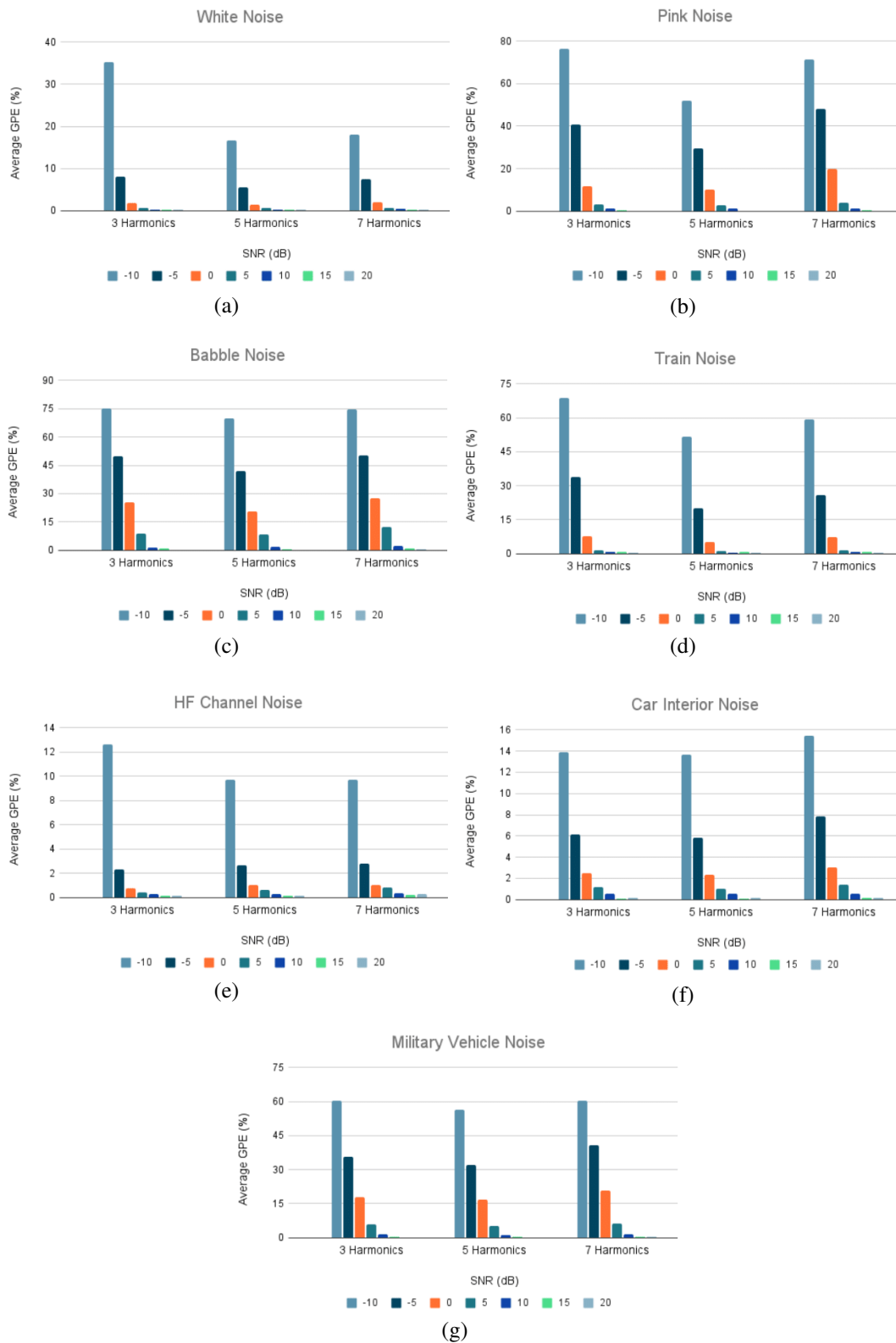


Figure 6. For female speakers in NTT database, the average GPE rate at various harmonic levels on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

3.3. Performance comparisons

Pitch extraction efficacy in environments with high levels of noise was compared between the suggested approach and the traditional approaches (YIN [21], PEFAC [25], and BaNa [34]). BaNa proved to be the most successful pitch extractor in noisy environments after nine different methods were tried in [12]. We examine seven forms of noise, namely white, pink, babble, train, HF channel, car interior, and military vehicle noises. Except for the length of the frame and quantity of inverse density functional theory (IDFT) points for PEFAC and BaNa, all the factors of the existing techniques were identical to those of the proposed technique. Yang *et al.* [34] states that 2^{16} points were utilized for the IDFT points, and that the frame duration for BaNa was set to 60 ms. This environment is ideal for BaNa. BaNa was implemented utilizing the source code provided in [38]. According to the advice in [25], hamming window function and 90 ms were utilized as the window function and frame length respectively. The source code uses 2^{13} as the value for the IDFT points. The PEFAC implementation's source code was compiled from [39]. To authenticate our proposed idea, we have also utilized the source code of BaNa [38]. Figures 7 and 8 display the average GPE rate of the three algorithms and proposed idea for identifying the pitch on the NTT and KEELE databases, respectively with various forms of noise.

According to Figure 7, our findings indicate that it provides a lower GPE rate at low SNRs (-10 to 5 dB) than that of other conventional methods (YIN, PEFAC, and BaNa) at almost all noise cases (from Figures 7(a)-(f)) except military vehicle noise (Figure 7(g)). When it comes to military vehicle noise, the PEFAC method provides a lower GPE rate at low SNRs and the proposed idea shows a lower GPE rate than that of BaNa. On the other hand, the proposed method may benefit from the high SNRs (10 to 20 dB) in all noises (from Figures 7(a)-(g)) without negatively affecting other conventional methods.

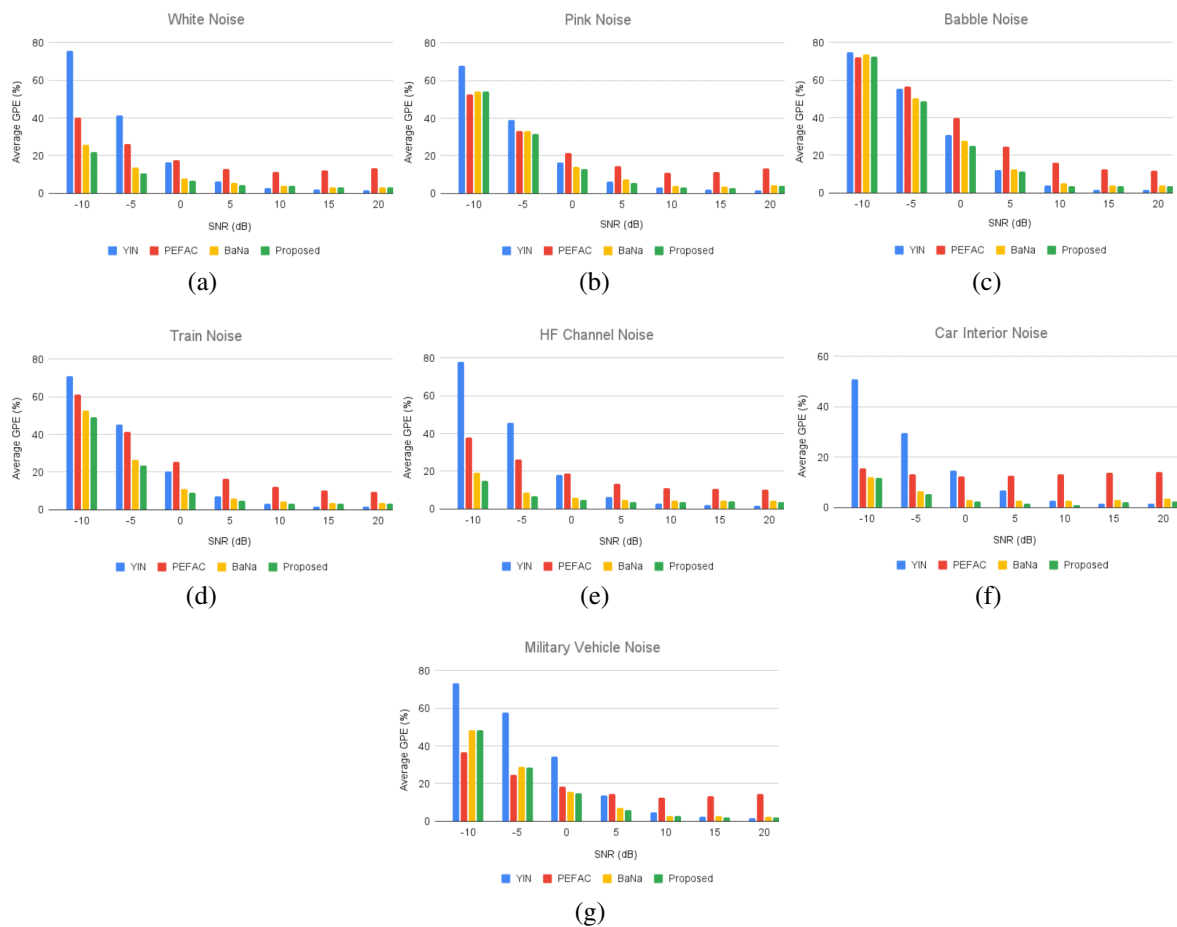


Figure 7. Average GPE rate in NTT database on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

To validate the performance of the Proposed idea more reliably, we also consider the KEELE database. Figure 8 shows the average GPE rates for an average of male and female speakers, respectively. Pitch values originally obtained from laryngograph signals re available in the KEELE database. Upon closer inspection, we discovered that there are some gaps. Consequently, the original pitch values aren't particularly precise. This is evident in the resulting GPE percentages. The GPE percentages of high SNR (20 dB) in Figure 8 is substantially greater than high SNR (20 dB) in Figure 7. This is due to the KEELE database's original pitch values being less accurate. Figure 8 (from Figures 8(a)-(g)) demonstrates a pattern akin to that in Figure 7 for all methods. From a performance comparison aspect, the YIN and PEFAC methods have comparatively low performance at low SNRs of -10 dB and -5 dB in the babble noise as in Figure 8(c) and military vehicle noise as in Figure 8(g) cases, respectively.

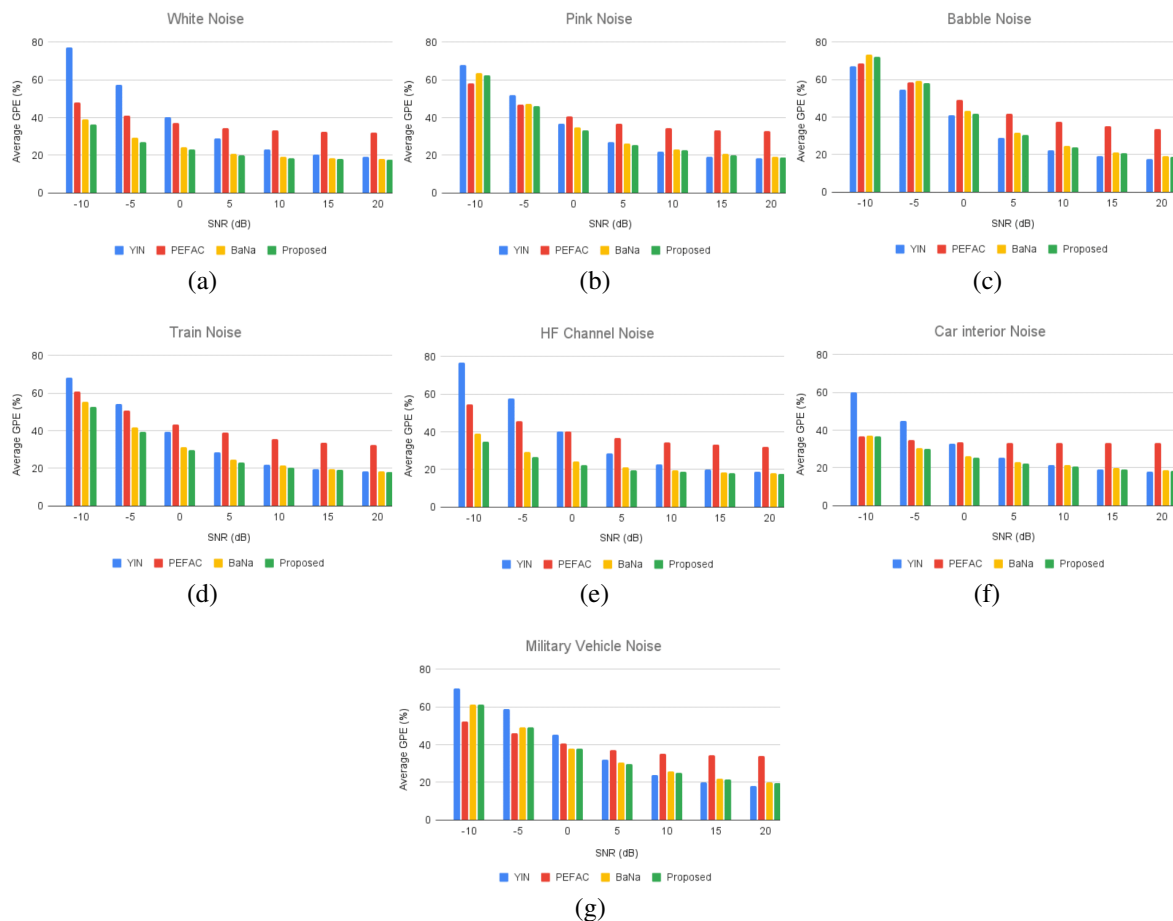


Figure 8. Average GPE rate in KEELE database on (a) white noise, (b) pink noise, (c) babble noise, (d) train noise, (e) HF channel noise, (f) car interior noise, and (g) military vehicle noise

4. CONCLUSION

The intrinsic distinctions between the features of male and female voice signals cause fundamental frequency extraction techniques to differ in accuracy for various speaker types. Different speakers will therefore experience the effects of different harmonics differently. This research presents an experimental and analytical analysis of various harmonics applied to both speakers. According to the aforementioned data, our suggested method outperforms existing algorithms, when speech is impacted by noise. We carried out experimental assessments to compare the performance of the proposed idea against BaNa, PEFAC, and YIN on two speech databases, including KEELE and NTT. Our results provide conclusive evidence that it achieves the lowest GPE rate for all noise and SNR levels examined while accounting for 3 harmonics for male speakers and 5 harmonics

for female speakers. Therefore, our research shows that it is more resilient than other conventional methods without any complicated post-processing according to their noise type and SNRs. Future research may look into developing a new method for extracting pitch that will be exceptionally resilient to extremely low SNR cases across various real-world noise case which will be more effective in speech processing applications.

REFERENCES

- [1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*. Wiley, 2006. doi: 10.1002/0470031743.
- [2] C. Shahnaz, "Pitch extraction of noisy speech using dominant frequency of the harmonic speech model," Department of Electrical and Electronic Engineering, 2002.
- [3] Z.-H. Ling, Z.-G. Wang, and L.-R. Dai, "Statistical modeling of syllable-level F0 features for HMM-based unit selection speech synthesis," in *2010 7th International Symposium on Chinese Spoken Language Processing*, IEEE, Nov. 2010, pp. 144–147. doi: 10.1109/ISCSLP.2010.5684833.
- [4] S. Sakai and J. Glass, "Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, IEEE, 2003, pp. 712–717. doi: 10.1109/ASRU.2003.1318527.
- [5] L. Buera, J. Droppo, and A. Acero, "Speech enhancement using a pitch predictive model," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Mar. 2008, pp. 4885–4888. doi: 10.1109/ICASSP.2008.4518752.
- [6] S. Ananthkrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IEEE, 2007, pp. IV-873–IV-876. doi: 10.1109/ICASSP.2007.367209.
- [7] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for children's speech recognition," *Computer Speech & Language*, vol. 48, pp. 103–121, Mar. 2018, doi: 10.1016/j.csl.2017.10.007.
- [8] C. Wang, "Prosodic modeling for improved speech recognition and understanding," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.
- [9] S. Furui, "Research of individuality features in speech waves and automatic speaker recognition techniques," *Speech Communication*, vol. 5, no. 2, pp. 183–197, Jun. 1986, doi: 10.1016/0167-6393(86)90007-5.
- [10] K. D. Jang and O. W. Kwon, "Speech Emotion Recognition by Speech Signals on a Simulated Intelligent Robot," in *Proceedings of the KSPS conference*, The Korean Society Of Phonetic Sciences and Speech Technology, 2005, pp. 163–166.
- [11] H. Park, J. Y. Yoon, J. H. Kim, and E. Oh, "Improving perceptual quality of speech in a noisy environment by enhancing temporal envelope and pitch," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 489–492, May 2010, doi: 10.1109/LSP.2010.2044937.
- [12] L. Sukhostat and Y. Imamverdiyev, "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of Voice*, vol. 29, no. 4, pp. 410–417, Jul. 2015, doi: 10.1016/j.jvoice.2014.09.016.
- [13] B. Cardozo and R. Ritsma, "On the perception of imperfect periodicity," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 159–164, Jun. 1968, doi: 10.1109/TAU.1968.1161978.
- [14] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, Oct. 1976, doi: 10.1109/TASSP.1976.1162846.
- [15] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Mar. 2008, pp. 3933–3936. doi: 10.1109/ICASSP.2008.4518514.
- [16] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, Feb. 1977, doi: 10.1109/TASSP.1977.1162905.
- [17] 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, Oct. 1974, doi: 10.1109/TASSP.1974.1162598.
- [18] 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, Dec. 2009, doi: 10.1007/s11760-008-0072-5.
- [19] 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.
- [20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, 1993, pp. 97–110.
- [21] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002, doi: 10.1121/1.1458024.
- [22] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, Feb. 1967, doi: 10.1121/1.1910339.
- [23] H. Kobayashi and T. Shimamura, "A modified cepstrum method for pitch extraction," in *IEEE APCCAS 1998. 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings (Cat. No.98EX242)*, IEEE, pp. 299–302. doi: 10.1109/APCCAS.1998.743751.
- [24] M. A. F. M. R. Hasan, M. S. Rahman, and T. Shimamura, "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.
- [25] 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, Feb. 2014, doi: 10.1109/TASLP.2013.2295918.
- [26] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, Jan. 1988, doi: 10.1121/1.396427.
- [27] B. Li and X. Zhang, "A pitch estimation algorithm for speech in complex noise environments based on the radon transform," *IEEE*





- Access, vol. 11, pp. 9876–9889, 2023, doi: 10.1109/ACCESS.2023.3240181.
- [28] Z. Mnasri, S. Rovetta, and F. Masulli, “A novel pitch detection algorithm based on instantaneous frequency for clean and noisy speech,” *Circuits, Systems, and Signal Processing*, vol. 41, no. 11, pp. 6266–6294, Nov. 2022, doi: 10.1007/s00034-022-02082-8.
- [29] F. Huang and T. Lee, “Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 99–109, Jan. 2013, doi: 10.1109/TASL.2012.2215589.
- [30] W. Chu and A. Alwan, “SAFE: a statistical approach to F0 estimation under clean and noisy conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 933–944, Mar. 2012, doi: 10.1109/TASL.2011.2168518.
- [31] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, “SPICE: self-supervised pitch estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020, doi: 10.1109/TASLP.2020.2982285.
- [32] S. Singh, R. Wang, and Y. Qiu, “DeepF0: end-to-end fundamental frequency estimation for music and speech signals,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 61–65. doi: 10.1109/ICASSP39728.2021.9414050.
- [33] W. Wei, P. Li, Y. Yu, and W. Li, “Harmof0: logarithmic scale dilated convolution for pitch estimation,” May 2022, [Online]. Available: <http://arxiv.org/abs/2205.01019>.
- [34] 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, Dec. 2014, doi: 10.1109/TASLP.2014.2352453.
- [35] F. Plante, G. F. Meyer, and W. A. Ainsworth, “A pitch extraction reference database,” in *4th European Conference on Speech Communication and Technology (Eurospeech 1995)*, ISCA: ISCA, Sep. 1995, pp. 837–840. doi: 10.21437/Eurospeech.1995-191.
- [36] “Multi-lingual speech database for telephony.” 1988. Distributed by NTT Advanced Technology Corp., Jpn. .
- [37] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993, doi: 10.1016/0167-6393(93)90095-3.
- [38] Wcng, “Wireless communication networking group.” [Online]. Available: <http://www.ece.rochester.edu/projects/wcng/code.html>
- [39] M. Brookes, “Voicebox toolkit.” [Online]. Available: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

BIOGRAPHIES OF AUTHORS







Arpita Saha     received her B.Sc. (engineering) degrees in Information and Communication Technology from Comilla University, Cumilla, Bangladesh in 2023. In 2018, she admitted as a student in the Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh. Her current research interests include speech analysis and digital signal processing. She can be contacted at email: arpitasaha2041@gmail.com







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







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



Moinur Rahman     received his B.Sc. (engineering) and M.Sc. (engineering) degrees in Information and Communication Technology from Comilla University, Cumilla, Bangladesh, in 2018 and 2019, respectively. In 2022, he joined as a lecturer in the Department of Computer Science and Engineering, The People's University of Bangladesh, 3/2 Asad Avenue, Dhaka, Bangladesh. Now he is currently serving as a lecturer in the Department of Information Technology, University of Information Technology and Sciences, Baridhara, Dhaka, Bangladesh from March 2023. His current research interests include speech analysis and digital signal processing. He can be contacted at email: pranta7907@gmail.com.



Any Chowdhury     received her B.Sc (engineering) degrees in Information and Communication Technology from Comilla University, Cumilla, Bangladesh in 2023. In 2018, she admitted as a student in the Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh. Her current research interests include speech analysis and digital signal processing. She can be contacted at email: anychowdhury1998@stud.cou.ac.bd.