Pitch Extraction for Speech Signals in Noisy Environments

（雑音環境下での音声信号のピッチ抽出）

Md. Saifur Rahman

A Dissertation Submitted to
the Graduate School of Science and Engineering
in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY
in
Mathematics, Electronics and Informatics

Supervisor: Professor Tetsuya Shimamura

Saitama University, Japan

September 2020

To my beloved daughter

Contents

List of Figures

List of Tables

List of Symbols

$x(n)$	Clean speech signal
m	Lag number
T	Time period
$\phi_{xx}(m)$	ACF of clean speech signal
N	Frame length
$v(n)$	Noise
$\phi_{vv}(m)$	ACF for noise
σ_v^2	Noise variance
$y(n)$	Noisy speech signal
$\phi_{yy}(m)$	ACF of noisy speech signal
$\psi(m)$	AMDF of noisy speech signal
$\zeta(\tau)$	WAF
λ	Small positive constant
$d(\tau)$	Difference function
$C(n)$	Cepstrum of clean speech signal
F	Frequency points
$\phi_{yy-wl}(m)$	Windowless ACF
$C_{wlacf}(n)$	Windowless ACF based CEP
$e(l)$	Error rate
l	Frame number
$F_{est}(l)$	Estimated fundamental frequency
$F_{true}(l)$	Ground true fundamental frequency
η	Clipping threshold level
$s(n)$	Voiced speech signal
n	Discrete time
a_i	Amplitude of each sinusoid
R	Number of sinusoids
F_0	Fundamental frequency
w	Angular frequency
$\delta(w)$	Direc delta function
$y_f(n)$	Framed signal
$P_f^S(w)$	Power spectrum
$P_{f,l}^y(k)$	Band pass filtered power spectrum
M	Subframe length

D	Frame shift
$\bar{P}_f^y(k)$	Accumulated power spectrum

Acknowledgements

First and foremost, all praises belong to Allah, *The Almighty*, who has given me the opportunity and uncountable blessings on me to complete the thesis.

I would like to express my deepest and sincere gratitude to my supervisor, Professor Tetsuya Shimamura for his continuous support in every step of my Ph.D, for being patient and supportive during the up and downs of this period of my life, for being always there to give encouragement, invaluable guidance and motivation. Without his inspiration and support my dream have never been true.

I would like to give thanks to my doctoral committee members Professor Takashi Komuro, Professor Yoshinori Kobayashi and Professor Jun Ohkubo for their valuable comments and suggestions to improve the quality of my thesis. Also, I would like to express my special thanks to Dr. Yosuke Sugiura, for the valuable guidance, suggestions and help throughout my research work.

I am also grateful to all my family members, especially, my parents Dr. Md. Sydur Rahman & Mrs. Hasina Begum, and sisters Umma Mustab Shira & Umma Sadia Tabassum for their endless prayers, love, and constant encouragement. I am also grateful to my in-laws for their prayers, love, moral support and kind cooperation.

My special heartfelt gratitude goes to my beloved wife, Dr. Nargis Parvin who has provided me a sweet home full of understanding, patience, and encouragement during the stressful times. Without her sacrifices, I would never be able to complete the PhD study. My sweet gratitude goes to my daughter, Tajrian Mahdia whose cute smiling face gave me the energy to work harder during this study period.

Finally, I would like to thank to all of my teachers and friends in home and abroad for their mental support.

Abstract

The pitch period is defined as the inverse of the fundamental frequency of the excitation source from the voiced speech signal. The pitch period (in short, pitch) or fundamental frequency is a prominent parameter of speech and highly applicable for speech-related systems such as speech coding, speech recognition, speech enhancement, speech synthesis and so on. The pitch and fundamental frequency so as to give the same meaning, while the pitch is inherently interpreted as the perception of the fundamental frequency. The pitch is generated from the vibration of the vocal cord causing periodicity in the speech signal.

Pitch extraction has proven to be a difficult task even for speech in a noise-free environment. The clean speech waveform is not really periodic; it is quasi-periodic and non-stationary. Although a large number of pitch extraction methods have been reported to deal with the noise-free environment. On the contrary, the least number of researchers attempt to extract the pitch in noisy environments. Under noisy environments, the periodic structure of the speech signal is destroyed so that the pitch extraction becomes an extremely complicated task. Therefore, the reliability and accuracy of the pitch extraction methods face real challenges in noisy environments.

From the above observations, the objective of this dissertation is to develop some approaches which are effective to handle the speech signals in the real application without any complicated post processing where speech signals are corrupted by noise. Some conventional state-of-the art approaches rely on a complicated post processing technique for pitch extraction. In this dissertation, we focus on simple and efficient approaches that are proposed and implemented to solve the factors that degrade the performance of pitch extraction methods.

In this dissertation, firstly, we propose the use of fourth-root spectrum instead of log spectrum for increasing the pitch extraction accuracy in noisy environments. To get clear harmonics, lifter and clipping operations are followed. When the resulting spectrum is transformed in the time domain by means of discrete Fourier transform, the pitch extraction is robust against narrow-band noise. When the above resulting spectrum is amplified by a power calculation and transformed in the time domain, the pitch extraction is robust against wide-band noise. These properties are investigated through exhaustive experiments in a variety of noise types. Computational time to be required is also studied. The experimental results based on above properties demonstrate the effectiveness of the new approaches for

improving the performance of the pitch extraction. Also, the performance of this method sometimes deteriorates by the windowing effect. This method utilizes Hanning window function which does not better perform to extract pitch in the noisy environments.

To improve the performance of the extraction accuracy, the second approach considers an advancing trend of recent techniques for pitch extraction of speech in noisy environments, windowing effects are discussed analytically, and it is insisted that the Rectangular window should be proactively used instead of the popular Hanning or Hamming window. In a variety of noise environments, a performance comparison of the conventional pitch extraction methods is conducted, and as a result, we take a standpoint to support the autocorrelation (ACF) method. Incorporating accumulation techniques, three types of pitch extraction approaches are developed. Through experiments, it is shown that the proposed approaches have the potential to provide better performance for pitch extraction without relying on a complicated post processing technique.

Chapter 1

Introduction

1.1 Background

Speech is composed of spoken words and sentences which is the most natural communication mode in daily life. This is the main form to use humans and interacts with each other. The speech is a sound wave of air, which is generated random air fluctuations from the lungs through the vocal cords and in vocal tract. According to the vibration of vocal cord or not, the speech is defined as the voiced or unvoiced speech. Voiced speech is produced when the vocal cords vibrate at a particular frequency. This frequency is referred to as the fundamental frequency (pitch) of the speech signal. On the contrary, the generation process of unvoiced speech does not involve the use of the vocal cords. Therefore, the fundamental frequency is the absent in the unvoiced signal. The fundamental frequency is perceived as pitch level, a low value of the fundamental frequency is perceived as a low pitch and a high value of fundamental frequency is perceived as a high pitch [1-6].

The most important purpose of speech is communication. In speech communication systems, it is difficult to accurately focus on the representation of the speech signal in a convenient form, and the preservation of the message content in the speech signal. The acoustic features, vocal tract shape, formant frequencies, and bandwidths and pitch are associated with the representation of speech signals and have profound applications in speech recognition, synthesis, and coding [7-14]. The performance of the above application systems are highly influenced by the accuracy of pitch extraction.

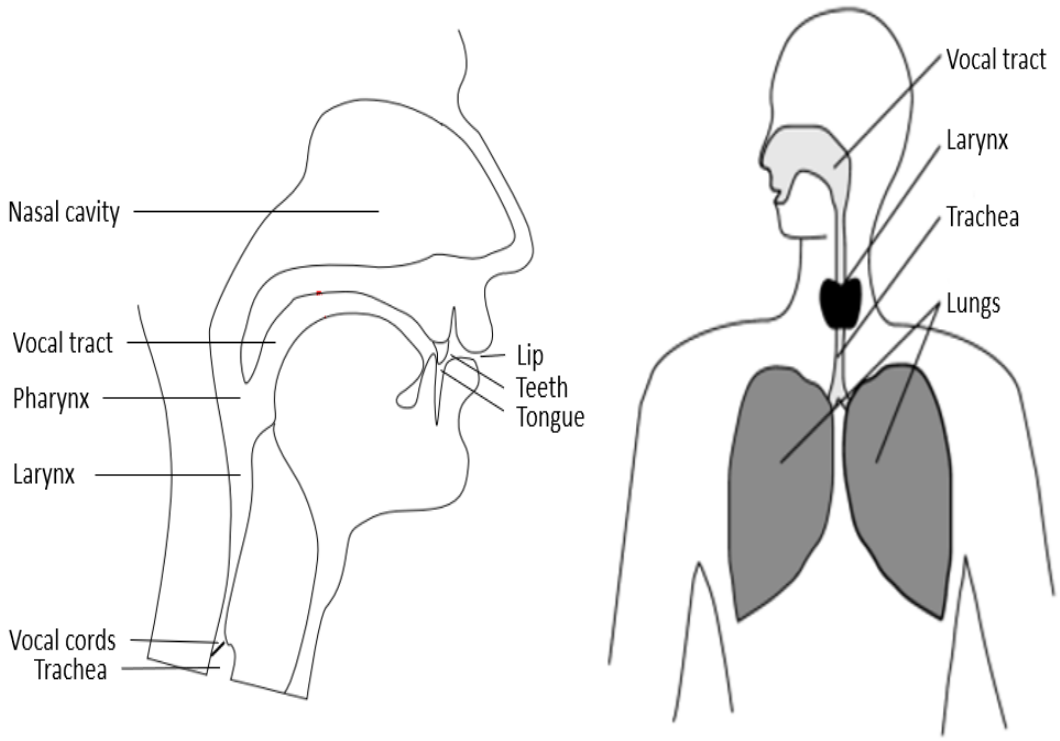

Figure 1.1: Schematic model for human speech production with different parts.

Pitch is one of the powerful speech analysis techniques and has the ability to represent the speech signal accurately and efficiently. The pitch changes in response to the stress, intonation, emotion, and size of the vocal cords at any given instant. The stress and intonation parameter of speech is most important for the pitch to identify the phoneme. In the case of the emotional state of the speaker, sometimes joy produces wide range of pitch, while sadness produces a narrow range of the pitch. Since, the average size of vocal cords in male speaker is larger than that of the female speaker, the average pitch of an adult male will be lower than females for the same utterance. The possible pitch range for men is found between 50-250 [Hz], while for women the range falls between 120-500 [Hz] [15].

1.2 Speech Production Mechanism

Figure 1.1 [2][16] shows the schematic diagram of the human speech production. The main components of the human speech production system are: the lungs, trachea, vocal cords, larynx, pharynx, vocal tract, nasal cavity. The components can be described as follows, by using the muscle force, the lungs generate the air

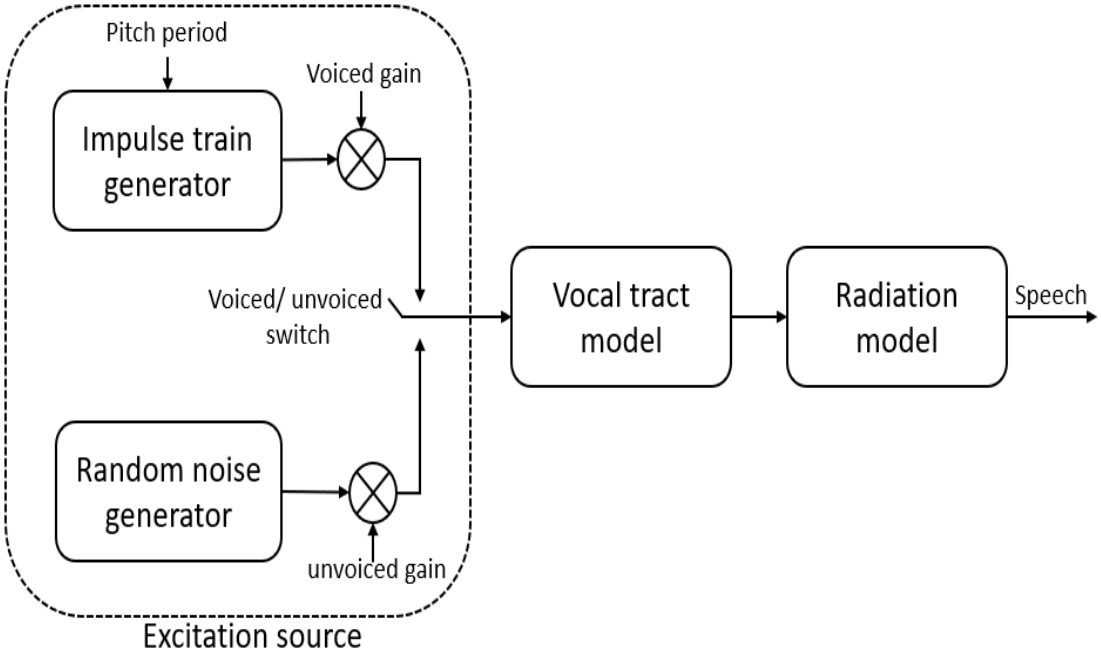

Figure 1.2: Source-filter model for speech production.

is streaming through the larynx, where it can be modulated and transferred to the vocal tract. Depending on the activities of the vocal folds, the air produce a quasi-periodic wave which is said to be voiced or unvoiced. On the other hand, vocal tract including the pharynx, oral cavity, and nasal cavity, which can be modeled as an acoustic tube with variable resonator. For both voiced and unvoiced generation, the vocal tract behaves as filter, which provides the pressure wave from the vocal cords and then amplifies to the lips or the nostrils. Speech is simply as an acoustic wave that is radiated from the system, when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction in the vocal tract [2].

According to the mode of excitation, speech sound can be divided into three categories such as voiced, unvoiced, and silence. Voiced sound are produced by forcing air through the trachea. After that vocal folds vibrate in a relaxation oscillation, there by producing quasi-periodic pulses of air which excites the vocal tract. The vocal cords vibrate at a particular frequency. This frequency is represented as the pitch or fundamental frequency. For example, all vowels or the consonant $/m/$ are of this type. Unvoiced sound are generated without any vibration of the vocal cords, simply forcing air through the constriction at a high enough velocity to produce turbulence. This creates a noise source to excite the vocal tract. In unvoiced speech, the pitch or fundamental frequency is absent. For example, $/f/$

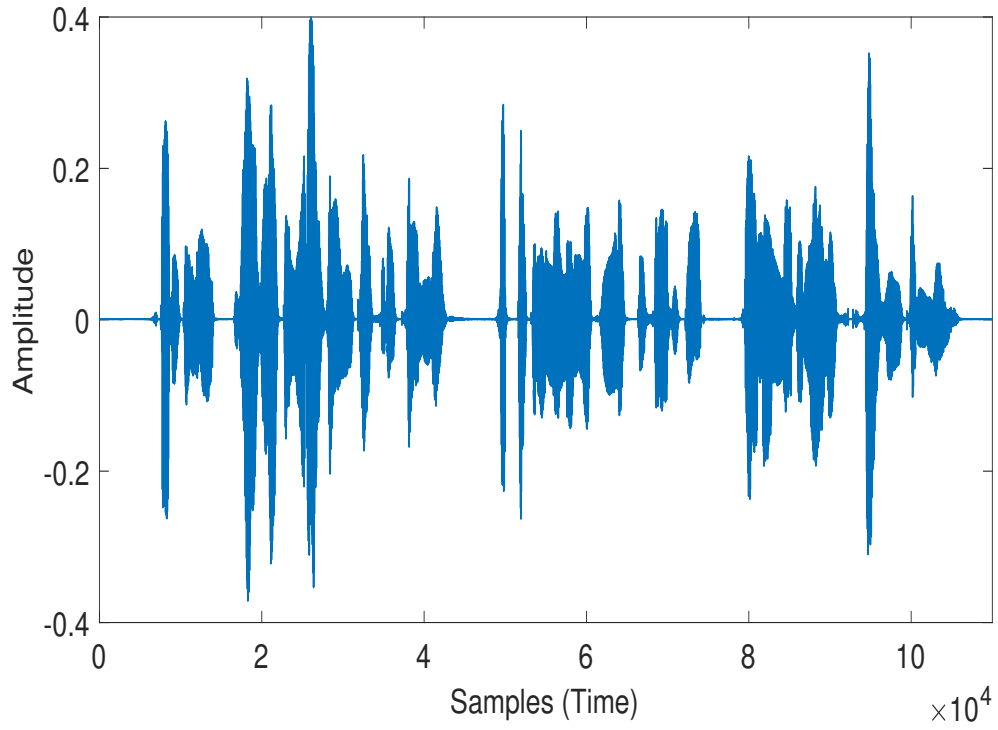

Figure 1.3: Full duration of clean speech signal.

and /s/ are of this type. In the case of silence, where no speech is produced.

The production of speech can also be viewed as a filtering operation in which a excitation source excites a vocal tract filter is known as source-filter model as shown in Fig. 1.2 [17]. In Fig. 1.2, the excitation source is the combination of the impulse train generator with the period according with the vibration of the vocal cords or a randomly distributed noise generator. Thus, speech signals can be computed as the convolution of time-varying vocal-tract system with time-varying excitation source. As a result, speech signals are time-variant in nature which is represented as in Fig. 1.3.

1.3 Challenges of Pitch Extraction

Many researchers have been concentrated on research and development about pitch extraction from a speech signal for more than 50 years. No one can able to develop the closed-form solution or error-free method, which is highly efficient and effective for pitch extraction up to now. The factors are involved to make difficulty for extracting the accurate and reliable pitch from the speech signal, which is as

follows.

- The glottal excitation waveform is a quasi-periodic which is not a perfect periodic train of pulses. It is difficult, even for a researcher, to find the accurate pitch period of a speech waveform, which changes drastically both in period and structure of the speech waveform depending on time.
- Another most important factor is to face the difficulty to extract accurate pitch, when glottal excitation waveform is interacted with the parameters of the vocal tract [18][19]. Such interactions are harmful to pitch extraction during rapid variations of the articulators as well as simultaneously rapidly changing the parameters of the vocal tract.
- For the purpose of pitch extraction, speech is combined into two states; a voiced state with a harmonic structure and an unvoiced state with a noise-like structure. The voiced/unvoiced states also affect the extraction accuracy of pitch.
- To increase the extraction accuracy, we face another challenging error, which is the choice of beginning and ending locations of the pitch period during voiced speech segments. For this reason, the spurious pitch period arises.
- Speech selection can also complicate the pitch extraction and result in increased extraction errors. For example, it is extremely difficult to distinguish between unvoiced speech and low level voiced speech in pitch extraction.
- In practical applications, the background noise can also affect the performance of the pitch extraction. Particularly in the case of mobile communication environments where noise affected is the general scenario. On the other hand, the signal generation process is also affected by the background noise.

1.4 Motivation of the Thesis

It should be pointed out that the accuracy of the pitch extraction methods is affected when the speech signal is corrupted by noise. Also, if the noise level is high, the extraction accuracy is also affected drastically. Most of the researchers had invented the pitch extraction method in clean speech conditions. When clean speech is contaminated by noise, especially in real-world noise, few numbers of

researchers concentrated on pitch extraction. Moreover, these methods provide satisfactory performance only at moderately low to high levels of SNR by relying on the complicated post-processing technique. Also, these methods require the high processing time. Actually, the processing time is also a growing demand for practical applications. From the above point of view, it is essential to concentrate on pitch extraction research where the real world noise is present which indicates the low level of SNR. It also should be focused on processing time. Our motivation in this dissertation is that we investigate and develop simple and efficient pitch extraction methods that are used in real-world noise including low SNRs without relying on any complicated post-processing.

1.5 Organization of the Thesis

The background, speech production mechanism, challenges of pitch extraction, and motivation of this thesis have been reviewed in this chapter. The rest of the thesis is organized as follows.

Chapter 2 describes the segmentation of each frame speech signal in speech analysis. After that, this thesis presents more details about the operation of some time domain and frequency domain based pitch extraction methods.

Chapter 3 presents the detailed development of the proposed pitch extraction method using the concept of the fourth-root spectrum in noisy speech. Firstly, the general pitch estimation methods and motivation of the proposed methods are described. Then, the theoretical concept of the FROOT and FROOT+ pitch extraction method is presented. The proposed pitch extraction methods are performed by measuring the percentage of average gross pitch error (GPE). In these experiments, Chapter 3 includes the preliminary experiments for the selection of the constant value in the clipping threshold level, and discusses the analysis from the obtained evaluation results. Finally, Chapter 3 explains the the noise effect in each method, processing time and the conclusion are presented.

Chapter 4 continues the related literature review, which is followed by the description of the pitch extraction methods. Then, Chapter 4, analytically describes the motivation of using the Rectangular window and explains the proposed accumulation based methods by utilizing the Rectangular window. It also discusses the performance of the proposed methods with that of the conventional methods by measuring the percentage of GPE. Finally, processing time and conclusion are presented.

Chapter 5 contains a summary of the conclusion and the suggestions for the future work of the thesis.

Chapter 2

Discussion on Pitch Extraction Methods

2.1 Background

Pitch is extracted from the analysis of speech signal where speech analysis is generally performed using short-time analysis in time and frequency domains. For this reason, the speech signal should be experimented in a short-term analysis, where the signal is divided into short time by using the window function (5-100 [ms]). The short window of a signal is called frame. By multiplying the input signal with a window function, the windowed signal also goes to zero at the border such that the discontinuity at the border becomes invisible as shown in Fig. 2.1.

Traditionally, different window functions are used in speech processing depending on the situations. In the field of pitch extraction, most of the researchers are used the Rectangular, Hanning, and Hamming windows [2], respectively, which are represented by,

$$Win_{rec}(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

$$Win_{han}(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n/(N-1)) & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$$Win_{ham}(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

By utilizing the short-time windowed speech signal, we can easily extract the

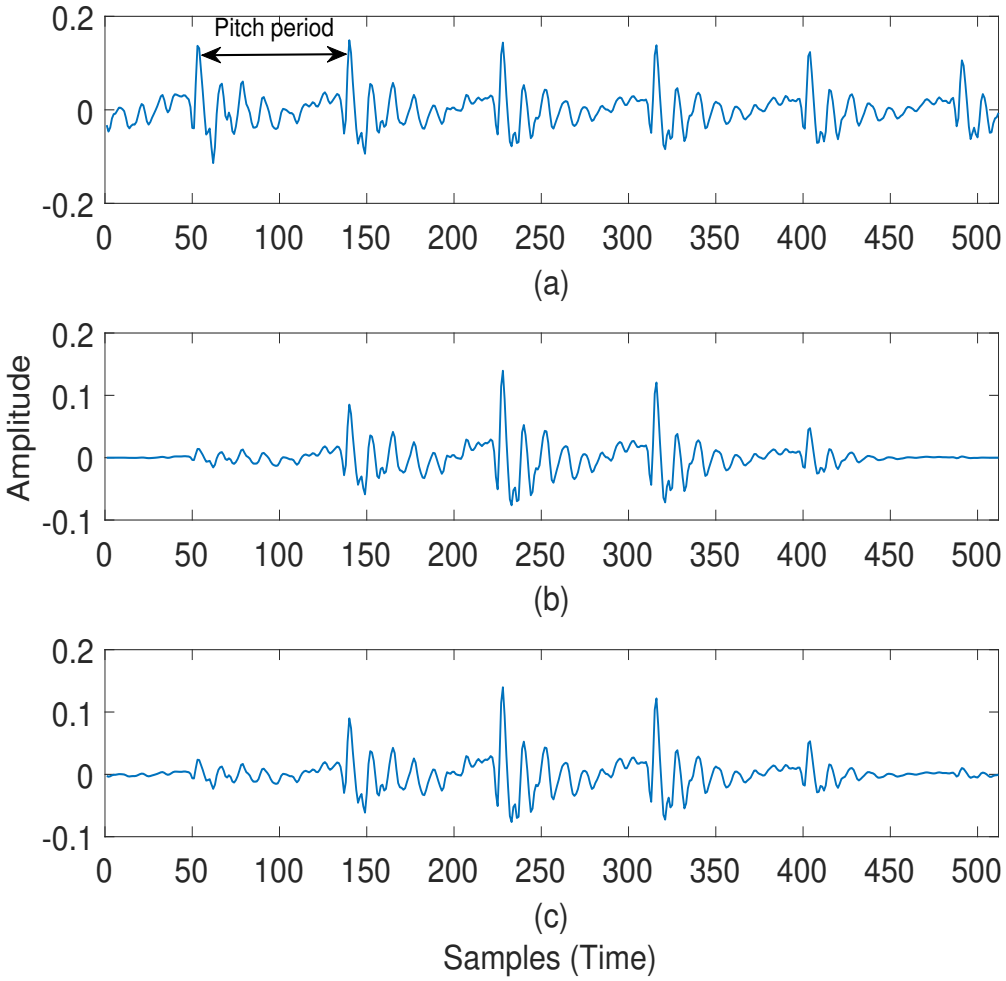

Figure 2.1: Framed clean speech signal (a) using Rectangular window, (b) using Hanning window, (c) using Hamming window.

pitch. However, pitch extraction has proven to be a difficult task, even for speech in a noise-free environment. The clean speech waveform is not really periodic; it is quasi-periodic and highly non-stationary which is shown in Fig. 2.2. In this figure, pitch period is present, where the fundamental frequency or pitch is the reciprocal of the pitch period. On the other hand, when the speech signal is corrupted by noise i.e noisy speech signal, which is shown in Fig. 2.3. In noisy environments, the reliability and accuracy of pitch extraction algorithms face real challenges. Under noisy conditions, the speech peaks is affected by noise peaks. Therefore, the periodic structure of the speech signal is destroyed. So that, the pitch extraction becomes an extremely complicated task. Therefore, a lot of pitch

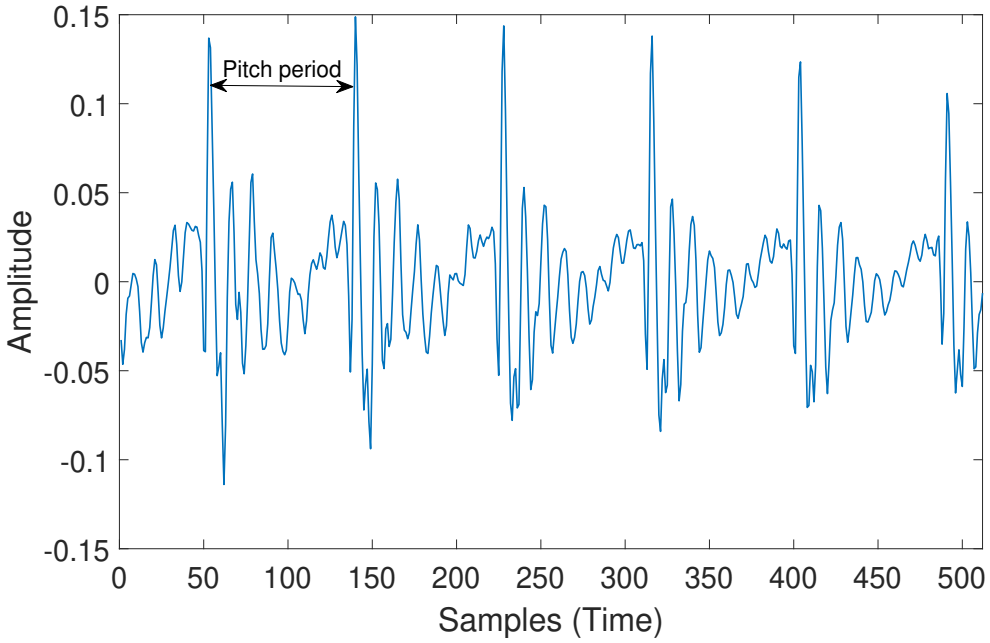

Figure 2.2: Framed clean speech signal.

extraction methods have been addressed for clean speech while pitch extraction methods from noisy speech has been addressed only by a few researchers.

2.2 Pitch Extraction Methods

This section presents some well known pitch extraction algorithms that appear in the literature. These pitch extraction methods have been utilized either in the time domain or frequency domain or in both domains of the speech signal, up to now. In the time domain pitch extraction methods [18-29], the measurements are directly applied on the speech waveform. Therefore, peak position, valley point, zero-crossing, autocorrelation, and number of measurement influence the time domain pitch extraction methods. Traditionally, time domain methods are efficient for extracting the accurate pitch period when quasi-periodic signal has been processed by reducing the effect of vocal tract characteristics. On the other hand, frequency domain pitch detectors [30-35] depend on the speech signal when it maintains the periodicity in the time domain. After that, the spectrum of the signal in frequency domain consists of series of impulses and its harmonics which is emphasized to extract the more accurate pitch period of the signal. Hybrid pitch detector methods [36-37] are used both time domain and frequency domain

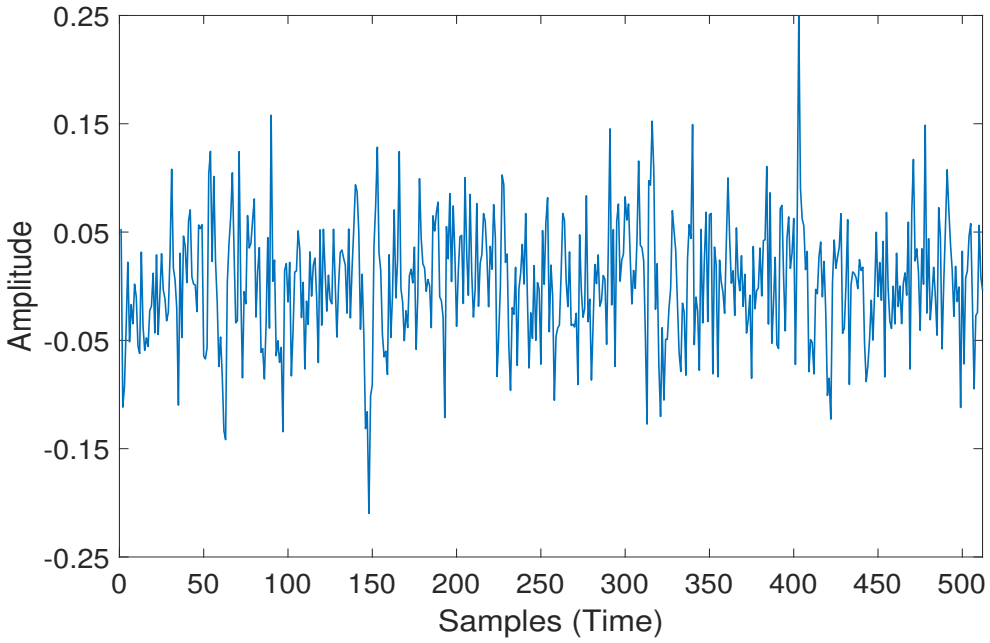

Figure 2.3: Framed noisy speech signal (0 [dB], white noise).

to extract the accurate pitch. For example, some time hybrid pitch extraction methods utilize frequency domain technique to reduce the effect of vocal tract characteristics and then time domain technique is applied to extract the more accurate pitch period.

In order to compare the performance of the proposed method in Chapters 3 and 4, the conventional pitch extraction algorithms were additionally implemented which are the time domain (such as autocorrelation function (ACF) [24], average magnitude difference function (AMDF) [25], weighted autocorrelation function (WAF) [26], and YIN [27]) and frequency domain based methods (cepstrum (CEP) [29-31], modified CEP (MCEP) [32], and windowless autocorrelation function based CEP (WLACF-CEP) [33]). These methods are chosen because they are highly effective against the random noise as well as reduce the vocal tract characteristics. In this chapter, the above methods are discussed.

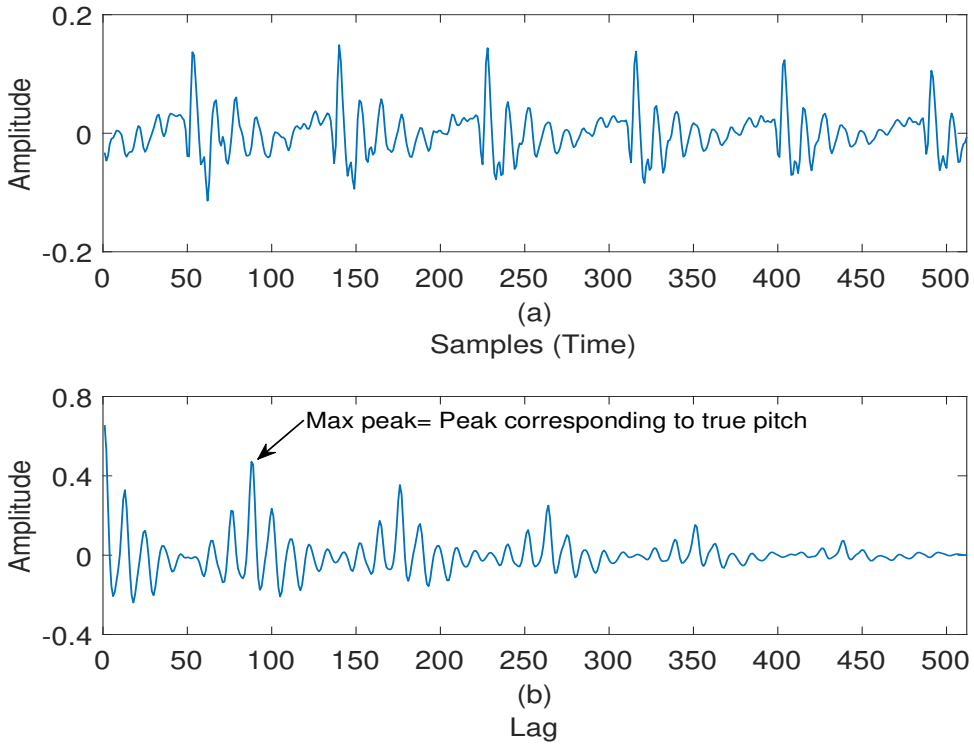

Figure 2.4: ACF in clean speech signal (a) input signal (b) ACF in (a).

2.3 Methods in Time Domain

2.3.1 Autocorrelation Function (ACF)

Autocorrelation function (ACF) [24] is one of the correlation-based method which is highly efficient for extracting the pitch period, because it can measure similarity between signals by simple computation, and represents a large peak at the lag corresponding to the pitch period. For a given signal $x(n)$, defined for all n, the ACF is defined as

$$\phi_{xx}(m) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+m) \tag{2.4}$$

where N is the frame length and m is the lag number. If $x(n)$ is periodic with period T , then the ACF, $\phi_{xx}(m)$ is also periodic with the period of T and shows peaks at the locations of kT where k is an integer ($k = 0, 1, 2, 3....$). Thus, the periodic signal is also a periodic signal with the period, the period of the autocorrelation function reflects the period of input signal. The ACF uses the location of the

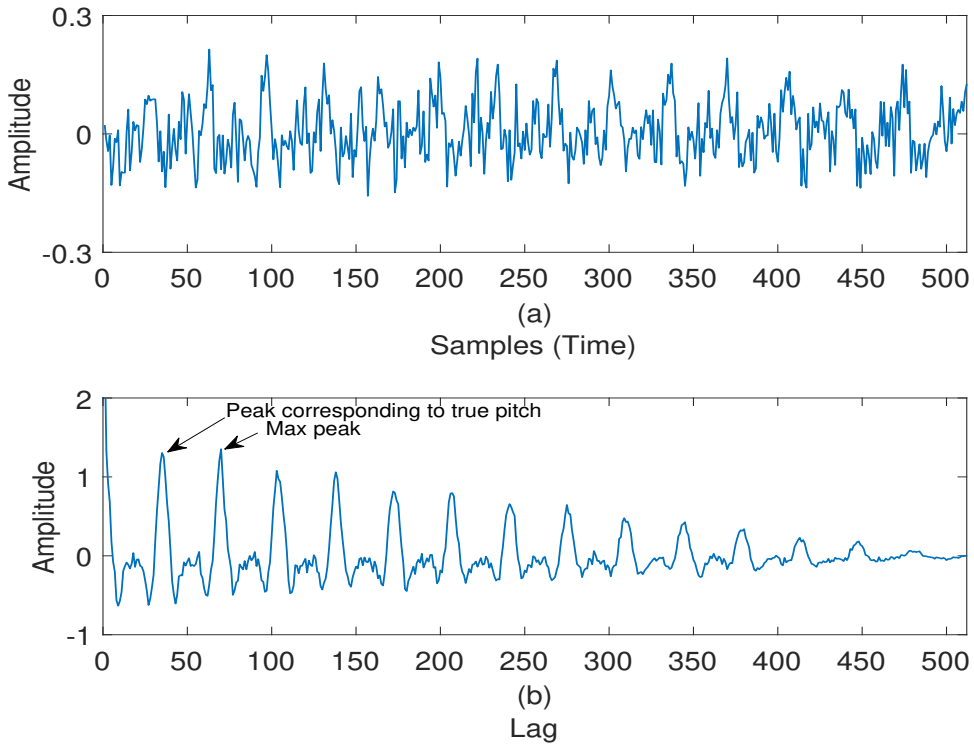

Figure 2.5: ACF in noisy speech signal (a) noisy speech signal at 0 dB SNR (white noise) (b) ACF in (a).

second largest peak relative to the largest peak (at $m = 0$) to obtain an estimate of the pitch period such as shown in Fig. 2.4.

In general, the white noise $v(n)$ is uncorrelated with signal $x(n)$ for all values of m except for $m = 0$. The ACF of $v(n)$, $\phi_{vv}(m)$ is defined as

$$\phi_{vv}(m) = \begin{cases} \sigma_v^2 & \text{for } m = 0, \\ 0 & \text{for } m \neq 0, \end{cases} \tag{2.5}$$

where σ_v^2 is the noise variance of $v(n)$. The autocorrelation of a white noise process will be an impulse function at lag $m = 0$. In the presence of noise, the autocorrelation of the noisy speech $y(n)$, $\phi_{yy}(m)$ is computed as

$$\phi_{yy}(m) = \phi_{xx}(m) + \phi_{vv}(m) \tag{2.6}$$

In (2.6), such method are robust against random noise, because only the first lag of noise is affected with clean speech. Thus, it can reduce the high-frequency

components and emphasize the low-frequency components to extract the accurate pitch. On the other hand, the pitch extraction performance of the ACF-based methods are degraded when clean speech is corrupted by color noise. In the ACF of color noise, the clean speech is affected by noise with the increases of the values of lag, m. $\phi_{yy}(m)$ in (2.6) can also be written as

$$\phi_{yy}(m) = \phi_h(m) * \phi_p(m) \tag{2.7}$$

where $\phi_h(m)$ and $\phi_p(m)$ are the autocorrelation functions of the vocal tract and the vocal source, respectively, and $*$ denotes the convolution. Eq. (2.7) indicates that the $\phi_{yy}(m)$ is highly influenced by the vocal tract information $\phi_h(m)$, which makes difficult to detect more appropriate pitch as shown in Fig. 2.5.

2.3.2 Average Magnitude Difference Function (AMDF)

Correlation based processing also includes the average magnitude difference function (AMDF) method [25]. The AMDF, $\psi(m)$ treats as a difference between the original speech signal and its delayed version, which is defined as

$$\psi(m) = \frac{1}{N}\sum_{n=0}^{N-1} |y(n) - y(n+m)| \tag{2.8}$$

It shows almost similar properties with the ACF. ACF exhibits the peaks at delays corresponding to the pitch period. While the difference signal always exhibits deep nulls at delays corresponding to the pitch period of voiced sounds having a quasi-periodic structure as shown in Fig. 2.6. The magnitude of the global minimum of AMDF is severely affected by intensity variation and background noise of the speech signal causing pitch extraction errors as shown in Fig. 2.7.

2.3.3 Weighted Autocorrelation Function (WAF)

The autocorrelation function (ACF) [24] is a straightforward computation one in the time domain and shows effectiveness against wide-band random noise such as white noise. The ACF corresponds to a correlation calculation between the input speech signal and its delayed version in the time domain, but it is also obtained by the inverse Fourier transform of the power spectrum of the speech signal in the frequency domain which satisfy the Wiener–Khinchin theorem. The ACF is, however, affected by the characteristics of the vocal tract. Thus, spurious peaks are

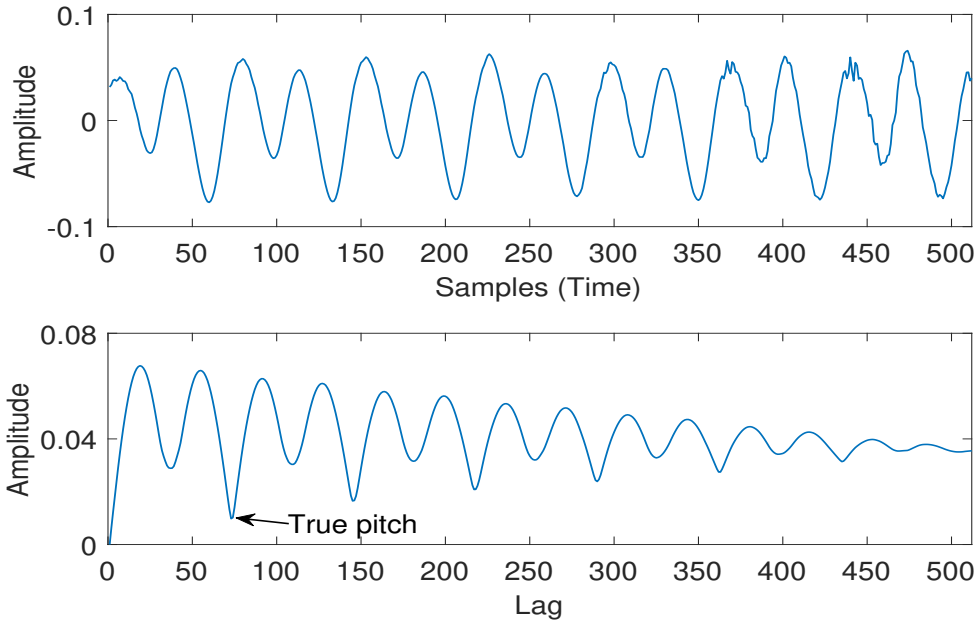

Figure 2.6: AMDF in clean speech signal (a) input signal (b) AMDF in (a).

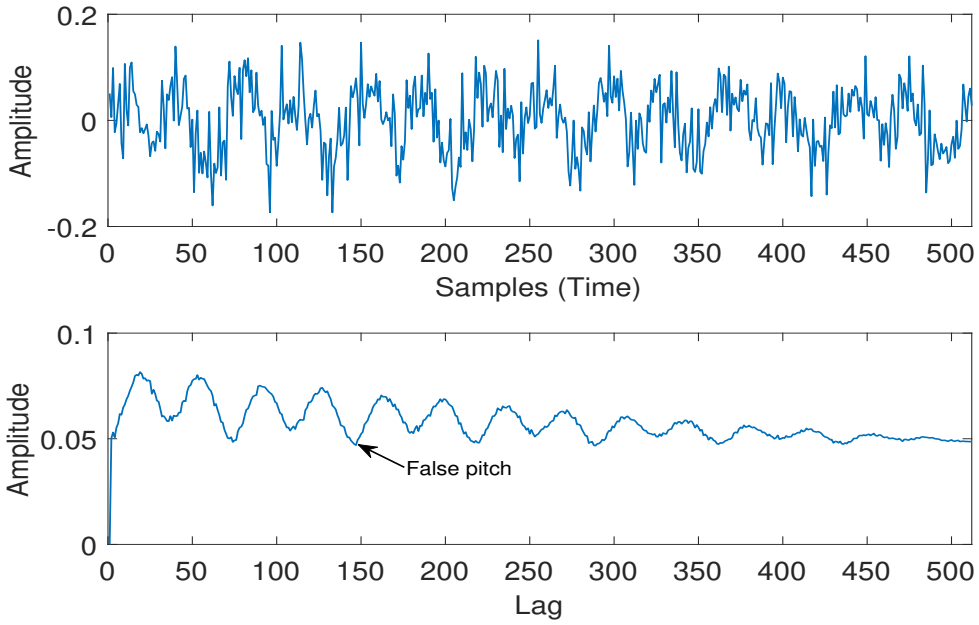

Figure 2.7: AMDF in noisy speech signal (a) noisy speech signal at 0 dB SNR (white noise) (b) AMDF in (a).

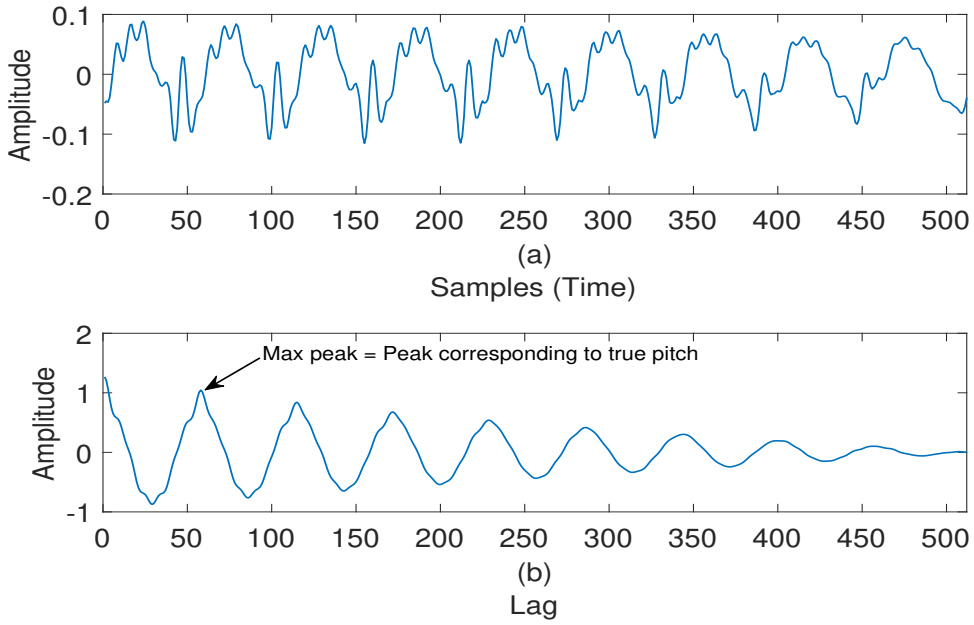

Figure 2.8: WAF in clean speech signal (a) input signal (b) WAF in (a).

also sometimes introduced in the spectrum in noisy or even in noiseless conditions. These peaks also sometimes makes true peak selection a difficult task. On the other hand, the AMDF [25] gives rise to a falling tendency with the increase of the delay number, which makes difficult to detect the appropriate valley point, therefore, inaccurate pitch extraction will be obtained. To solve this problem, WAF can emphasize the true peak and suppresses the non-pitch peaks as well as noise components by dividing ACF with the AMDF, which is represented as WAF. In WAF [26], the ACF and AMDF are used as numerator and denominator parts, respectively. For a noisy signal $y(n)$, $n = 0, 1, ...N-1$, at a frame, the WAF, $\zeta(\tau)$ is denoted as

$$\zeta(\tau) = \frac{\phi_{yy}(\tau)}{\psi(\tau) + \lambda} \tag{2.9}$$

where λ is a small positive constant to avoid the division by 0.

In WAF, the ACF creates a strong peak at the location which corresponds to the pitch period. On the other hand, the inverse of the AMDF, 1/AMDF, also creates a strong peak at the same location, because the AMDF itself creates a deep notch at the same location. Therefore, these peaks are emphasized in the WAF. Furthermore, the components of noise included in noisy speech may be distributed in the waveform of the ACF for a certain range of lags. In the AMDF

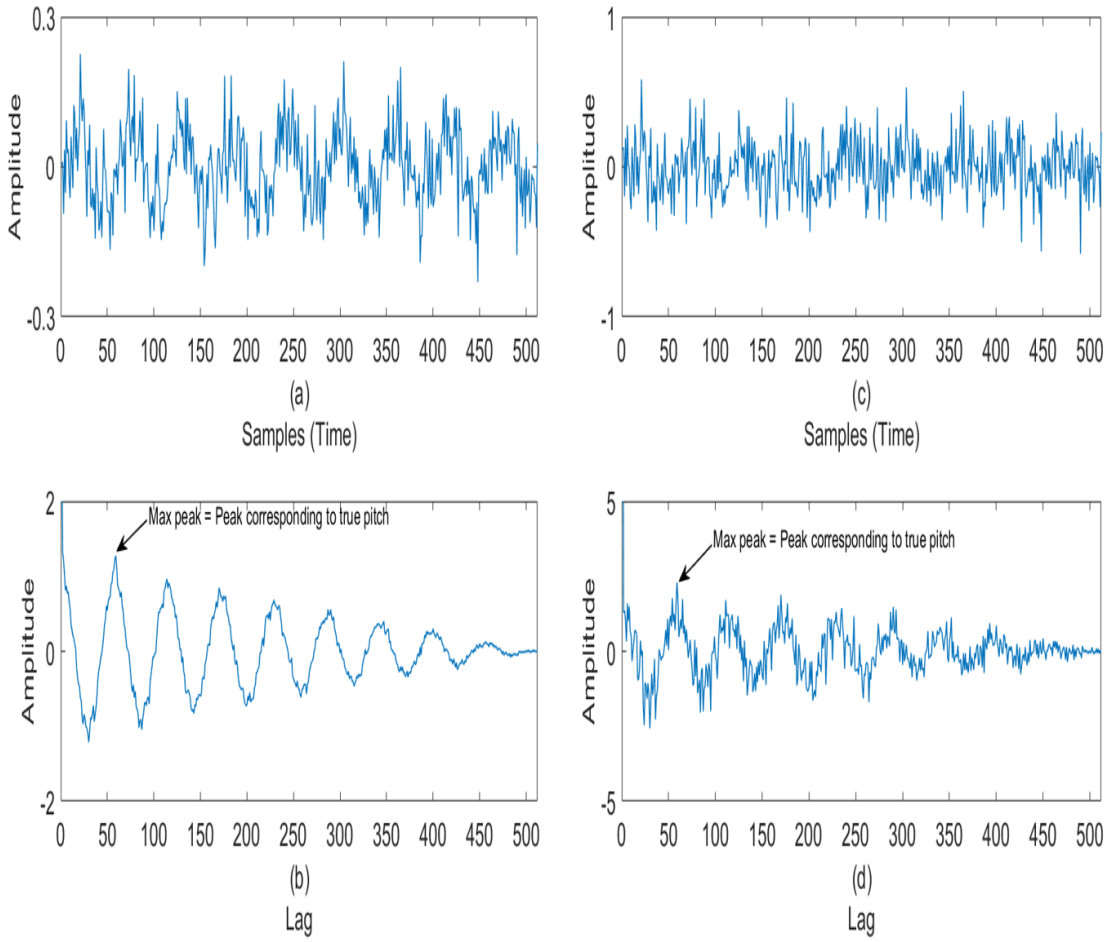

Figure 2.9: WAF in noisy speech signal (a) input signal at 0 dB SNR (white noise) (b) WAF in (a) (c) input signal at -10 dB SNR (white noise) (d) WAF in (c)

case, also a similar phenomenon is expected in the waveform of the AMDF (actually the inverse of the AMDF). However, these noisy components would behave in a different manner in the ACF and AMDF, which are uncorrelated as validated in [26]. Therefore, a multiplication of the two functions; the ACF and 1/AMDF, suppresses the noise components each other. This also leads to suppress the unwanted peaks created in the ACF and 1/AMDF each other. Hence, the WAF efficiently emphasizes the pitch peak and provides an excellent performance of pitch extraction in noisy environments as shown in Fig. 2.9 (b). Contrary, at high SNR level (-10 [dB]) as shown in Fig. 2.9 (d), the WAF emphasizes to extract the accurate pitch period but it is affected by the vocal tract characteristics, causing the error rate is increased.

2.3.4 YIN Method

Among the correlation based pitch extraction methods in time domain, YIN [27] is one of the most important approach which is concentrated in speech and music signals. The YIN method is introduced by A. de Cheveigne and H. Kawahara in 2002, which is followed some steps.

- Difference function: Finding the difference function which provides the almost similar properties of the ACF. The difference function is highly emphasized to reduce pitch extraction errors by utilizing square of the difference between the original speech signal and its delayed version. The difference function, $d(\tau)$ is defined as

$$d(\tau) = \sum_{n=0}^{N-1} (y(n) - y(n+\tau))^2 \tag{2.10}$$

- Cumulative mean normalized difference function: In order to utilize the quasi-periodic nature of speech signal for pitch extraction, the YIN algorithm normalizes the difference function by its cumulative mean, is represented as

$$d'(\tau) = \begin{cases} 1, & \text{for } \tau = 0, \\ \frac{d(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d(j)} & \text{otherwise,} \end{cases} \tag{2.11}$$

- Another number of steps in YIN algorithm are absolute threshold, parabolic interpolation, and best local search, respectively, to determine the list of candidates, true pitch extraction of the whole period, and search the analysis point for the better estimation, respectively.

YIN is a successful and robust time-domain algorithm for pitch extraction in noise free environment, but in noisy environments, YIN algorithm is not more effective.

2.4 Methods in Frequency Domain

2.4.1 Cepstrum (CEP) Method

In the CEP [29]-[31] method, the pitch is extracted by utilizing the inverse discrete Fourier transform (IDFT) of the log-amplitude spectrum, which is more effective

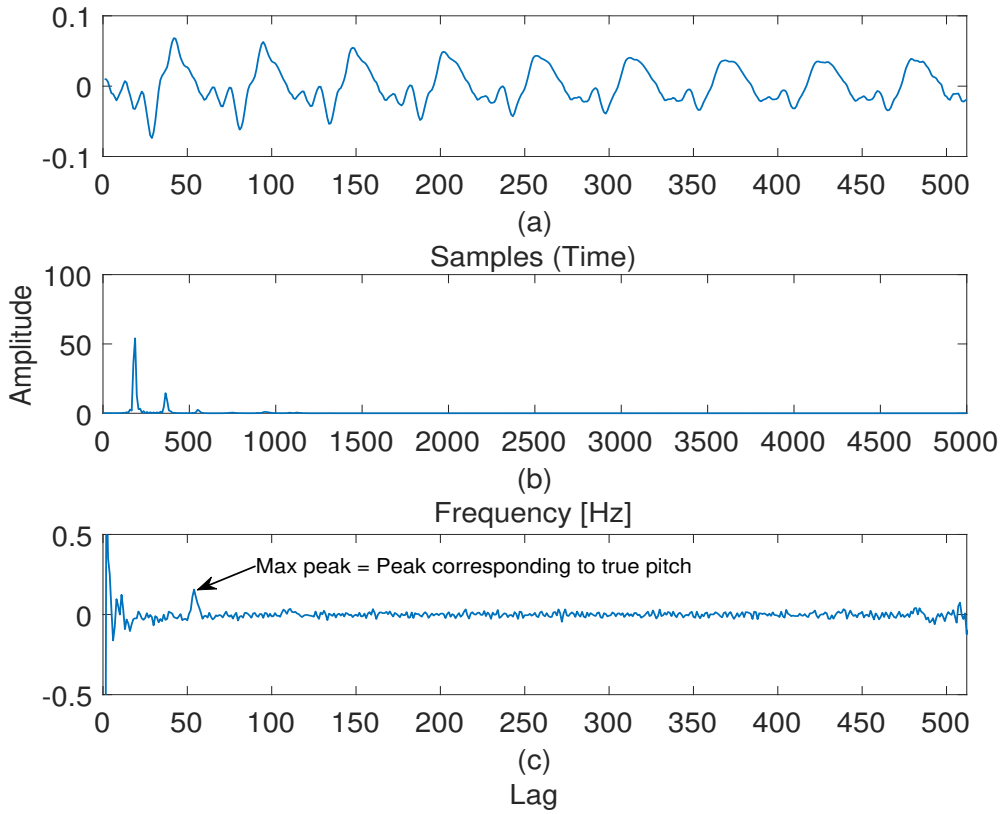

Figure 2.10: CEP in step-by-step (a) clean speech signal (b) log-amplitude spectrum in (a), (c) CEP output.

in clean speech. A speech signal may be modelled as the convolution of excitation source information, $e(t)$ and vocal tract information and a glottal input $v(t)$ in time domain. Alternatively, in frequency domain, the spectrum of the speech signal is the product of excitation source information, $E(f)$ and vocal tract information and a glottal input $V(f)$. The speech signal can be expressed as in time domain and frequency domain, respectively,

$$x(t) = e(t) * v(t) \tag{2.12}$$

$$X(f) = E(f).V(f) \tag{2.13}$$

The CEP of $x(t)$, $C(n)$ can be obtained as

$$C(n) = \frac{1}{F}\sum_{f=0}^{F-1} log|X(f)|e^{\frac{j2\pi fn}{F}} \tag{2.14}$$

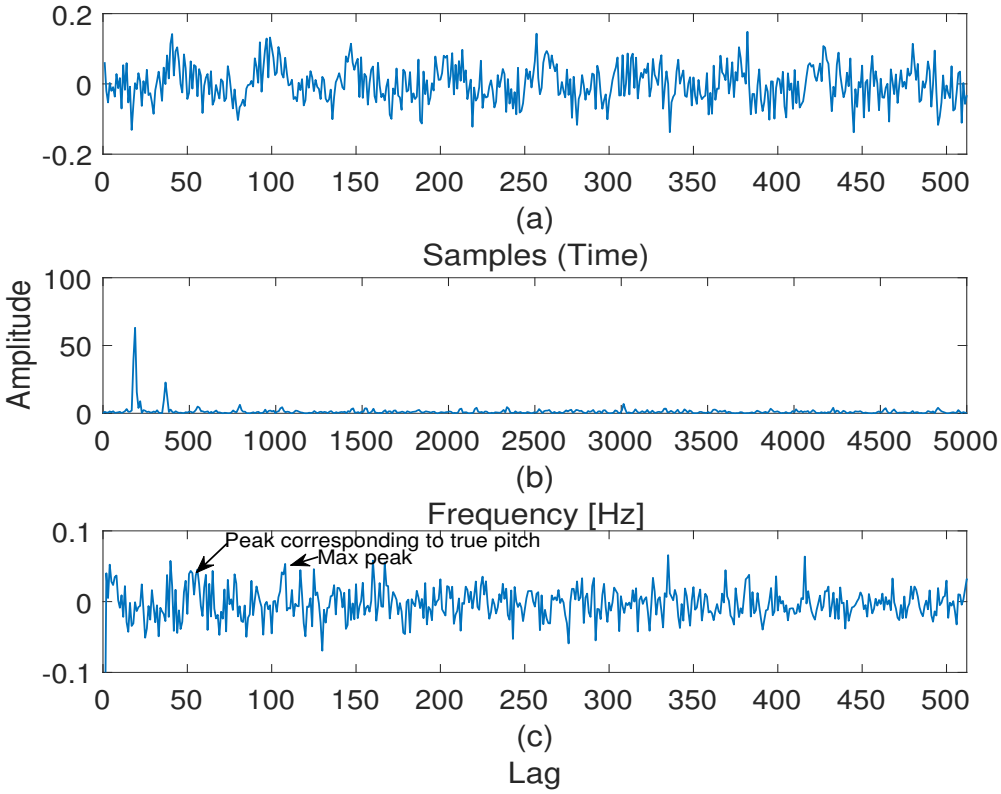

Figure 2.11: CEP in step-by-step (a) input signal at 0 [dB] SNR (white noise) (b) log-amplitude spectrum in (a), (c) CEP output.

where $X(f)$ is the DFT of $x(t)$ with F frequency points. By utilizing the properties of (2.12) and (2.13), it can be represented as in (2.14),

$$C'(n) = \frac{1}{F}\sum_{f=0}^{F-1}[log|E(f)| + log|V(f)|]e^{\frac{j2\pi fn}{F}} \tag{2.15}$$

From the above observations in (2.14), One of the advantages of CEP method is that the logarithm operation compresses the spectral diversity of $|X(f)|$ and easily separates the excitation source information and vocal tract information, respectively. The logarithmic function of CEP method has the effect of shifting the vocal tract characteristics to low-quefrency parts. Utilizing high quefrency parts, CEP can extract the pitch without being affected by the characteristics of vocal tract. Therefore, the CEP based methods clearly express the harmonic structure of the speech signal under clean speech conditions as shown in Fig. 2.10.

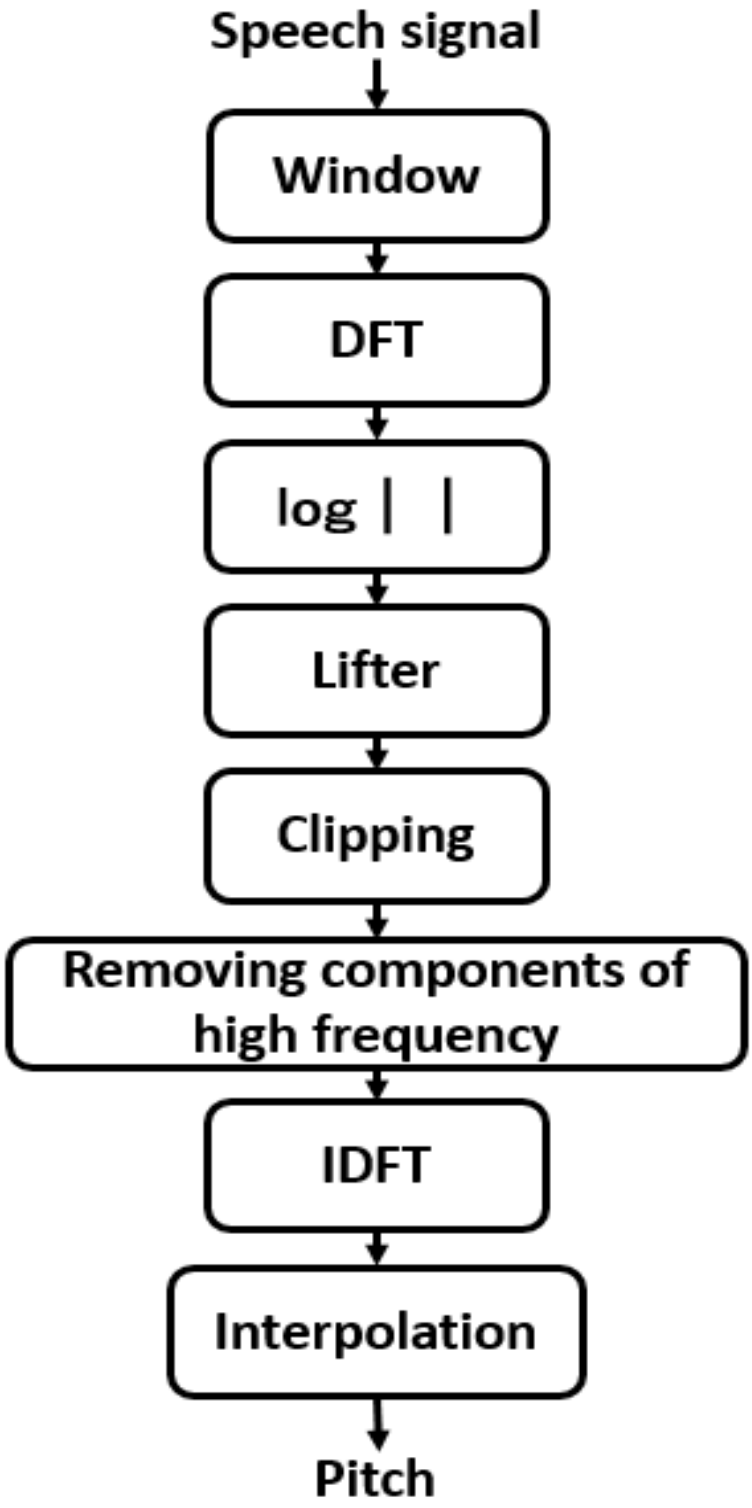

Figure 2.12: Block diagram of MCEP method.

On the other hand, when clean speech is corrupted by noise, the CEP method can be represented as

$$C'(n) = \frac{1}{F}\sum_{f=0}^{F-1} log|X(f) + V'(f)|e^{\frac{j2\pi fn}{F}} \tag{2.16}$$

where $V'(f)$ corresponds to the DFT of additive noise. According to Eq. (2.16), the addition of $log|V'(f)|$ can destroy the periodicity of $log|X(f)|$ at low SNRs. The CEP obtained from the clean speech signal which is shown in Fig. 2.10, where detection of pitch peak is more accurate. On the other hand, after corruption with white noise at 0 [dB] SNR is shown in Fig. 2.11, where the detection of the true pitch peak has failed.

2.4.2 Modified Cepstrum (MCEP) Method

Traditionally, the CEP method is very simple to implement and more efficient in clean speech environment. But, in noisy environment, the log-amplitude spectrum is highly affected by noise. From these observations, the MCEP [32] method utilizes the lifter and clipping operations which is shown in Fig. 2.12. These operations are applied on the log spectrum output in the frequency domain, to enhance the extraction accuracy of pitch. Lifter is a flattening operation which is effective to eliminate the effect of vocal tract characteristics and followed to the clipping operation. After that, clipping operation is applied on liftered spectrum to reduce the effect of noise components. However, the noise components are also present among the harmonics of the clipping output in the high frequency region. Therefore, MCEP method is used the high frequency components on clipped spectrum. After this operation, inverse discrete Fourier transform (IDFT) is applied to the low frequency part and then pitch is extracted by searching a location of the peak in the quefrency domain. The MCEP is comparatively insensitive to vocal tract effects, but it is sensitive to noise characteristics, resulting in extraction errors in highly noisy environment.

2.4.3 Windowless Autocorrelation Function based Cepstrum (WLACF-CEP) Method

The windowless ACF (WLACF) based CEP (WLACF-CEP) [33] utilizes both the properties of WLACF and CEP as shown in Fig. 2.13. The WLACF of the signal

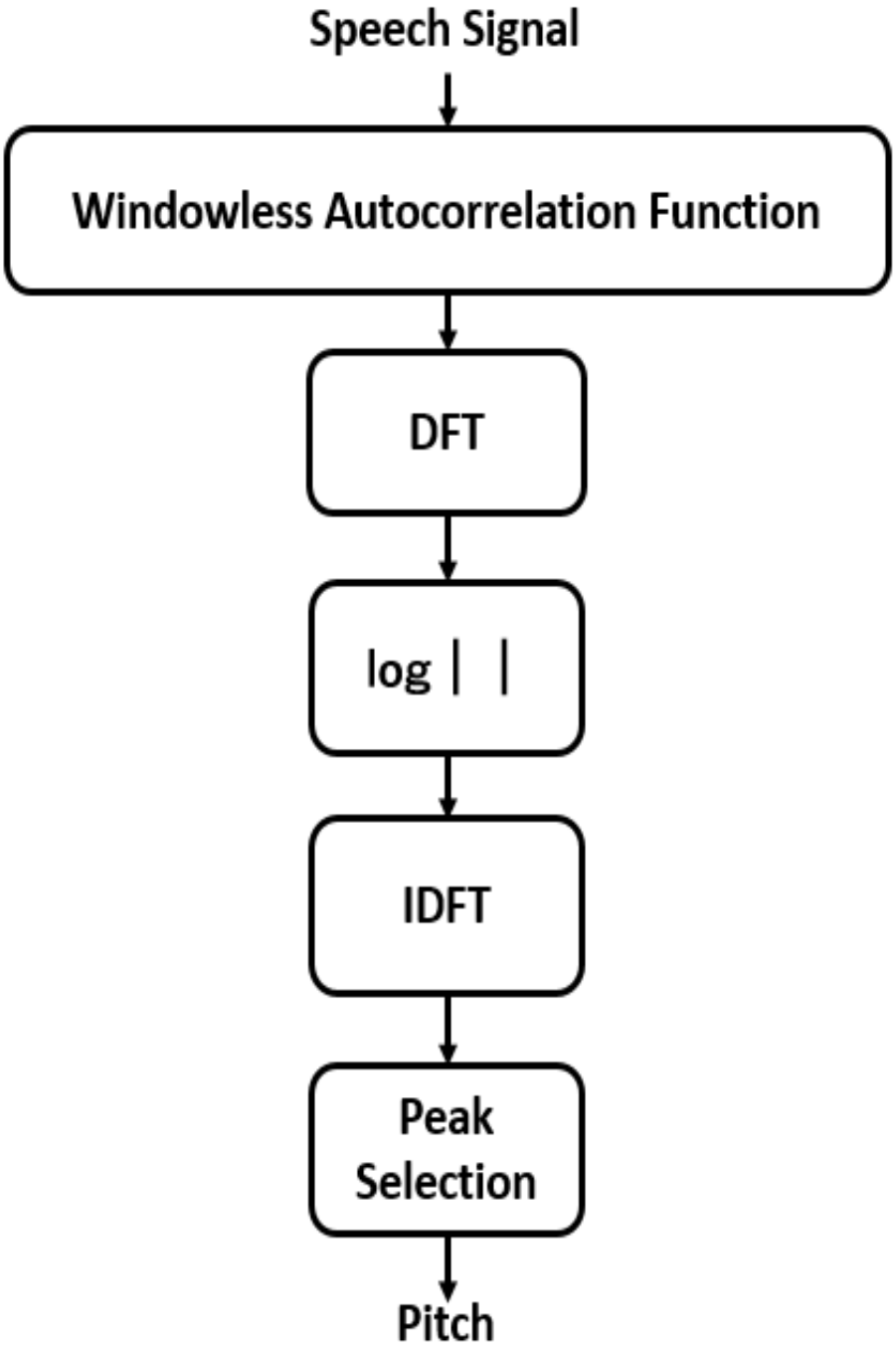

Figure 2.13: Block diagram of WLACF-CEP method.

is a noise compensated equivalent of the signal, while sustaining the periodicity of

the speech signal. The WLACF, $\phi_{yy-wl}(m)$ is defied as

$$\phi_{yy-wl}(m) = \frac{1}{N}\sum_{n=0}^{N-1} y(n)y(n+m) \tag{2.17}$$

for $y(n)$, $n = 0, 1, 2, ..., 2N-1$. In the case of $\phi_{yy-wl}(m)$, $m = 0, 1, 2, ..., N-1$,. In (2.15), $y(n+m)$ is not zero outside N . Therefore, the WLACF shows the strongest in periodicity with accurate pitch peaks. However, the WLACF sometime faces the pitch extraction error due to the effect of the vocal tract characteristics. To combat this problem, a noise free speech signal with periodicity is applied to the CEP to enhance the accuracy of pitch extraction by reducing the effect of vocal tract characteristics. Therefore, the developed WLACF-CEP method, $C_{wlacf}(n)$, as

$$C_{wlacf}(n) = \frac{1}{M}\sum_{f=0}^{M-1} log|\phi_{yy-wl}(f)|e^{\frac{j2\pi fn}{M}} \tag{2.18}$$

The WLACF-CEP method behaves better against various types of noise which is in [33]. But, in noisy environments, the CEP based method does not perform well because the speech peaks are influenced by the noise peaks in the frequency domain.

2.5 Summary

In this chapter, based on time and frequency domains, different methods for extracting pitch have been discussed. In addition the effect of noise on the ACF and AMDF are explained. In noisy environment, these methods are not effective than the WAF and YIN techniques, which are also analyzed detail in time domain. On the other hand, some important methods are discussed in frequency domain. The following Chapters 3 and 4 present two techniques based on the CEP and ACF methods, respectively. The proposed techniques are more effective by utilizing the properties of CEP and ACF which improve the performance of the pitch extraction accuracy in noisy environments.

Chapter 3

Pitch Extraction Using Fourth-Root Spectrum in Noisy Speech

In this chapter, we propose techniques which are based on the CEP method. The CEP method utilizes the log spectrum which maintains the strong periodicity in clean speech. However, in case of noisy speech, the periodicity of log spectrum is corrupted in high frequency domain, and unnecessary peaks arise. From the above point of view, we present the use of the fourth-root spectrum instead of the log spectrum for pitch extraction in noisy environments. To obtain clear harmonics, lifter and clipping operations are performed. When the resulting spectrum is transformed into the time domain by the discrete Fourier transform, pitch detection is robust against narrow-band noise. When the same spectrum is amplified by power calculation and transformed into the time domain, pitch detection becomes robust against wide-band noise. These properties are investigated through exhaustive experiments in various noises. The required computational time is also studied.

3.1 Problem Formulation

The pitch period is defined as the inverse of the fundamental frequency of the excitation source from a voiced speech signal. The pitch period (in short, pitch) or fundamental frequency is a prominent parameter of speech and highly applicable for speech-related systems such as speech analysis-synthesis, speech coding, speech enhancement, and speaker identification systems. The performance of these systems is significantly affected by the accuracy of pitch or fundamental frequency extraction. In this study, we treat pitch and fundamental frequency as having the

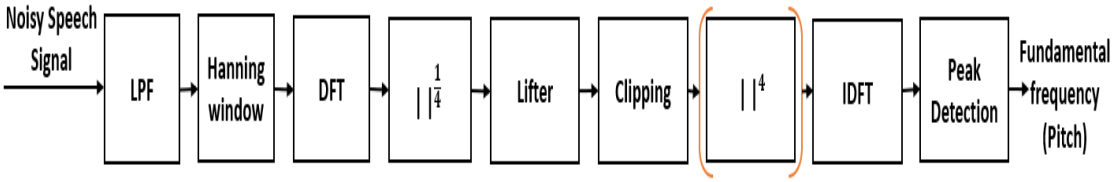

Figure 3.1: Block diagram of FROOT and FROOT+ methods

same meaning, while pitch is inherently interpreted as the perception of fundamental frequency.

Pitch extraction has proven to be a difficult task, even for speech in a noise free environment [18][39]. A clean speech waveform is not really periodic; it is quasi-periodic and highly nonstationary. On the other hand, when a speech signal is corrupted by noise, it is difficult to maintain the reliability and accuracy of pitch extraction algorithms. Under noisy conditions, the periodic structure of the speech signal is destroyed so that pitch extraction becomes an extremely complex task. Among the conventional pitch extraction methods, the autocorrelation function (ACF) [24] is straightforward to compute in the time domain and shows robustness against wide-band random noises such as white noise. The ACF corresponds to a correlation function between the input speech signal and its delayed version in the time domain, but it is also obtained by the inverse Fourier transform of the power spectrum of the speech signal. The ACF is, however, affected by the characteristics of the vocal tract. To reduce the effect of the vocal tract, many algorithms have been developed that rely on the properties of the correlation function [25]-[32]. For example, YIN [27] focuses on the relationship between the conventional ACF and the difference function, and utilizes a cumulative mean function of the difference function to reduce the error rate in pitch extraction. The average magnitude difference function (AMDF) [25] is a simplified version of the ACF, which treats the difference between the speech signal and its delayed version. In [28], the AMDF was combined with linear predictive analysis to eliminate the effect of the vocal tract. Correntropy [40] has similar properties to the ACF and correntropy has a kernel function to transform the original signal into a high-dimensional reproducing kernel Hilbert space (RKHS) in a nonlinear manner. This transformation preserves the characteristics of the periodic signal. Higher-order statistics [41] are also used to enhance the resolution of pitch extraction. However, the performance of correntropy and higher order-statistics in noisy environments is unsatisfactory. In [42], harmonic sinusoidal autocorrelation (HSAC) was proposed. The symmet-

ric average magnitude sum function (SAMSF) was utilized to generate a periodic impulse train to extract the pitch. The resulting pitch extractor based on least squares and optimum value finding (searching) is too complex to implement because it requires post-processing. In [43], dominant harmonic reshaping from the normalized autocorrelation function (NACF) [44] of noisy speech was performed and the empirical mode decomposition (EMD) of the resulting NACF waveform was implemented where an iterative operation could not be avoided. The method in [43] is also complicated and results in a long computation time. In [45], the auditory filterbank decomposed the speech signal into subbands. Then, the NACF was applied to the subband signals, which were encoded to extract the pitch. The NACF reduces the variations in signal amplitude more than the ACF does. The approach in [45] is very effective, but it inherently relies on a sophisticated post-processing technique to compensate for the pitch extraction errors.

In highly noisy environments, the two correlation-based methods, ACF and AMDF, are inferior to the weighted autocorrelation function (WAF) [26]. The WAF also focuses on the ACF, but it is weighted by the inverse of the AMDF, resulting in an excellent pitch extractor in noisy environments. Most of the ACF-based pitch extraction methods are effective in white noise. However, the pitch extraction performance of the ACF-based methods is degraded when clean speech is corrupted by color noise.

In the frequency domain, one of the most widely used techniques employs the cepstrum (CEP), which was originally proposed in [30] and improved in [31]. In the CEP method, the pitch is extracted by applying the inverse Fourier transform to the log-amplitude spectrum, which is also effective. The logarithmic function involved in the CEP shifts the vocal tract characteristics to low-quefrency parts. Utilizing high-quefrency parts, we can extract the pitch without being affected by the characteristics of the vocal tract. The modified CEP (MCEP) in [32] further involves the liftering and clipping operations on the log spectrum, which is used to remove the characteristics of the vocal tract as well as to eliminate the unnecessary notches of spectral valleys that correspond to noise in the log spectrum. The MCEP also removes the high-frequency components to increase the pitch extraction accuracy. The ACF of the log spectrum (ACLOS) [38] also utilizes the liftering and clipping operations on the log spectrum. Then, the ACF is applied to the resulting log spectrum. The ACLOS emphasizes the periodicity of harmonics in the spectrum.

The CEP-based methods clearly express the harmonic structure of the speech

signal under no-noise conditions. However, in noisy environments, the CEP-based methods do not always perform well because the speech peaks are affected by the noise peaks in the frequency domain. A spectral harmonic technique was proposed in [36]. In this method, a bank of bandpass lifters is used to flatten the spectrum. The ACF is applied in the spectrum domain to extract the pitch periodicity by reducing the effect of vocal tract characteristics. This approach may be effective but the overall procedure is too complex to implement.

Recently, two sophisticated approaches have been proposed [46][47]. The pitch estimation filter with amplitude compression (PEFAC) [46] is a frequency domain pitch extraction method, which utilizes sub-harmonic summation [48] in the log frequency domain. The PEFAC also includes an amplitude compression technique to enhance its noise robustness. On the other hand, BaNa [47] considers noisy speech peaks and provides a hybrid pitch extraction method that selects the first five spectral peaks in the amplitude spectrum of the speech signal. BaNa calculates the ratios of the frequencies of the spectral peaks with tolerance ranges and accurately extracts the pitch of the speech signal.

Although deep neural network (DNN)-based approaches exist [49][50] as a recent approach to pitch extraction, they typically require a tremendously long time for learning owing to the huge data size.

In this chapter, we propose the use of the fourth-root (FROOT) spectrum of noisy speech for pitch extraction. Motivated by the fact that the MCEP method is very simple to implement but provides an excellent pitch extraction performance in noisy environments, the MCEP method is improved. In the proposed method, which is referred to as the FROOT method, the fourth-root spectrum is used instead of the log spectrum in the MCEP method. The idea of the FROOT method has been reported in a conference [51], where a preliminary experiment was conducted and only limited results for narrow-band noise were shown. In this paper, we further extend the FROOT method for wide-band noise and investigate the performance of both the FROOT and extended FROOT methods in various noises. In wide-band noise, the noise energy is distributed over a wide range of frequencies. In this case, the FROOT method is corrupted in the high-frequency domain by the noise characteristics. However, the extended FROOT method additionally utilizes the fourth-power calculation (fourth-power spectrum) to present clear harmonics and emphasizes the pitch peak in the frequency domain, simultaneously suppressing the noise components included in noisy speech. In this paper, the extended FROOT method is referred to as the FROOT+ method.

The remainder of this chapter is organized as follows. Section 3.2 describes the principle of the FROOT and FROOT+ methods. In Sec. 3.3, we first show preliminary experiments. After that, we compare the FROOT and FROOT+ methods with conventional methods through experimental results and then discuss the performance and processing time for each method. Finally, we conclude this chapter in Sec. 3.4.

3.2 FROOT and FROOT+ Methods

Let us assume that the clean speech signal $x(n)$ is corrupted by noise, $v(n)$. The noisy speech signal $y(n)$ is expressed as

$$y(n) = x(n) + v(n) \tag{3.1}$$

Figure 3.1 shows a block diagram of the FROOT and FROOT+ methods. When the fourth-power calculation in parentheses is included, Figure 1 corresponds to the FROOT+ method. When this part is not included, it corresponds to the FROOT method. In the FROOT and FROOT+ methods, firstly we apply a low-pass filter (LPF) to the noisy speech signal because the LPF can eliminate the noise characteristics to increase the accuracy of pitch extraction. The LPF is often applied before the analysis of speech signals and filters out the high-frequency components of the noisy speech signal. We use an LPF with the telephone line cut-off frequency.

After windowing, we calculate the fourth-root spectrum. Here, we considered different spectral shapes of a speech signal as shown in Fig. 3.2. From Fig. 3.2, we can observe that the periodicity of the log spectrum is destroyed by the noise. On the other hand, the fourth-root spectrum emphasizes the pitch harmonics in the low-frequency region as well as reduces the noise effect. For this reason, the fourth-root spectrum is used in the FROOT and FROOT+ methods.

However, the fourth-root spectrum is sometimes affected by vocal tract characteristics. To overcome this problem, the operation of flattening is effective. Therefore, we apply a lifter to the fourth-root spectrum by multiplying a filter in the quefrency domain and then converting the liftering result back to the frequency domain. Basically, the vocal tract information is present at the lower part in the quefrency domain. At the higher part in the quefrency domain, the pitch information is present. Therefore, we apply a high-pass lifter (HPL) to eliminate the effect

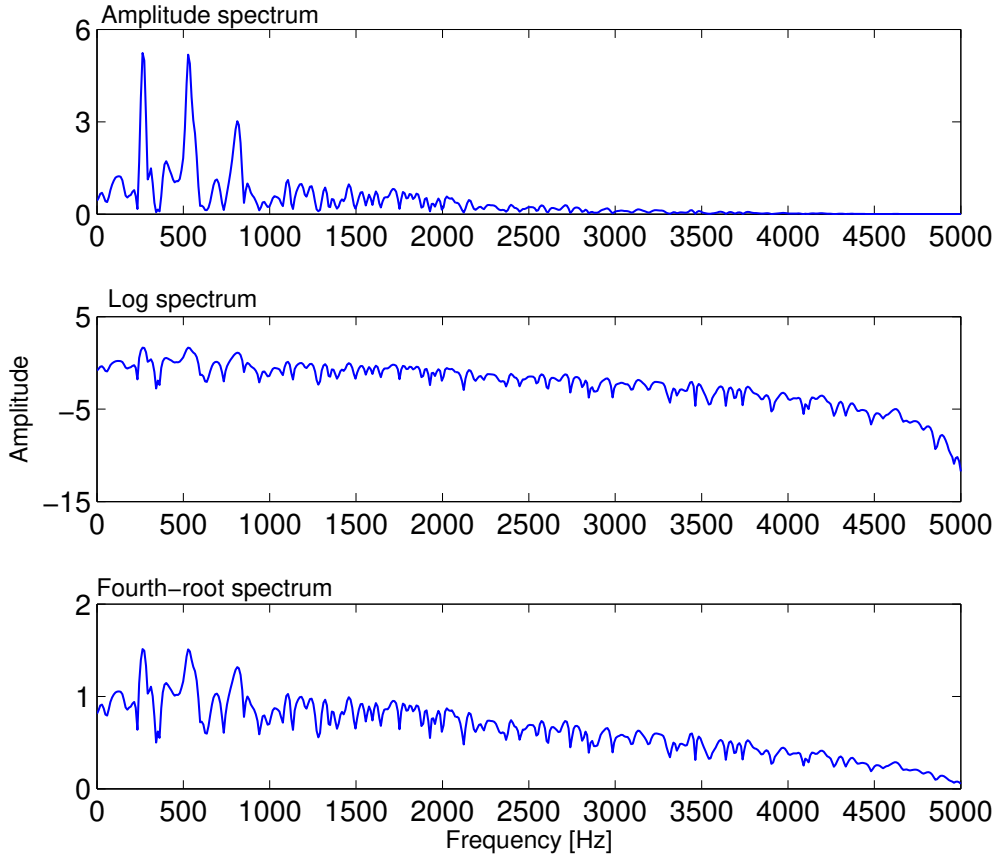

Figure 3.2: Different spectral shapes of speech signal at SNR=0 [dB] (white noise)

of the vocal tract information and simultaneously eliminate the noise components contained at the lower part in the quefrency domain. The cutoff quefrency level of the HPL should be small to reduce the effect of the vocal tract characteristics. Experimentally, we found that the cutoff quefrency level of 2.5 [ms] (25 samples for the sampling rate of the NTT database) for the HPL preserves the high periodicity more reliably than that with a higher cutoff quefrency level at the lifter output. Some examples are shown in Fig. 3.3. Therefore, when the FROOT and FROOT+ methods were used in the experiments in Sec. 3.3, the cutoff quefrency level of 2.5 [ms] for the HPL was used. However, after the lifter operation, we observed that noise components are present between the harmonics. Therefore, a clipping operation is also applied to the lifter output, which reduces the effect of the noise using an accurate clipping threshold level. The selection of the clipping threshold level is described in Sec. 3.3.

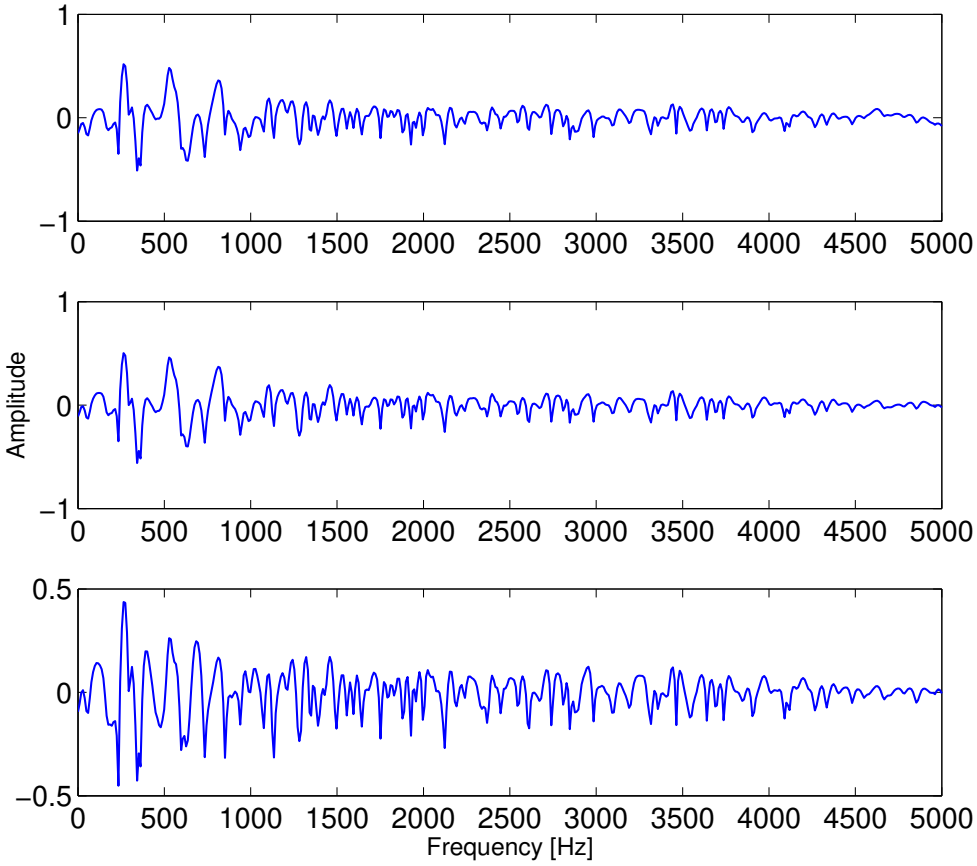

Figure 3.3: Waveforms of different liftering outputs using cutoff quefrency levels of (a) 1 [ms] (b) 2.5 [ms] and (c) 4 [ms] at SNR=0 [dB] (white noise) in NTT database

After the above process, in the FROOT+ method, a power calculation is performed (in the FROOT method, this part is omitted). Figure 3.4 shows an example of which power factor is suitable for the clipping output to reduce the noise components in the FROOT+ method. In Fig. 3.4, we observe that the noise components are reduced by increasing the power factor. However, as the power factor increases, the effect of the formant characteristics of the vocal tract sometimes also increases. Therefore, undesired peaks arise. From Fig. 3.4, we selected four as the power factor value for the FROOT+ method, which is the most effective value for reducing the noise. This is the reason why the fourth-power calculation is drawn in Fig. 3.1.

After this process, for both the FROOT and FROOT+ methods, the inverse discrete Fourier transform (IDFT) is applied and the resulting spectrum is trans-

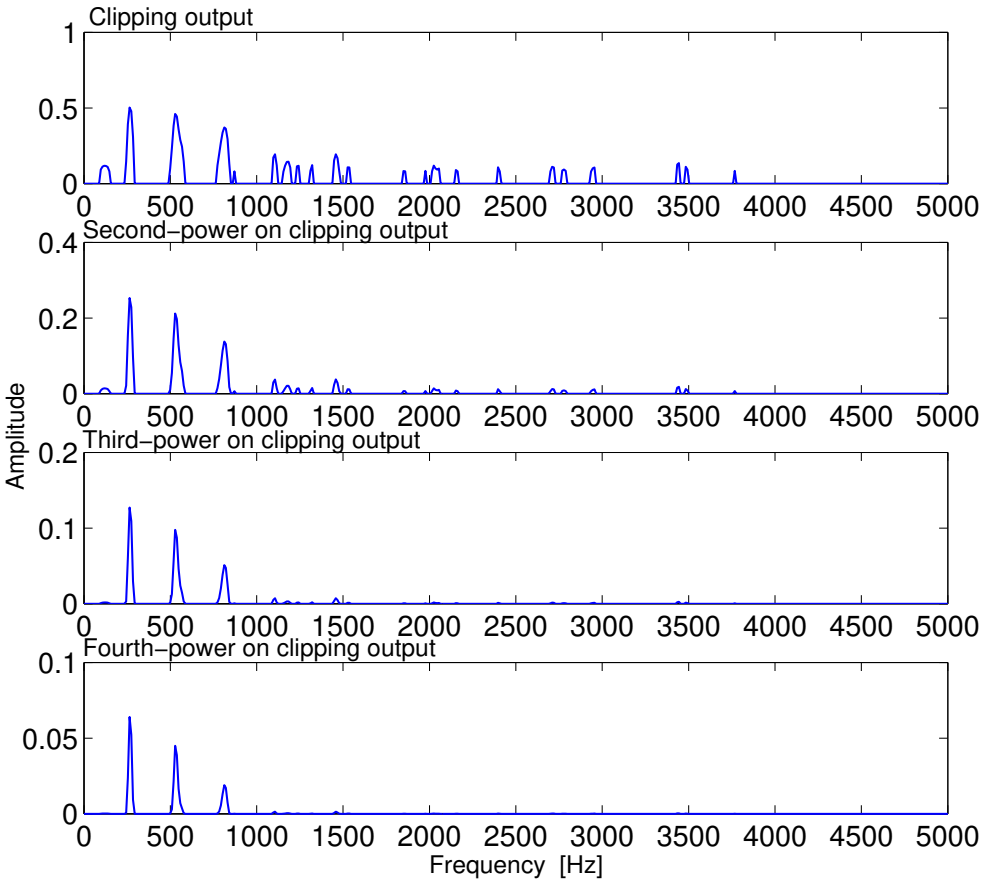

Figure 3.4: Relation between the clipping output and the power factor for female speech at SNR=0 [dB] (white noise)

formed into the time domain, where a peak corresponding to the pitch peak is detected.

Figure 3.5 illustrates how to extract the pitch period by using the FROOT and FROOT+ methods in narrow-band noise (car interior noise) and in wide-band noise (white noise). In the narrow-band noise (Fig. 3.5(a)), we observe that the energy level of the first three peaks provides almost the same amplitude in the low-frequency region of the fourth-root spectrum. The pitch information exists in this region, but some peaks are undesired. When the fourth-power calculation is applied to the clipping output, the undesired peaks are enhanced. This leads to the FROOT+ method producing a pitch detection error. However, the FROOT method gives correct pitch detection without undesired peaks. In contrast, in the wide-band noise (Fig. 3.5(b)), the harmonic peaks maintain their periodicity in the low-frequency region at the clipping output. When the fourth-power calculation is

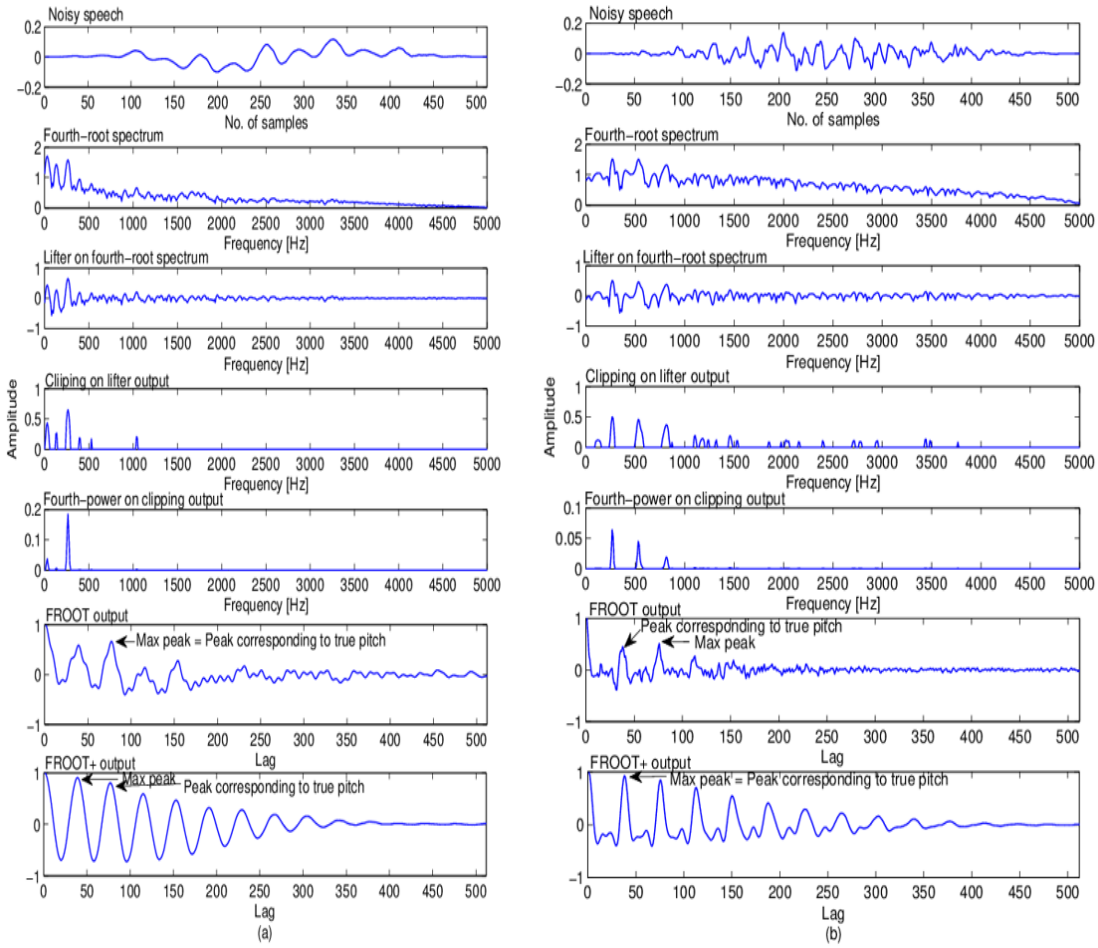

Figure 3.5: Processing in step-by-step for FROOT and FROOT+ methods, (a) at SNR=0 [dB] (car interior noise) (b) at SNR=0 [dB] (white noise)

applied to the clipped spectrum, the noise effect is suppressed. Otherwise, the noise components remain in a wide frequency range. Therefore, the FROOT+ method accurately detects the pitch peak. The FROOT method leads to a detection error in this case.

3.3 Experiments

We conducted experiments on speech signals.

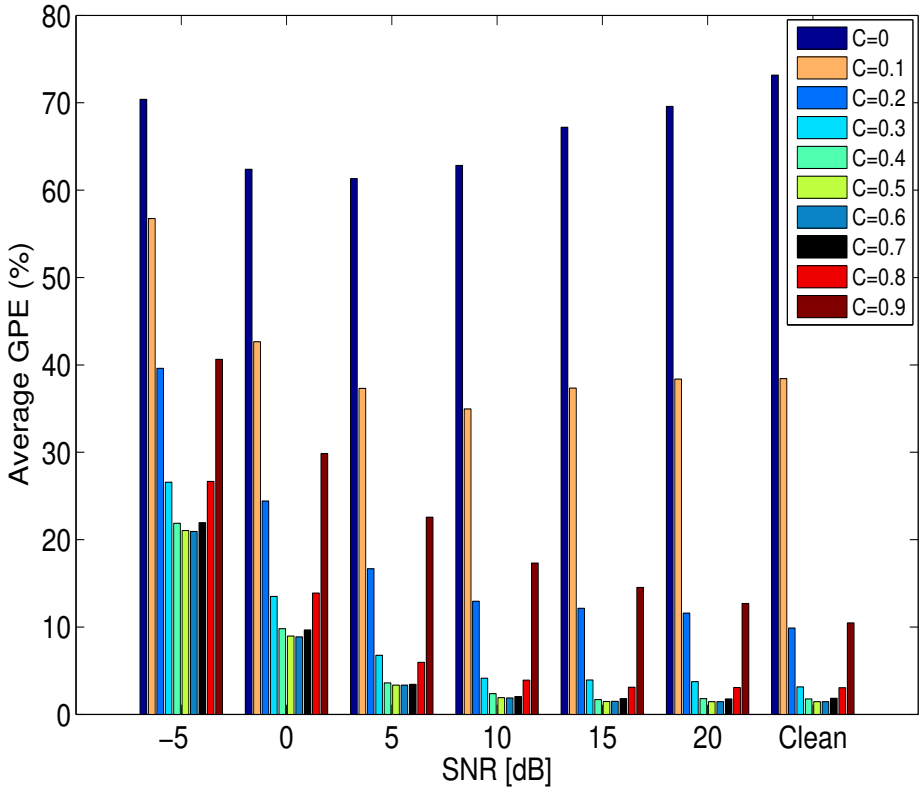

Figure 3.6: Relation between clipping constant level (C) and GPE at different SNRs (male speakers)

3.3.1 Experimental conditions

Speech signals were taken from two databases: NTT [52] and KEELE [53]. In the NTT database, which was developed by NTT Advanced Technology Corporation, the speech materials are 11 [s] long and are spoken by four male and four female Japanese speakers for each sentence; the speech signals were sampled at a rate of 10 [kHz]. From the KEELE database, we utilize five male and five female English speech signals. The total length of the ten speakers' speeches is about 6 [m]. The speech signals were sampled at a rate of 16 [kHz]. To generate noisy speech signals, we added different types of noise to the speech signals in both databases. White noise with zero mean and unit variance was generated by a computer and added to the speech signals with amplitude adjustment. Pink, babble, factory, HF channel, car interior, and military vehicle noises were taken from the NOISEX-92

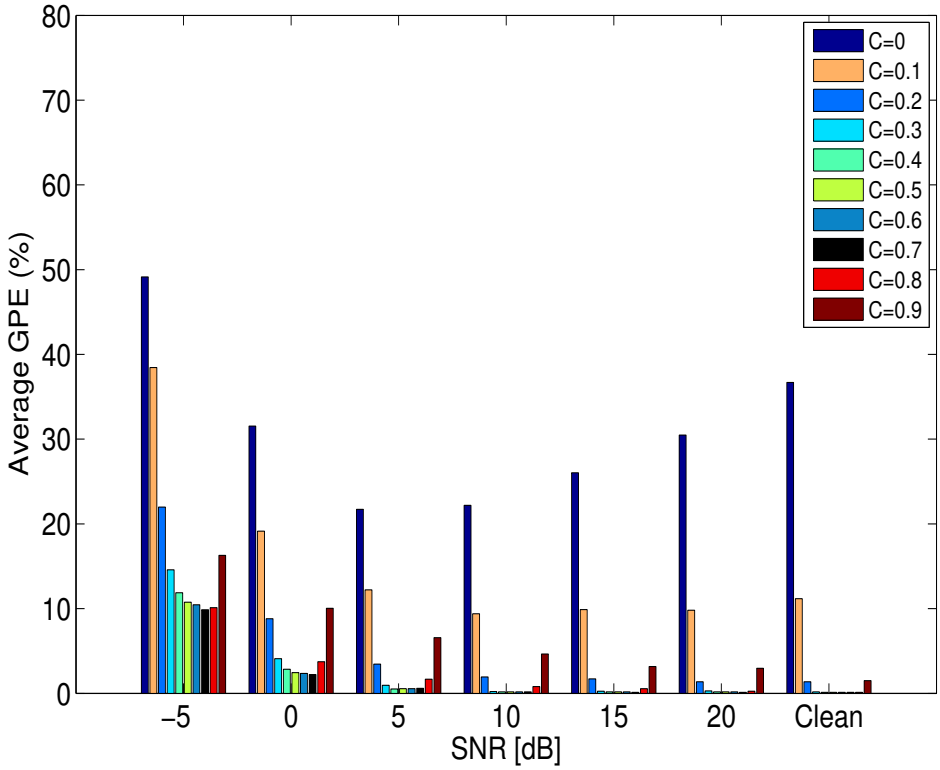

Figure 3.7: Relation between clipping constant level (C) and GPE at different SNRs (female speakers)

database [54] with a sampling frequency of 20 [kHz], and train noise was taken from the Japanese Electronic Industry Development Association (JEIDA) noise database [55] with a sampling frequency of 8 [kHz]. These noises were resampled with sampling frequencies of 10 [kHz] and 16 [kHz], respectively, when they were added to the speech data in the NTT and KEELE databases. The SNR was set to -5, 0, 5, 10, 20, and ∞ [dB], and the other experimental conditions for pitch extraction were

- frame length: 51.2 [ms], except for BaNa;
- frame shift: 10.0 [ms];
- window function: Hanning;
- band limitation of LPF: 3.4 [kHz];

- DFT (IDFT) length: 1024 points for the NTT database and 2048 points for the KEELE database except for BaNa.

The following pitch extraction error $e(l)$ based on Rabiner's rule [18] was used for the evaluation of pitch extraction accuracy:

$$e(l) = F_{est}(l) - F_{true}(l) \tag{3.2}$$

where l is the frame number and $F_{est}(l)$ and $F_{true}(l)$ are the fundamental frequency extracted from the noisy speech signal and the ground truth fundamental frequency at the lth frame, respectively. If $|e(l)| > 10[\%]$ from the ground truth fundamental frequency, we classified the error as a gross pitch error (GPE) and calculated the GPE rate (as a percentage) over all the voiced frames included in the speech data. Otherwise, we classified the error as a fine pitch error (FPE) and calculated the mean value of the absolute errors. We detected and assessed only voiced parts in sentences for pitch extraction. To extract the pitch, we used the search range of $f_{max} = 50$ [Hz] and $f_{min} = 400$ [Hz], which corresponds to the fundamental frequency range of most people.

The ground truth information for the fundamental frequency at each frame is included in the KEELE database, while the true fundamental frequencies at each frame in the NTT database were measured in [38] by observing the speech waveforms carefully, which are used here. Therefore, the $F_{true}(l)$ values in (3.2) are known a priori in the evaluation.

3.3.2 Preliminary Experiments

For the FROOT and FROOT+ methods, it is important to set a constant parameter for the clipping threshold level, η , which is expressed as

$$\eta = \alpha_{min} + \mathrm{C}(\alpha_{\max} - \alpha_{\min}) \tag{3.3}$$

where α_{min} and α_{max} are respectively the minimum and maximum values of the fourth-root spectrum after the lifter operation, and C denotes a constant parameter. We conducted preliminary experiments to determine the optimal value of the clipping threshold level. For this purpose, we used the NTT database, because the size of its speech data is smaller than that of the KEELE database. We employed male and female speech signals corrupted by white noise. By adjusting the amount of noise to be added, a range of SNR of -5 [dB] to 20 [dB] was investi-

gated. Additionally, clean speech was also investigated. Figures 3.6 and 3.7 show the relationship between the clipping threshold level and average GPE rate of the FROOT+ method for four male and four female speakers, respectively. Here, we changed the clipping threshold level from 0 to 0.9. In Figs. 3.6 and 3.7, we observe that setting C = 0.6 - 0.7 for male and female speakers gives low GPE rates at almost all SNR levels.

In accordance with the results in Figs. 3.6 and 3.7, we select the constant parameter C = 0.6 commonly for both male and female speech signals to ensure a high extraction accuracy in the FROOT and FROOT+ methods.

3.3.3 Performance Comparison

The pitch extraction performance of the conventional methods (YIN [27] and BaNa [47]) and the FROOT and FROOT+ methods was investigated in noisy environments. In [56], BaNa was assessed as the best pitch extractor in noisy environments among nine methods that were compared. YIN was the second-best method in [49], where the best one was DNN-based. In this chapter, we consider eight types of noise, which are classified into two categories depending on their characteristics: wide-band noise and narrow-band noise. White, pink, babble, train, factory, and HF channel noises correspond to wide-band noise. Car interior and military vehicle noises correspond to narrow-band noise. The noise characteristics are discussed in detail in Sec. 3.3.4. For the FROOT and FROOT+ methods, we commonly used a cutoff quefrency level of 2.5 [ms] for the HPL. All parameters of the conventional methods were the same as those of the FROOT and FROOT+ methods, except for the frame length and the number of DFT(IDFT) points for BaNa. Specifically, for BaNa, the frame length was set as 60 [ms] and the number of DFT (IDFT) points was 2^{16} in accordance with [47] (this is the best setting for BaNa). The source code used to implement BaNa was taken from [57]. We implemented the YIN method on the basis of the algorithm described in [27]. In particular, for the YIN method, to confirm the validity of our code, we used the same parameter settings and GPE evaluation criteria as those in [49], and confirmed that the performance of our implemented YIN method provides a similar average GPE rate to the YIN method in [49] for white and babble noises in the KEELE database.

For pitch extraction, we cannot ignore the fact that the extraction performance is largely dependent on the speaker's characteristics, especially for low or high pitches [18][38][39], which are typical characteristics of male and female speech,

respectively. Additionally, different natures of additive noise such as wide-band or narrow-band, flat-spectral or not flat-spectral, and time-invariant or time-variant produce different results for pitch extraction [46][47][49][50]. This is due to the nonuniform phenomena invoked in a complex combination of speech harmonics, formant characteristics and the noise shape created in a framed voiced speech. Therefore, it is important to investigate the pitch extraction performance separately on male and female speech and separately on each noise type. For this reason, we precisely show the result for each case and discuss it later.

(A) NTT database case

Figures 3.8 and 3.9 show the average GPE rates of the four male and four female speech signals in the NTT database, respectively, with different noises. Each plot was obtained under each SNR level from -5 [dB] to ∞ [dB] (clean speech).

From Fig. 3.8, it is observed that in the case of wide-band noise, the average GPE rate of the FROOT+ method is lower than those of the other methods for the white, train, and HF channel noises at low SNRs. At high SNRs (>10 [dB]), the FROOT+ and FROOT methods have similar performance characteristics. At low SNRs of pink and factory noises, BaNa provides a lower error rate than the other methods. At high SNRs (>5 [dB]) of pink and factory noises, the FROOT+ and FROOT methods have similar performance characteristics but provide lower GPE rates than BaNa. In the babble noise case, the FROOT+ method has a lower GPE rate than the YIN method and BaNa, and competitive performance with the FROOT method. On the other hand, in the case of narrow-band noise, the FROOT method provides a lower GPE rate than the other methods at almost all SNR levels except for BaNa at low SNRs (<5 [dB]) of car interior noise.

From Fig. 3.9, the FROOT+ method has significantly better performance than the FROOT method in the wide-band noise case. However, BaNa has a lower GPE rate than the other methods in the wide-band noises except for pink noise. BaNa is still also better in the narrow-band noises, although the FROOT method has better performance than BaNa at low SNRs of car interior noise. In the pink noise case, the FROOT+ method has better performance than the conventional and FROOT methods at all SNRs.

Figures 3.10 and 3.11 show the average FPE for male and female speech data, respectively, in the NTT database. The FPE represents the degree of variation in detecting the pitch. The average FPE for all methods ranges approximately from 0.8 [Hz] to 6.2 [Hz]. In Fig. 3.10, we observe that the FPE of the FROOT+ method is better than those of most of the other methods but not the best. The

YIN method has excellent performance at low SNRs (<15 [dB]) in the case of wide-band noise and the FROOT method has the best performance at high SNRs (>15 [dB]). In the narrow-band noise case, the FROOT method is the best, and the FROOT+ method is typically the second best. On the other hand, in Fig. 3.11, we observe that BaNa performs better than the other methods at low SNRs (<5 [dB]) in wide-band noise. At high SNRs, the YIN method has the best performance and the FROOT and FROOT+ methods, and BaNa have similar performance characteristics. In the narrow-band noise case, the FROOT method is the best.

(B) KEELE database case

To validate the performance of the FROOT and FROOT+ methods in a more reliable manner, we also employed the KEELE database. Figures 3.12 and 3.13 show the average GPE rates for male and female speakers, respectively. The KEELE database provides the ground truth values of the fundamental frequency, which are obtained from laryngograph signals. We analyzed them and found that some discontinuities are present. Therefore, the ground truth values are not particularly accurate. This is reflected in the resulting GPE rates. In Figs. 3.12 and 3.13, the GPE rates of the clean speech are clearly higher than those of the clean speech in Figs. 3.8 and 3.9. This is due to the lower accuracy of the ground truth values in the KEELE database.

Figure 3.12 indicates a tendency similar to that in Fig. 3.8 for all methods. Figure 3.13 is also similar to Fig. 3.9 from a performance comparison aspect, although BaNa has comparatively low performance in the babble and car interior noise cases.

The average FPE performance characteristics for male and female speech data are shown in Figs. 3.14 and 3.15, respectively. Figure 3.14 is also similar to Fig. 3.10, but the FROOT and FROOT+ methods behave similarly, giving the best performance in almost all cases. However, Fig. 3.15 indicates a different tendency from Fig. 3.11. In particular, the performance of BaNa deteriorates and BaNa has the worst performance in all cases. However, the relationship between the performance characteristics of the FROOT, FROOT+ and YIN methods in Fig. 3.15 is similar to that in Fig. 3.11.

(C) Summary

Through the results in Figs. 3.8-3.15, we observe that the performance of each method has a similar tendency for both speech databases. To summarize, in the wide-band noise case, the FROOT+ method provides a low GPE rate in various types of noise over a wide range of SNRs, although BaNa is advantageous for female

speech. Regarding the FPE performance, the FROOT method is superior to BaNa. In the narrow-band noise case, the FROOT method has excellent performance in terms of GPE and FPE.

In the case of windowing effect for the FROOT and FROOT+ methods, Fig. 3.16 shows the average GPE rate on four male and four female speech signals in the NTT database with different noises. When the SNR is changed from -5 [dB] to infinity [dB] (clean speech case), each plot has been obtained under each SNR condition. Here, we have used the white, pink, babble, and car interior noises. In Fig. 3.16, only the FROOT and FROOT+ methods are compared. The FROOT and FROOT+ methods are used Hanning window function. On the other hand, FROOT-REC and FROOT+-REC methods are used the Rectangular window function. From the experimental results in Fig. 3.16, the usefulness of the Rectangular window in noisy environments has been found. When the FROOT and FROOT+ methods are used with the Rectangular window, it provides better GPE rates than the Hanning window based FROOT and FROOT+ methods especially at low SNRs (≤ 10 [dB]) at almost all noise cases except for low SNRs in car interior noise. At low SNRs in car interior noise, the Rectangular window based FROOT (FROOT-REC) method is competitive with Hanning window based FROOT (FROOT) method.

3.3.4 Discussion

We next discuss the performance of each method. Figure 3.17 shows long-term spectra of the different noises we employed. The spectra of the narrow-band noises (car interior and military vehicle noises) have the greatest amplitude in the frequency range of less than 200 [Hz], producing narrow-band peaks. On the other hand, the spectra of the wide-band noises (white, pink, babble, train, factory, and HF channel noises) are comparatively evenly spread.

The YIN [27] method is an ACF-based method. For such a method, pitch extraction is robust against wide-band random noises such as white noise but weak against narrow-band noises such as periodic noise. This is consistent with the results in Figs. 3.8-3.15. Car interior and military vehicle noises create sharp peaks in the low-frequency region as shown in Fig. 3.17. These peaks produce clear periodicity in the noise waveform, resulting in the degradation of the GPE rate. HF channel noise has wide-band characteristics. However, a typical peak exists at around 2600 [Hz]. When SNR is high, the peak is negligible. However,

when SNR becomes lower, the peak increases in magnitude and is expected to produce periodicity in the noise waveform. This is considered to be the reason why the GPE rate of the YIN method is often severely degraded at low SNRs in the HF channel noise case as shown in Figs. 3.8, 3.9, 3.12, and 3.13. In BaNa, the pitch of speech is found from some candidates and post-processing is also incorporated to accomplish accurate pitch extraction. BaNa is capable of overcoming the movement of distorted peaks in noisy cases by estimating the pitch by calculating the harmonic number with a permitted margin. Female speech consists of fewer harmonics in the first formant range and the energy of the voice speech is concentrated at these harmonics; thus, female speech is less affected by noise. In this case, BaNa is advantageous, as shown in Figs. 3.9 and 3.13 regardless of the noise type. In contrast, in the male speech, the speech energy is spread over many harmonics and is highly affected by noise. In this case, the performance of BaNa degrades, since the choice of more spectral harmonic peaks must be considered. Actually, the performance of BaNa is comparatively low as shown in Figs. 3.8 and 3.12, but it is still excellent at low SNRs by relying on the post-processing algorithm. However, a huge number of points of FFT is required to find each harmonic peak accurately in BaNa. The computation of several candidate pitches and that of the post-processing including the Viterbi algorithm are complex, resulting in a long computation time as shown in Sec. 3.3.5. On the other hand, for the FROOT and FROOT+ methods, the fourth-root spectrum makes the periodicity in the harmonic structure clear. In this process, the spectral peak of narrow-band noise is suppressed. Since this effect is combined with the following lifter and clipping operations, the FROOT method is robust against car interior and military vehicle noises as shown in Figs. 3.8, 3.9, 3.12 and 3.13. However, the FROOT+ method uses the fourth-power calculation after the clipping operation. In this process, the remaining spectral peak of narrow-band noise is enhanced again. Therefore, the performance of the FROOT+ method is lower than that of the FROOT method. However, in the wide-band noise that we employed, unnecessary peaks typically arise in the high-frequency region as noise components. These are suppressed by the fourth-power calculation as shown in Fig. 3.4. Therefore, the fourth-power calculation can be effectively applied to the clipped spectrum to reduce the noise effect in the case of wide-band noise.

3.3.5 Processing Time

In Table 3.1, we show the processing time per second of data for each method in the NTT database. We tested all methods on a PC with an Intel (R) Core(TM) i5-6400K CPU with 4 [GHz] clock speed and 8 [GB] of memory. For evaluation, we used five trials for each method then calculated the average processing time required to obtain reliable measurements. The computational time of the YIN method is reasonable because it uses the squared difference function to identify the pitch. BaNa has the longest processing time because of the large FFT size used to achieve a high frequency resolution. The processing times of the FROOT+ and FROOT methods are similar and shorter than those of the other methods, since the clipping and liftering operations are directly applied to the fourth-root spectrum.

Table 3.1: Processing time per second of speech

YIN	BaNa	FROOT	FROOT+
0.641	29.427	0.146	0.157

3.4 Summary

In this chapter, we proposed the use of the fourth-root spectrum to deal with the problem of pitch extraction from noise-corrupted speech signals. The FROOT and FROOT+ methods were derived from the liftered and clipped version of the fourth-root spectrum. The FROOT method can be switched to the FROOT+ method in a simple manner by embedding the fourth-power calculation after the liftered and clipped spectrum calculation. The FROOT+ method reduces the effect of vocal tract characteristics as well as suppresses the non-pitch peaks in the frequency domain, enhancing the pitch peak in the wide-band noise. On the other hand, the FROOT method behaves similarly to the FROOT+ method but results in a pitch extractor that is strongly robust against narrow-band noise. Through experiments, we confirmed that both methods are efficient and effective for extracting the pitch in a wide range of noise types when selected in accordance with the noise characteristics such as wide-band and narrow-band noises. Also, to improve the accuracy of the FROOT and FROOT+ methods in noisy environments, the use

of the Rectangular window is more effective instead of the Hanning or Hamming window.

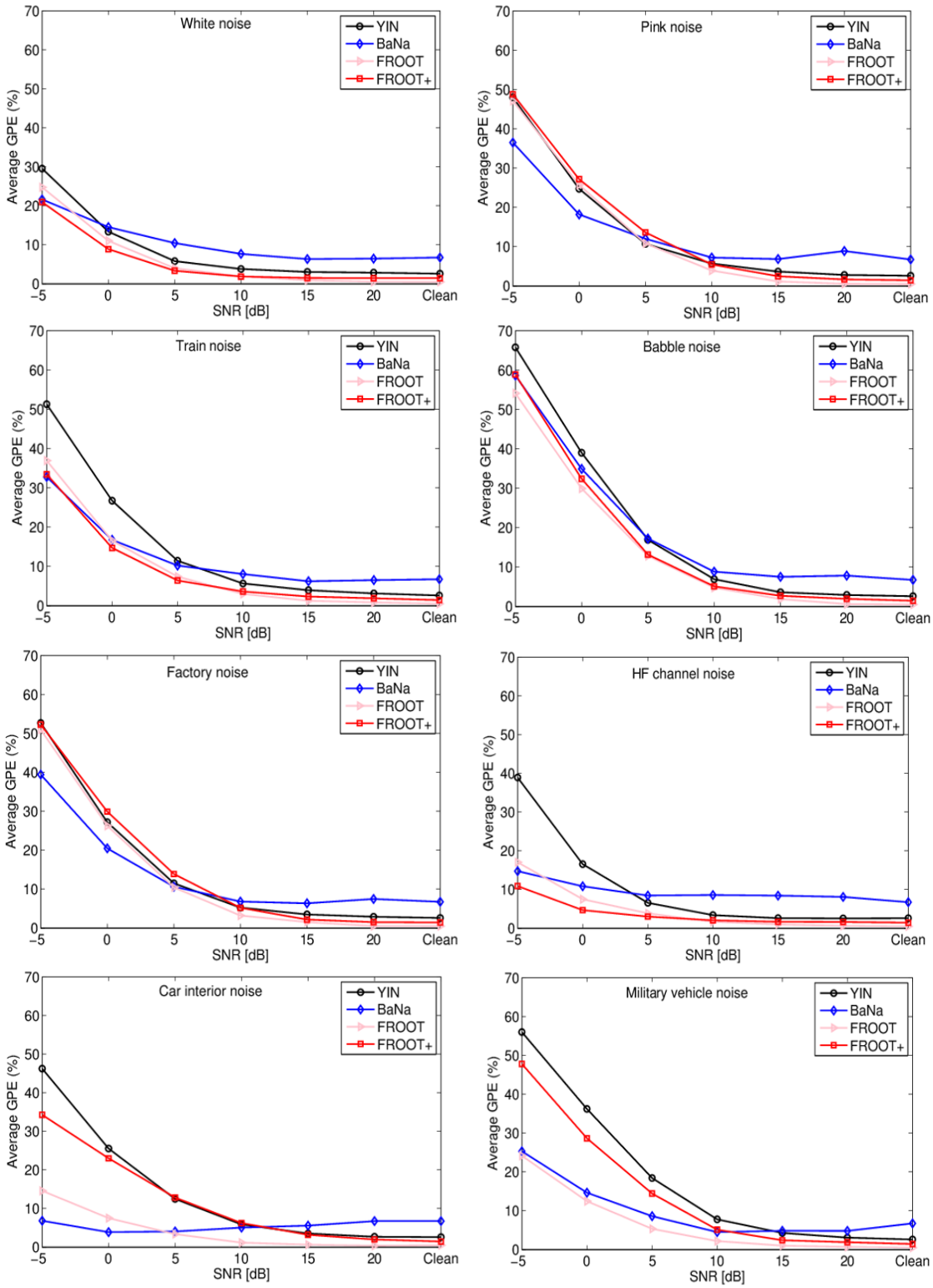

Figure 3.8: GPE for four male speakers with different types of noise under different SNR levels in NTT database

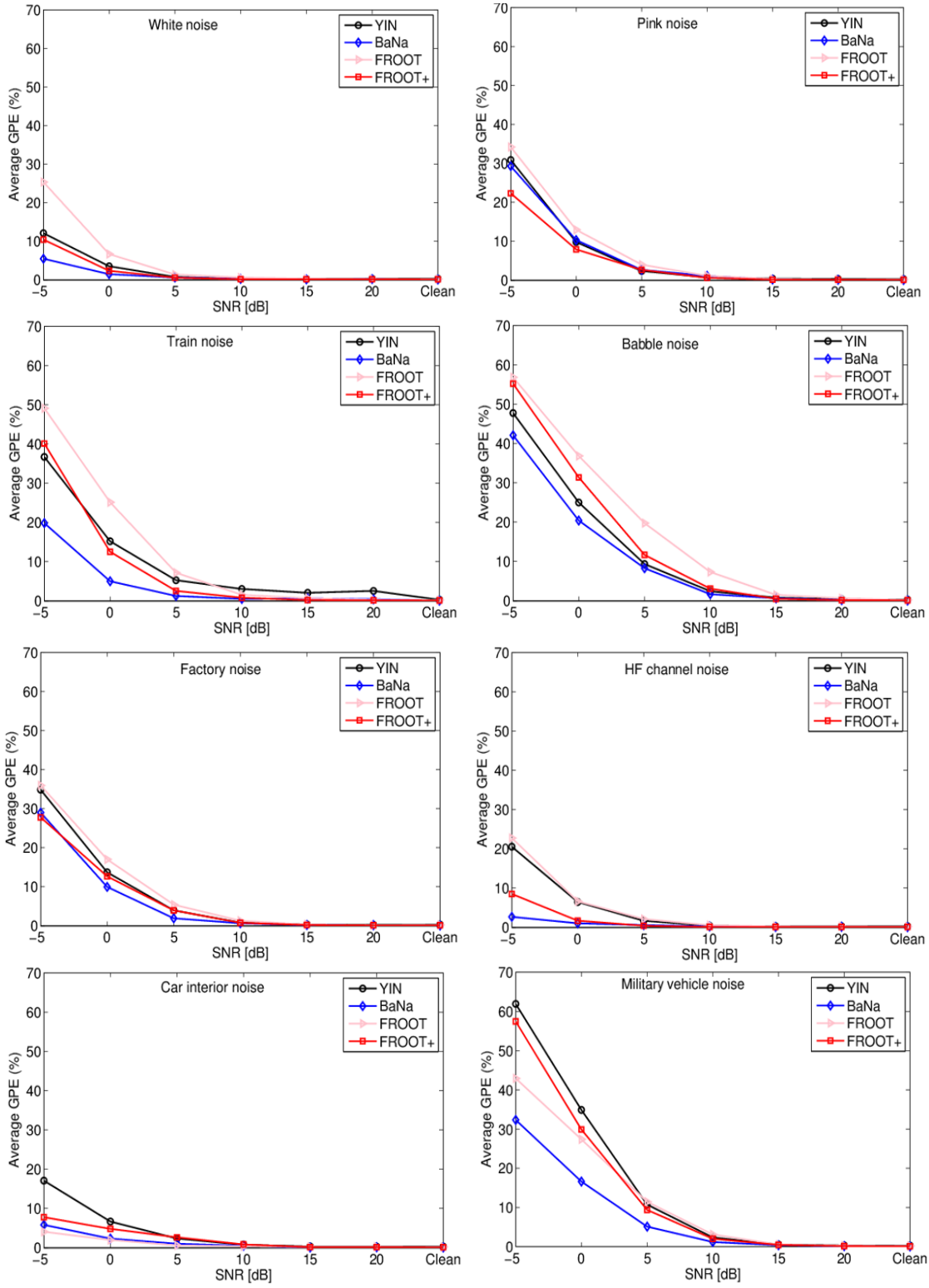

Figure 3.9: GPE for four female speakers with different types of noise under different SNR levels in NTT database.

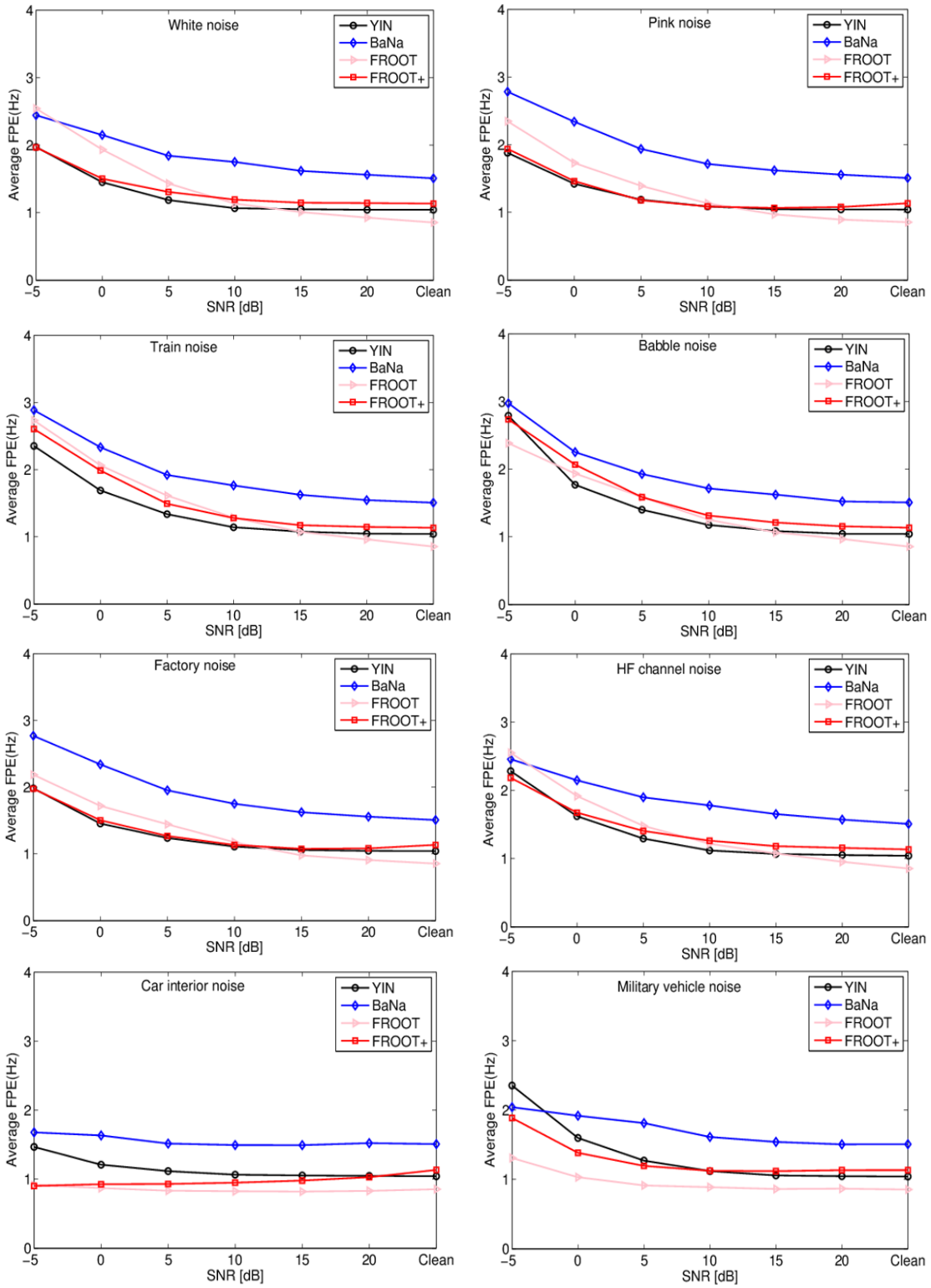

Figure 3.10: FPE for four male speakers with different types of noise under different SNR levels in NTT database

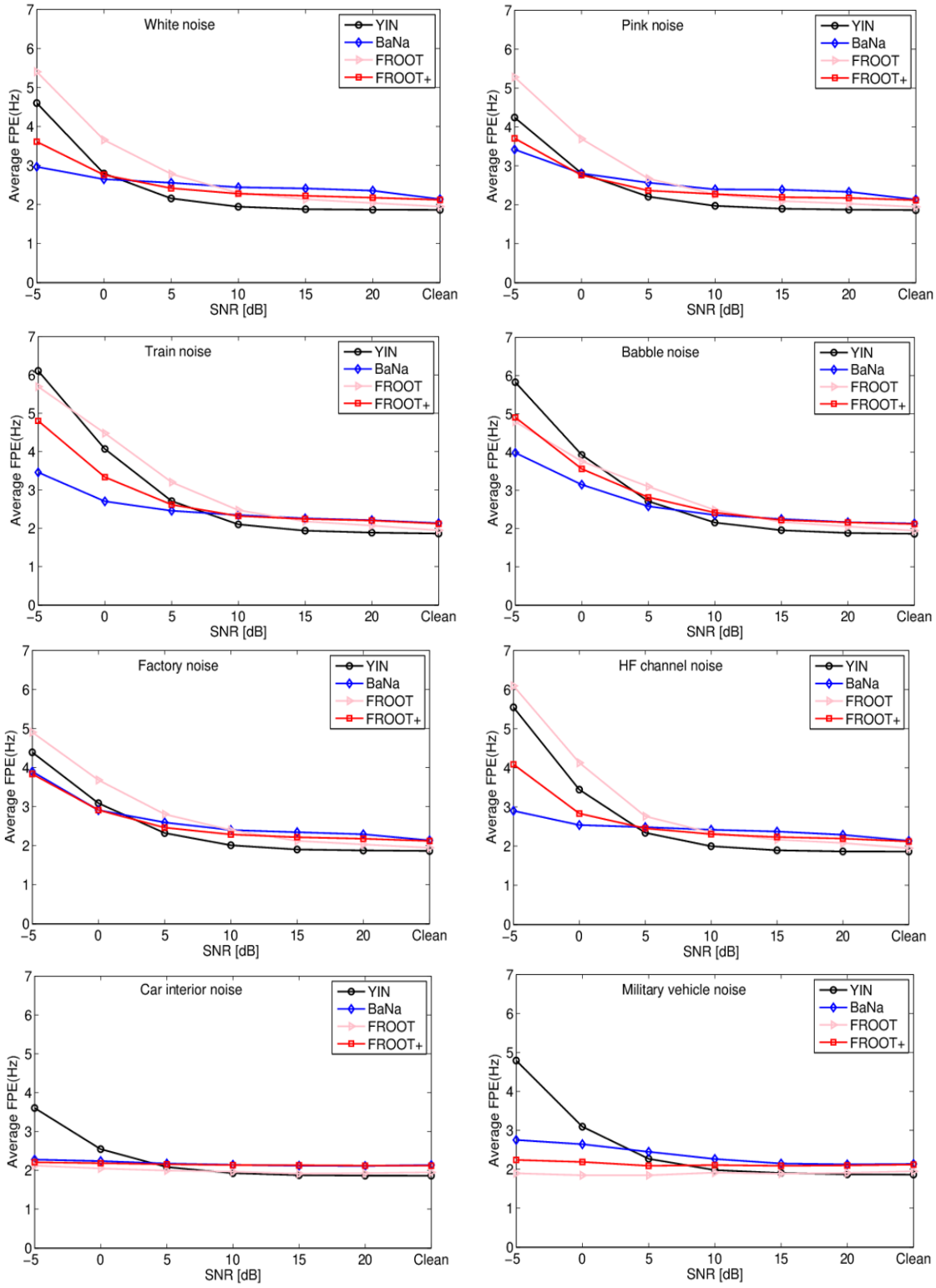

Figure 3.11: FPE for four female speakers with different types of noise under different SNR levels in NTT database

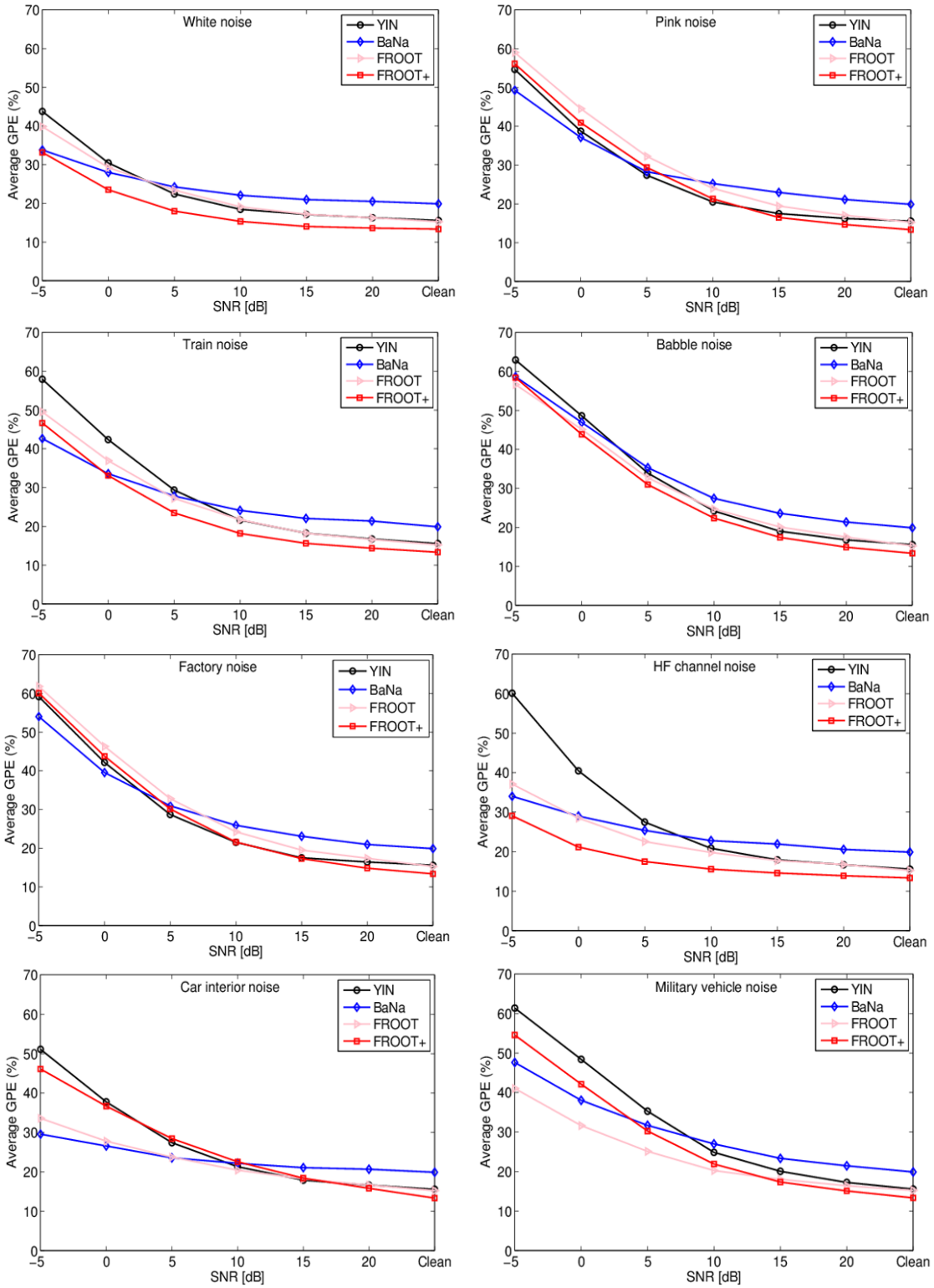

Figure 3.12: GPE for five male speakers with different types of noise under different SNR levels in KEELE database

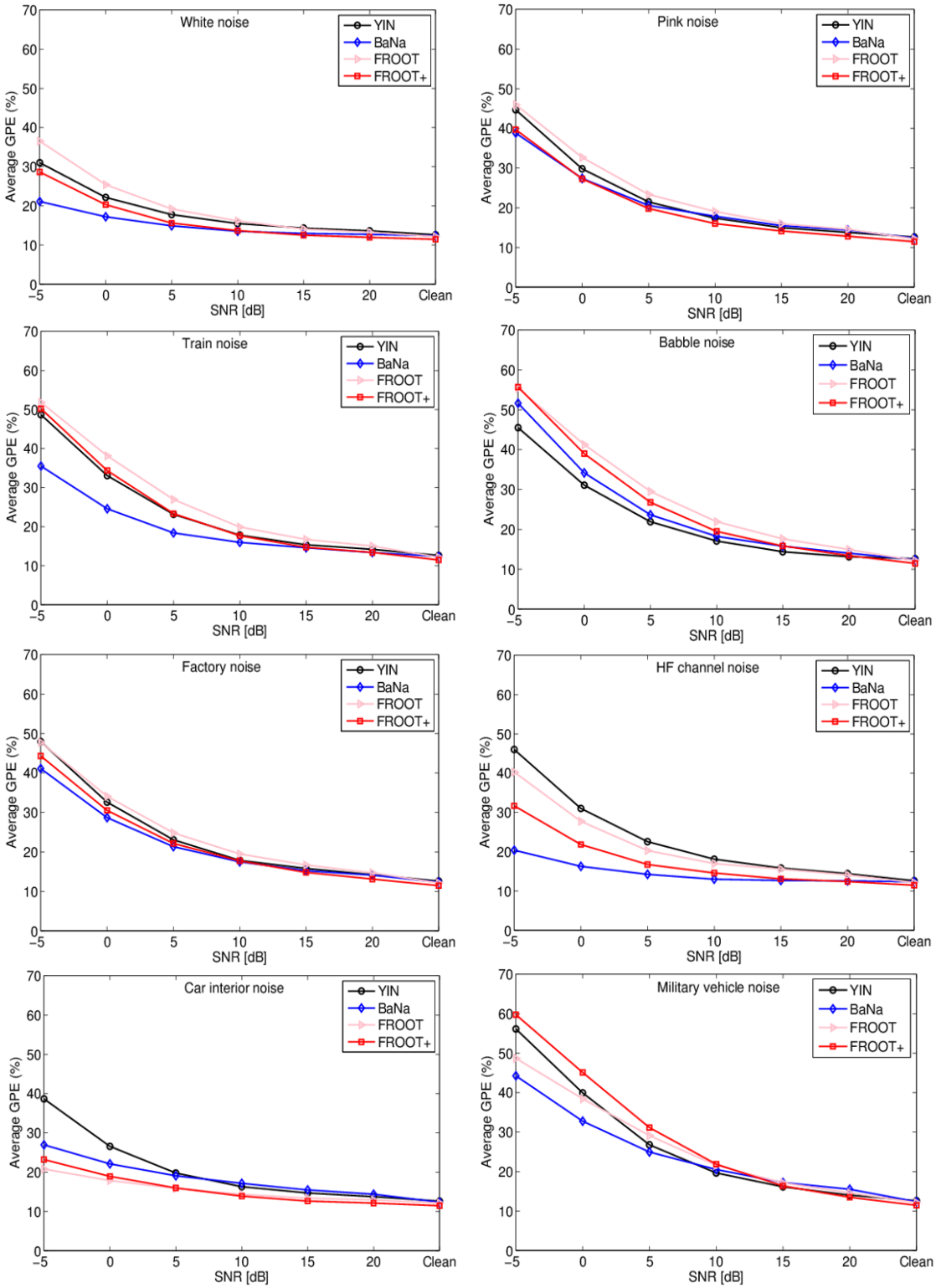

Figure 3.13: GPE for five female speakers with different types of noise under different SNR levels in KEELE database

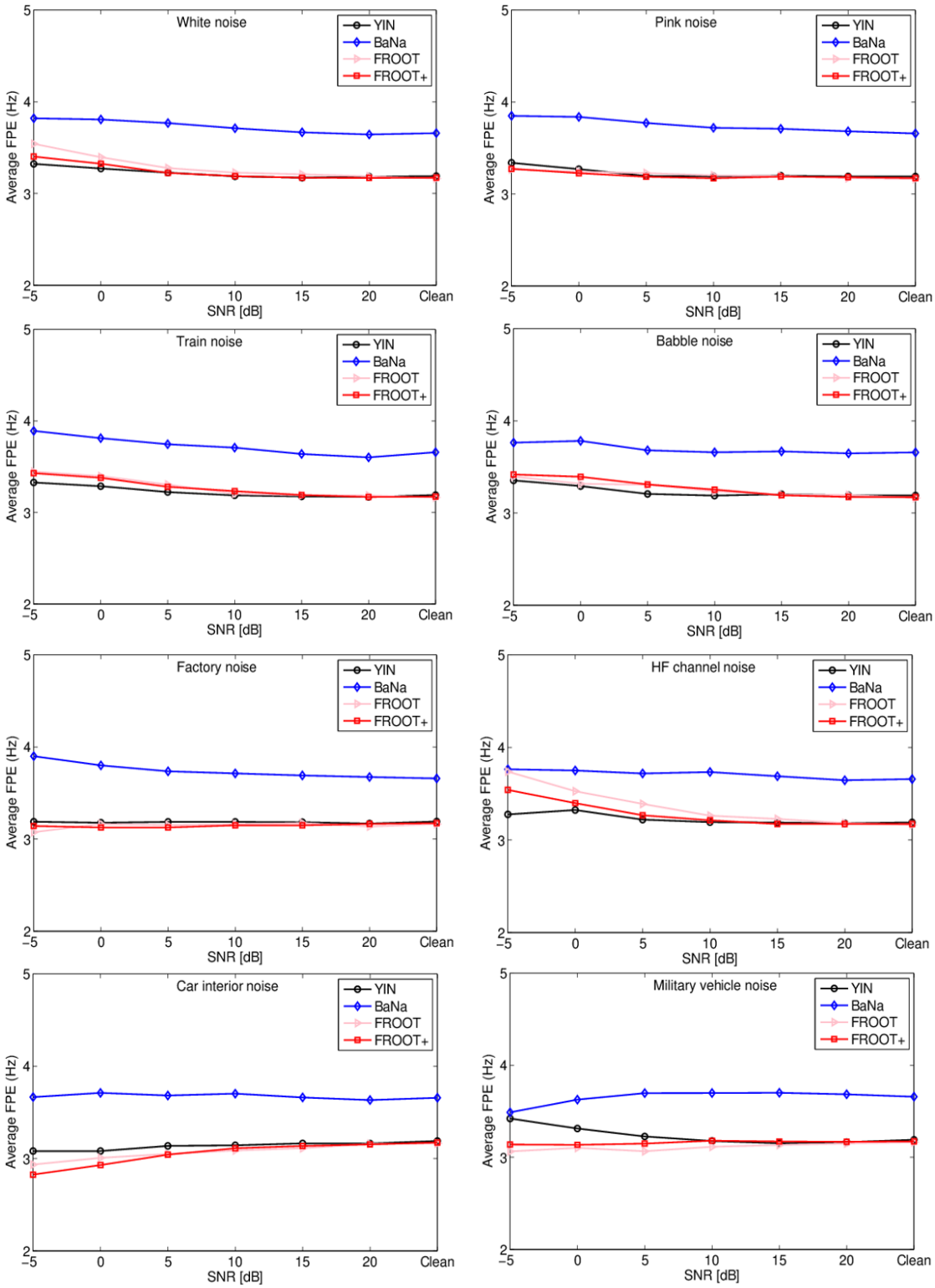

Figure 3.14: FPE for five male speakers with different types of noise under different SNR levels in KEELE database

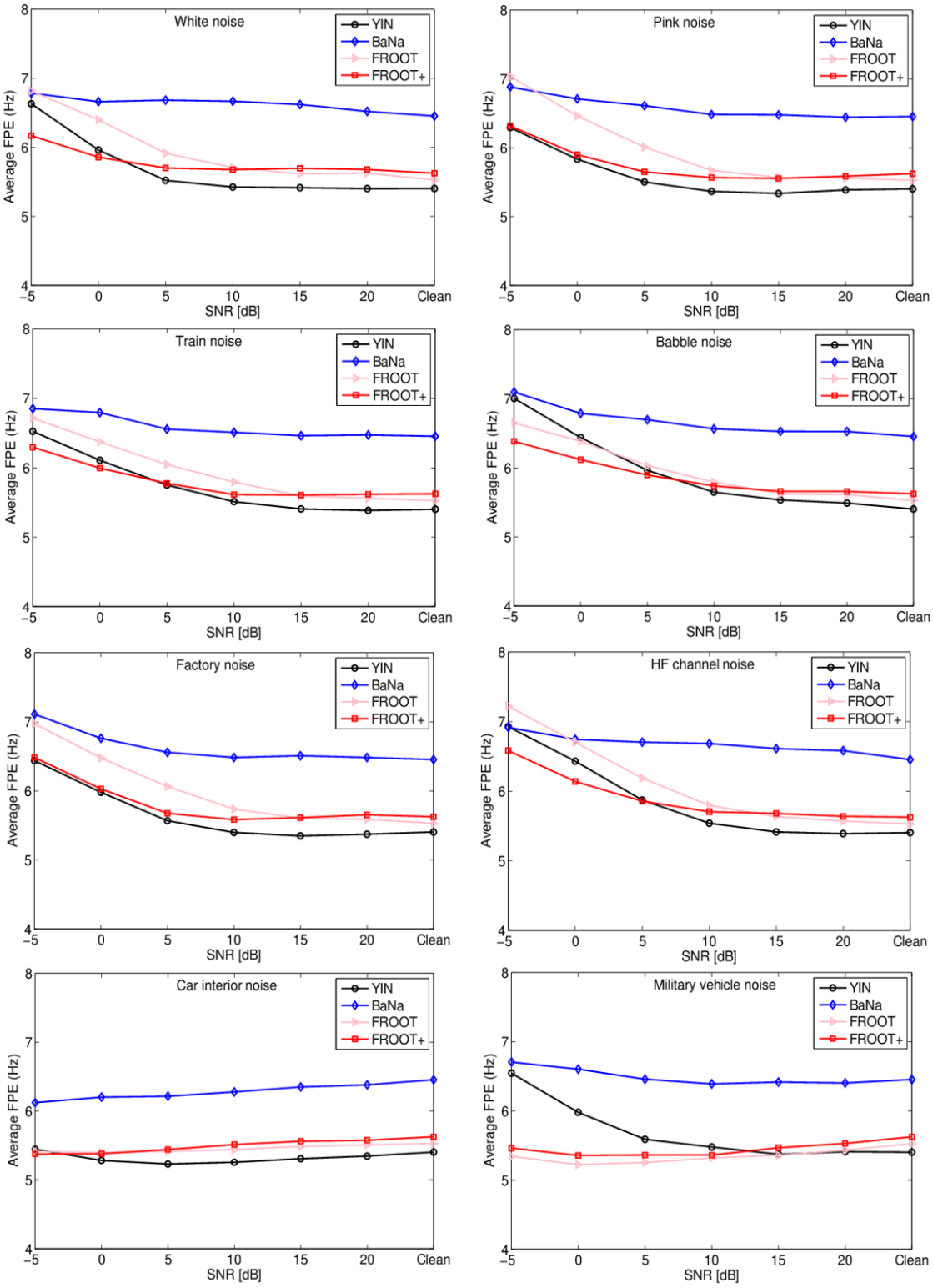

Figure 3.15: FPE for five female speakers with different types of noise under different SNR levels in KEELE database

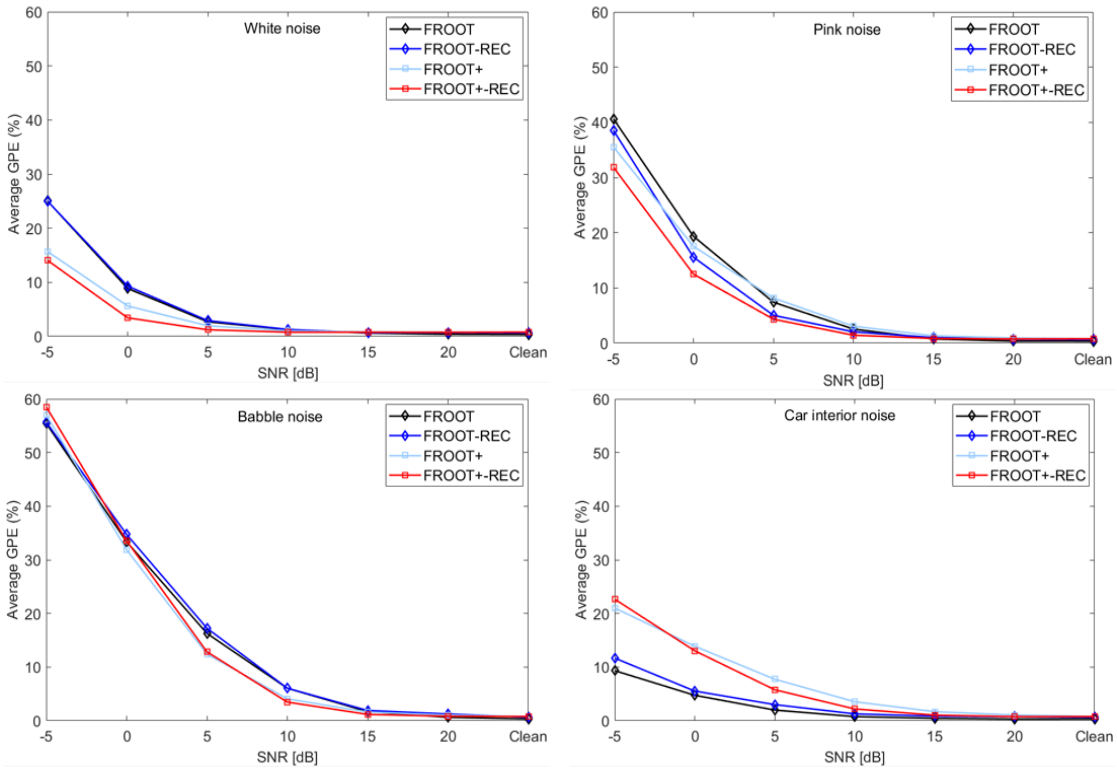

Figure 3.16: GPE for FROOT and FROOT+ methods with different types of noise under different SNR levels in the NTT database

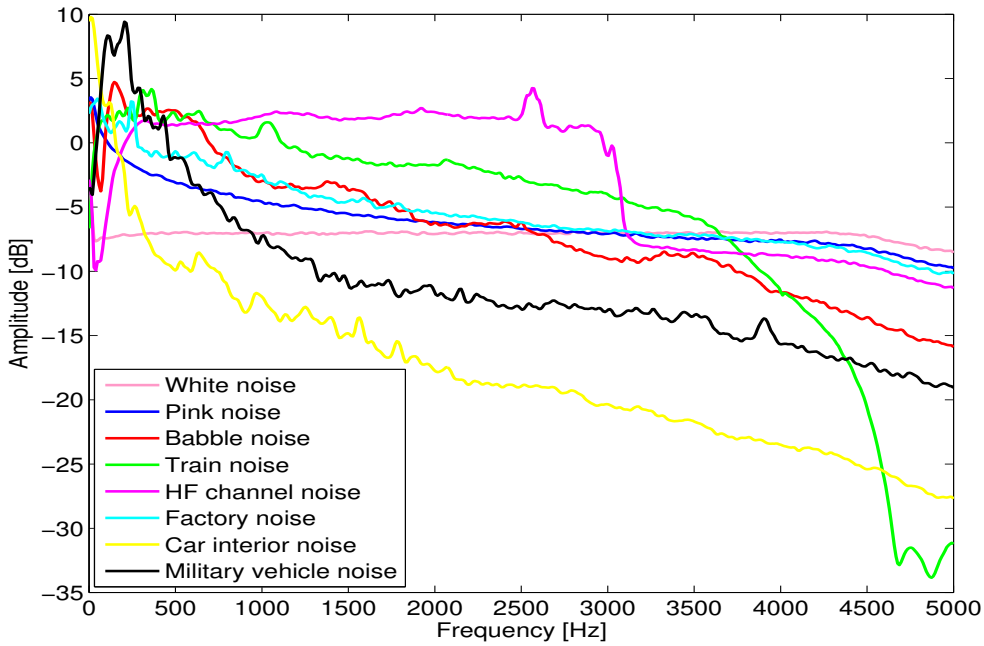

Figure 3.17: Long term spectra of different noises.

Chapter 4

Utilization of Windowing Effect and Accumulated Autocorrelation Function and Power Spectrum for Pitch Detection in Noisy Environments

In this chapter, three types of accumulation techniques are proposed by supporting the ACF based method to further improve the performance of the ACF. The accumulation based pitch detection methods utilize ACFs obtained from a filter bank and power spectra obtained from shorter subframes. These are effective to keep the speech harmonics and to suppress the noise components included in the noisy speech signal. To improve the accuracy of the pitch extraction methods in noisy environments, windowing analysis is more effective where the use of the Rectangular window is emphasized instead of the Hanning or Hamming window. Through experiments, it is shown that the three accumulation based approaches have the potential to provide better performance than recent state-of-the art methods for pitch detection without relying on a complicated post processing technique.

4.1 Introduction

Pitch is an important attribute of human speech, which is originated due to the vibration of vocal folds. Reliable detection of the pitch period (T_0) being the inverse of the fundamental frequency (F_0) from speech is required in a wide range of applications such as speech coding, speech recognition, speech enhancement,

speech synthesis and so on. Therefore, a large number of pitch detection methods have been addressed up to now [18][24-33][36-39][44-48][53][58].

Various pitch detection methods are operated in the time domain. The autocorrelation function (ACF) [24] and average magnitude difference function (AMDF) [25][28] are widely used to accurately detect the pitch by measuring the similarity between the original waveform and its delayed version. By using the properties of ACF and AMDF, a number of methods have been developed. The YIN method [27] uses a cumulative mean normalized square difference function of the speech signal. In [26], to improve the robustness of pitch detection in noisy environments, an improved version of ACF is addressed where the conventional ACF is weighted by the reciprocal of AMDF. Most of the ACF based pitch detection methods are effective in white noise. In general, the pitch detection performance of the ACF based methods is degraded when the clean speech is corrupted by color noise. Also, the ACF is affected by the characteristics of the vocal tract.

For reducing the vocal tract effect, several pitch detection methods are introduced in the frequency domain. The cepstrum (CEP) method [30][31] is one of such most popular methods. The CEP is obtained by the inverse Fourier transform of the log-amplitude spectrum. The logarithmic function in the CEP is used for separating the periodic components from the vocal tract characteristics in the speech signal. The CEP behaves accurately in a noiseless environment, but in noisy environments, the performance of the CEP is severely affected. To combat this problem, improved versions are addressed in the modified CEP (MCEP) [32] and the ACF of the log spectrum (ACLOS) [38], respectively. In [36], a filter bank approach is incorporated in the frequency domain. The windowless ACF (WLACF) based CEP (WLACF-CEP) [33] utilizes both the properties of WLACF and CEP. The WLACF is used for suppressing the noise from the noisy speech signal, while sustaining the periodicity of the speech signal. In the WLACF-CEP, a noise suppressed speech signal is applied to the CEP to enhance the accuracy of pitch detection. In [33], it is shown that the WLACF-CEP behaves robustly against various types of noise.

Recently, two sophisticated approaches have been addressed [46][47]. The pitch estimation filter with amplitude compression (PEFAC) [46] is a frequency domain pitch detection method, which utilizes its sub-harmonic summations [48] in the log frequency domain. The PEFAC also attempts an amplitude compression technique for enhancing its noise robustness. On the other hand, the BaNa [47] considers the noisy speech peaks and results in a hybrid pitch detection method that selects

first five spectral peaks in the amplitude spectrum of the speech signal. The BaNa calculates the ratios of the frequencies of the spectral peaks with tolerance ranges and accurately extracts the pitch of the speech signal.

Traditionally, most of the pitch detection methods are utilized with the Hanning or Hamming window for the segmentation at each frame. This is because the Hanning and Hamming windows are adequate for the purpose of keeping a good balance between sharp peaks of speech harmonics and noise suppression except for speech harmonics. Thus, even in the recent techniques of PEFAC [46] and BaNa [47], these windows are used. However, in both techniques, a comparatively longer frame length is proactively utilized. Specifically, a frame length of 90 [ms] in PEFAC and that of 60 [ms] in BaNa are set up, respectively. Increasing the frame length leads to creating narrowing peaks of speech harmonics. From this point of view, in this paper we set out to use the Rectangular window proactively instead of the Hanning or Hamming window, but keeping a standard length of frame such as 50 [ms] (or less). In noisy environments, a wider bandwidth of each harmonic peak created by involving the Hanning or Hamming window badly behaves due to the corrupted noise. Therefore, these window based methods result in pitch detection errors in noisy environments. To improve the accuracy of the pitch detection methods in noisy environments, the use of the Rectangular window is emphasized instead of the Hanning or Hamming window in this chapter. Supporting the ACF based method, three types of accumulation techniques are derived to further improve the performance of the ACF. The accumulation based pitch detection methods utilize ACFs obtained from a filter bank and power spectra obtained from shorter subframes. These are effective to keep the speech harmonics and to suppress the noise components included in the noisy speech signal.

The remainder of this chapter is organized as follows. Section 4.2 analytically describes the motivation of using the Rectangular window. Section 4.3 explains the proposed accumulation based methods. In Section 4.4, we compare the performance of the proposed methods with that of the conventional methods. Finally, we conclude this chapter in Section 4.5.

4.2 Motivation

Let us assume that a voiced speech signal, $s(n)$ is represented for simplicity by

$$s(n) = \sum_{i=1}^{R} a_i \cos(2\pi F_0 in) \tag{4.1}$$

where n is a discrete time, a_i is the amplitude of each sinusoid ($a_i > 0$), R is the number of sinusoids, and F correspond to the fundamental frequency. In the frequency domain, the speech signal is represented by

$$S(w) = \sum_{i=1}^{R} A_i \delta(w - w_0 i) \tag{4.2}$$

only in the positive frequency region, where $S(w)$ is the (discrete-time) Fourier transform and w corresponds to the angular frequency with $w_0 = 2\pi F_0$. $\delta(w)$ is the Dirac delta function and A_i is the amplitude level such as $A_i = \pi a_i$, where A_i is also positive.

When a window function, $w(n)$, $0 \leq n \leq N-1$, is used, the framed speech signal is given by $s_f(n) = s(n)w(n)$. In the frequency domain, the speech signal and window function are convolved as

$$S_f(w) = S(w) * W(w) \tag{4.3}$$

where $S_f(w)$ and $W(w)$ are the Fourier transforms of $s_f(n)$ and $w(n)$, respectively. From (4.2) and (4.3), we can obtain

$$|S_f(w)| = \sum_{i=1}^{R} A_i |W(w)| \delta(w - w_0 i). \tag{4.4}$$

If the power spectrum of $s_f(n)$ is represented in a sense of Periodogram, one of the non-parametric methods [59], then the resulting power spectrum, $P_f^S(w)$ is given as

$$P_f^S(w) = \frac{1}{N}|S_f(w)|^2 = \sum_{i=1}^{R} \frac{A_i^2}{N} |W(w)|^2 \delta(w - w_0 i). \tag{4.5}$$

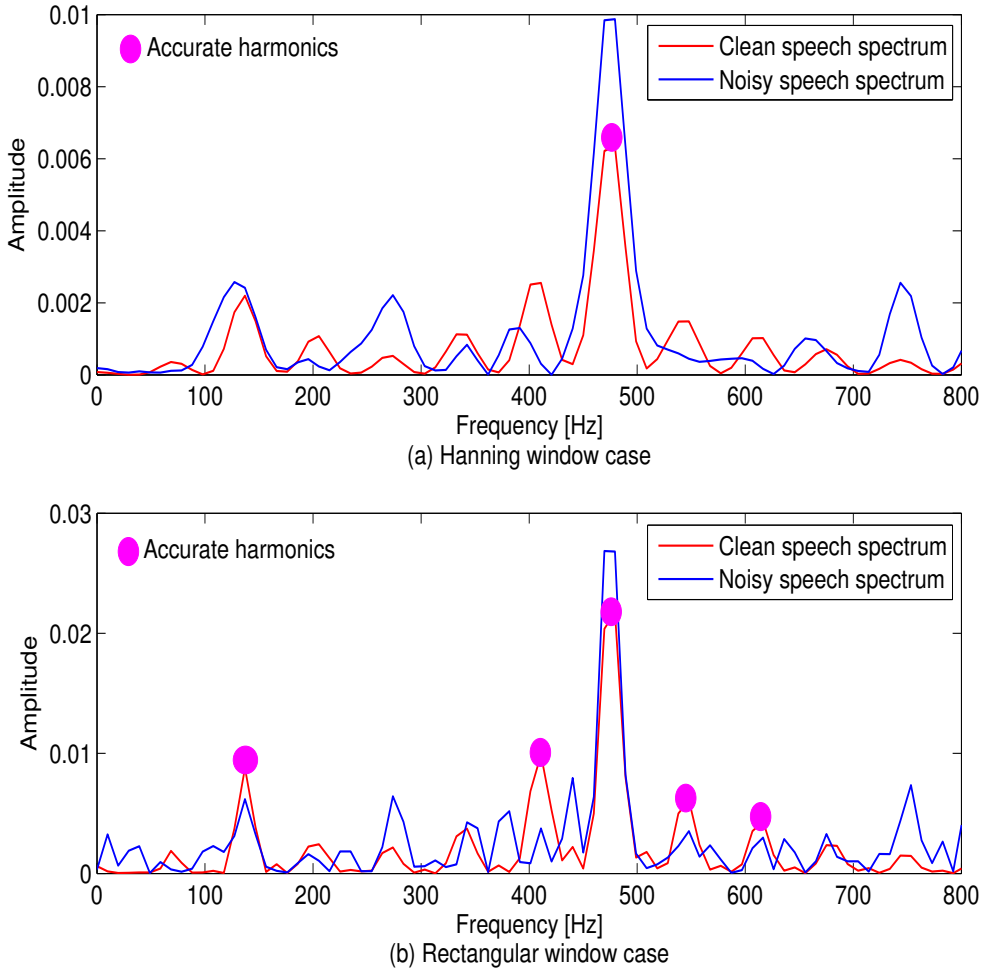

Figure 4.1: Harmonic characteristics of clean and noisy speech signals.

When the Hanning, Hamming, and Rectangular windows are represented by

$$w_{han}(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$$w_{ham}(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

$$w_{rec}(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

respectively, the $W(w)$ of each function is well known as

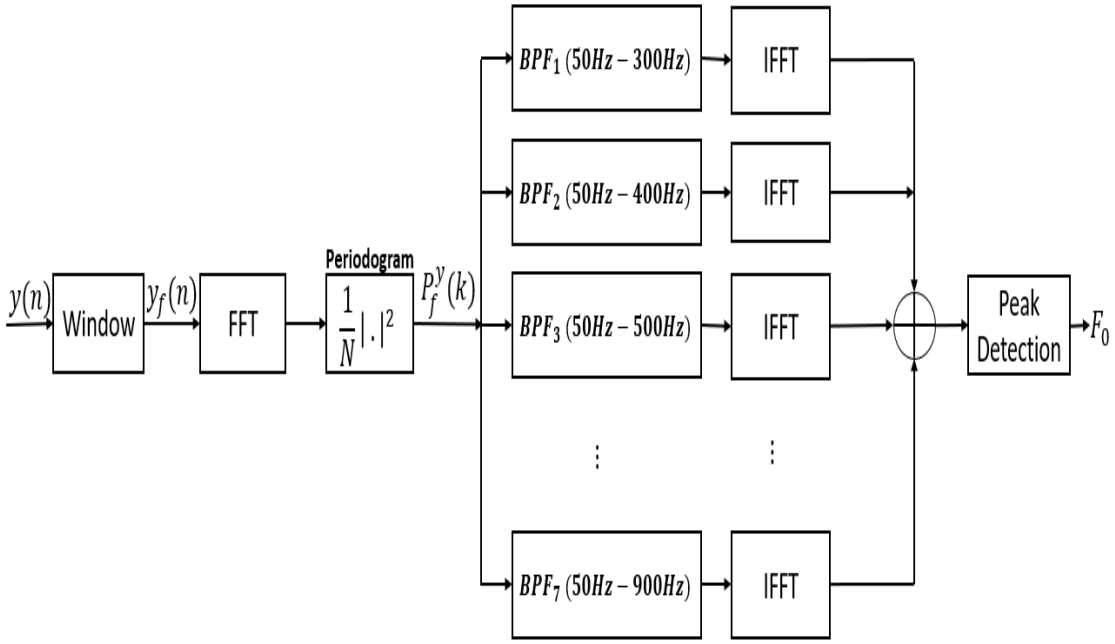

Figure 4.2: Block diagram of AACF approach.

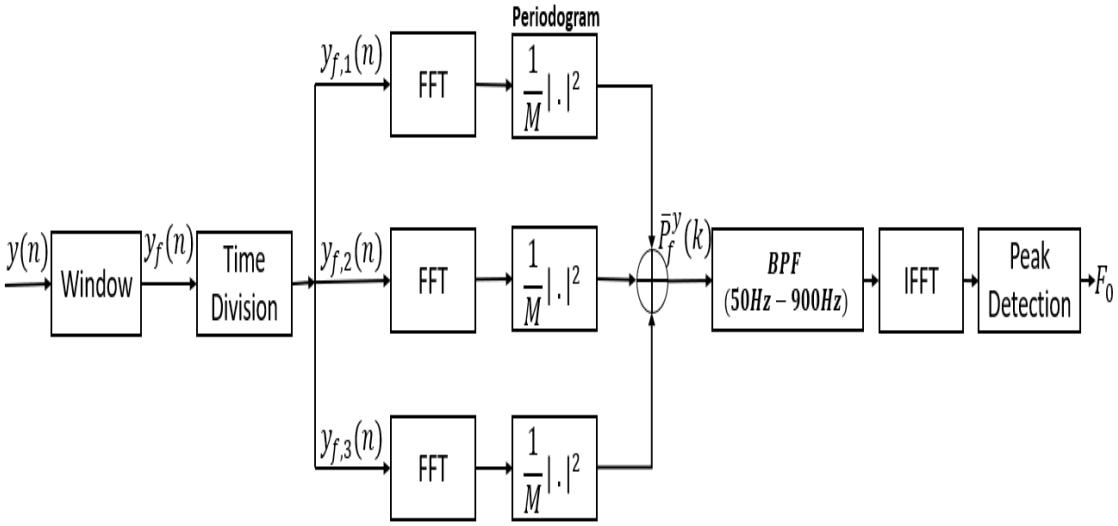

Figure 4.3: Block diagram of APS approach.

$$W_{han}(w) = 0.5W_{rec}(w) + 0.25W_{rec}(w - \frac{2\pi}{N}) + 0.25W_{rec}(w + \frac{2\pi}{N}) \tag{4.9}$$

$$W_{ham}(w) = 0.54W_{rec}(w) + 0.23W_{rec}(w - \frac{2\pi}{N}) + 0.23W_{rec}(w + \frac{2\pi}{N}) \tag{4.10}$$

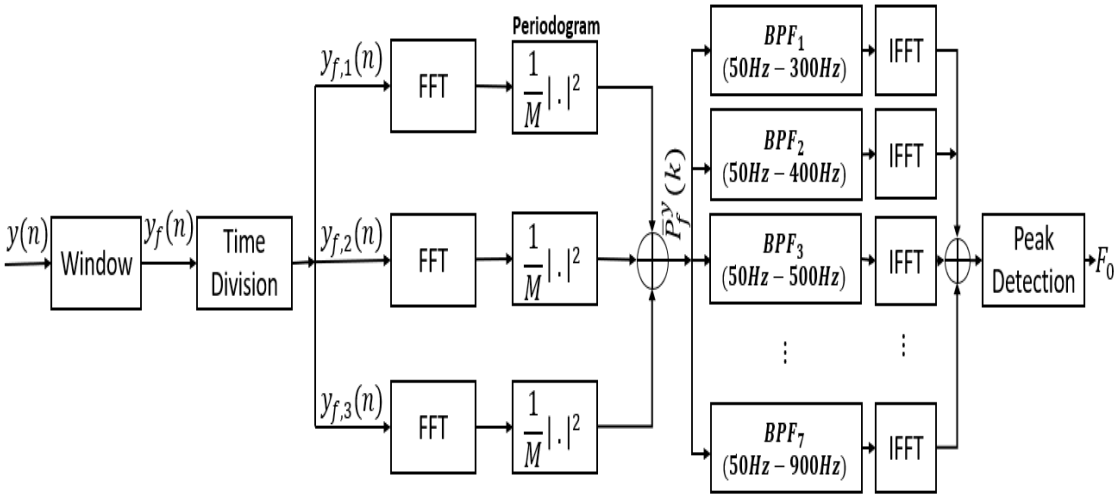

Figure 4.4: Block diagram of C approach.

$$W_{rec}(w) = \frac{\sin(wN/2)}{\sin(w/2)} \tag{4.11}$$

respectively [59][60]. The above expression of each $W(w)$ is a version shifted to the left by $(N-1)/2$ samples, that is, the case of the definition range being $-(N-1)/2 \leq n \leq (N-1)/2$. The bandwidth of the mainlobe of each window is $\frac{4\pi}{N}$, $\frac{4\pi}{N}$, and $\frac{2\pi}{N}$, respectively. Commonly, when N is increased, the mainlobe bandwidth is decreased. In this case, even if the speech signal is corrupted by noise, the affect of noise is suppressed. This is the motivation in [46][47], where a long length of N is used commonly.

In this chapter, we consider to use the Rectangular window instead of the Hanning and Hamming windows, and keep a standard length of N. By this strategy, it is expected that we could obtain a similar effect of noise suppression, since the mainlobe bandwidth of the Rectangular window is half of that of the Hanning or Hamming window. Figure 4.1 shows an example of comparison of windowing effects. A clean speech signal in the NTT database [52] and its noisy version (white noise corruption at 0 [dB] signal-to-noise ratio (SNR)) are used here, whose sampling frequency is 10 [kHz]. The amplitude spectrum of the framed signal, $|S_f(w)|$ in (4.4), is drawn for the Hanning and Rectangular windows, respectively. The corresponding noisy signal version is overlapped, respectively. In Fig. 4.1, only a part of the amplitude spectrum, 0 [Hz]-800 [Hz] frequency region, is drawn for clear visibility. When the harmonic peak location is not shifted for the clean and noisy speech spectra, a circle is marked. From Fig. 4.1, it is obvious that the Rectangular window case is more accurate preserving the speech harmonics. This

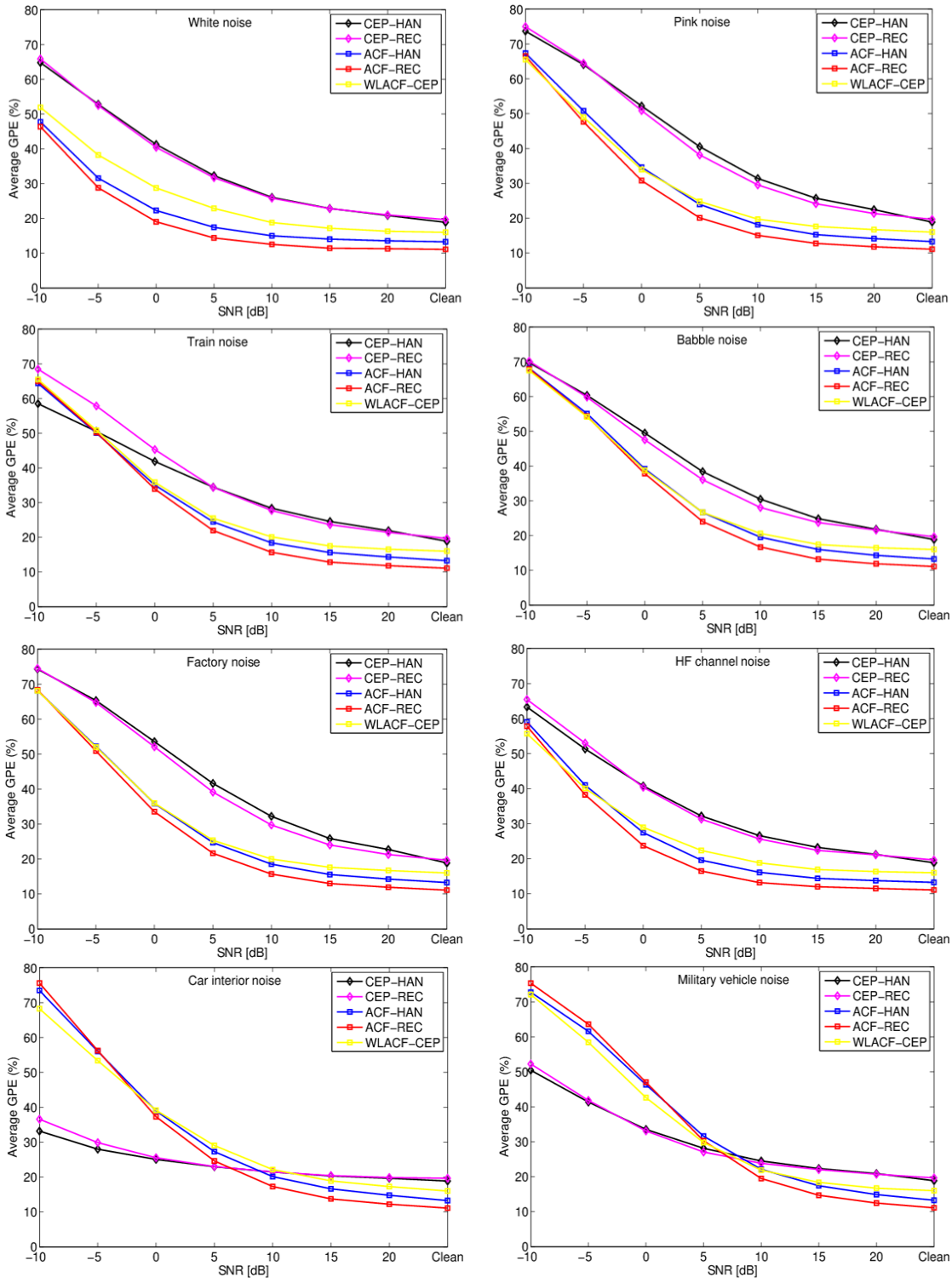

Figure 4.5: GPE for conventional methods with different types of noise under different SNR levels

property leads to more accurate pitch detection even in noisy environments.

4.3 Proposed Methods

Let us assume that the clean speech signal, $x(n)$, is corrupted by noise, $v(n)$. The noisy speech signal, $y(n)$, is expressed as

$$y(n) = x(n) + v(n). \tag{4.12}$$

In this chapter, based on the use of the Rectangular window, we further improve the ACF method. The frame length is set to 50 [ms]. This length is shorter than 90 [ms] in PEFAC [46] and 60 [ms] in BaNa [47], but it is a very often used level of frame length in the literature [44] (exactly speaking, 51.2 [ms] with 10 [kHz] sampling is used in [32-33][36][38][58], but this is almost the same as 50 [ms]).

We propose the following three approaches:
(1) Accumulated Autocorrelation Function (AACF) approach
(2) Accumulated Power Spectrum (APS) approach
(3) Combination (C) approach
Each approach is specifically explained in the next. The aim of each approach is to enhance the speech harmonics suppressing the noise components.

(1) AACF Approach

Figure 4.2 shows a block diagram of the AACF approach. The framed signal, $y_f(n), 0 \leq n \leq N-1$, is transformed into the frequency domain by means of Periodogram calculation with the fast Fourier transform (FFT). The resulting spectrum, power spectrum of $y_f(n)$, is represented by $P_f^y(k)$ where k is the frequency bin number related with a discrete representation of w, w_k. In the following each band pass filter (BPF) with different passband, the components of $P_f^y(k)$ except for the passband are forced to zeros directly in the frequency domain. Each band pass filtered power spectrum $P_{f,l}^y(k)$, $l = 1, 2, ..., 7$ where l corresponds to the BPF number such as $BPF_1, BPF_2, ..., BPF_7$, is transformed into the autocorrelation domain by means of the IFFT. The resulting each ACF, $r_{f,l}^y(m)$, $l = 1, 2, ..., 7$ where m corresponds to a lag, is accumulated as

$$\bar{r}_f^y(m) = \sum_{l=1}^{7} r_{f,l}^y(m) \tag{4.13}$$

for each lag number. Then, by finding the maximum peak location in the funda-

mental frequency range most of people have, we detect the fundamental frequency of $y_f(n)$.

(2) APS Approach

Figure 4.3 shows a block diagram of the APS approach. The framed signal, $y_f(n), 0 \leq n \leq N-1$ is divided into three subframes through the time division part as

$$y_{f,1}(n) = y_f(n), \quad 0 \leq n \leq M-1 \tag{4.14}$$

$$y_{f,2}(n-D) = y_f(n), \quad D \leq n \leq D+M-1 \tag{4.15}$$

$$y_{f,3}(n-2D) = y_f(n), \quad 2D \leq n \leq 2D+M-1 \tag{4.16}$$

where M is an integer which corresponds to the subframe length and D is a frame shift sample. In general, it is desired that $2D+M-1$ is set so as to be equivalent to N. In Section 4.4, the length of M and that of D are set to 30 [ms] and 10 [ms], respectively.

From each subframe $y_{f,j}(n)$, $j = 1, 2, 3$, $0 \leq n \leq M-1$, each power spectrum is calculated as $P^y_{f,1}(k)$, $P^y_{f,2}(k)$, and $P^y_{f,3}(k)$. The three power spectra are accumulated for each frequency bin as

$$\bar{P}^y_f(k) = \sum_{j=1}^{3} P^y_{f,j}(k). \tag{4.17}$$

The accumulated power spectrum $\bar{P}^y_f(k)$ is band pass filtered by forcing to zeros except for the passband 50 [Hz] - 900 [Hz]. The resulting power spectrum is inverse Fourier transformed by the IFFT, and by finding the maximum location on the resulting ACF, the fundamental frequency of $y_f(n)$ is detected.

(3) C Approach

In this approach, the APS approach is combined with the AACF approach as shown in Fig. 4.4. The accumulated power spectrum $\bar{P}^y_f(k)$ is calculated from the framed signal $y_f(n)$, which is used instead of the power spectrum $P^y_f(k)$ in Fig. 4.2.

4.4 Experiments

To investigate the performance of the accumulation based approaches, we conducted experiments on speech signals.

4.4.1 Experimental Condition

Speech signals are taken from the KEELE database [53]. We utilize five male and five female speech signals spoken in English from the KEELE database. The total length for the ten speakers' speeches are about 6 [m]. These speech signals were sampled at a rate of 16 [kHz]. To generate noisy speech signals, we added different types of noise to the speech signals. White noise with zero mean and unit variance was generated by a computer and added to the speech signals with amplitude adjustment. Pink noise, babble noise, factory noise, HF channel noise, car interior noise, and military vehicle noise were taken from the NOISEX-92 database [54] with a sampling frequency of 20 [kHz], and train noise was taken from the Japanese Electronic Industry Development Association (JEIDA) noise database [55] with a sampling frequency of 8 [kHz]. These noises were resampled with a sampling frequency of 16 [kHz], respectively, when they are added to the speech data in KEELE database. The SNR was set to -10, -5, 0, 5, 10, 20, ∞ [dB] and the other experimental conditions for fundamental frequency extraction were

- frame length: 50 [ms] except for PEFAC and BaNa;
- frame shift: 10 [ms];
- window function: Rectangular and Hanning except for PEFAC;
- DFT (IDFT) length: 2048 points except for PEFAC and BaNa;

The following fundamental frequency extraction error $e(l)$ was used for the evaluation of fundamental frequency extraction accuracy based on Rabiner's rule [18];

$$e(l) = F_{est}(l) - F_{true}(l) \tag{4.18}$$

where l corresponds to the frame number, $F_{est}(l)$ and $F_{true}(l)$ are the fundamental frequency extraction from the noisy speech signal, and the true fundamental frequency at the l-th frame, respectively. If $|e(l)| > 10[\%]$ from the ground truth fundamental frequency, we recognized the error as gross pitch error (GPE) and calculated the GPE rate (in percentage) over the total voiced frames included in the speech data. We detected and assessed only voiced parts in sentences for the fundamental frequency extraction. For extracting the fundamental frequency, we used the search range of $f_{max} = 50$ [Hz] and $f_{min} = 400$ [Hz], which corresponds to the fundamental frequency range most of people have.

Table 4.1: GPE for PEFAC with different types of noise under different SNR levels.

	Noise type							
SNR[dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	44.48	**57.50**	60.66	70.53	62.88	52.21	**30.70**	**50.38**
-5	36.13	**43.77**	47.68	57.89	49.13	40.73	28.26	**42.45**
0	31.10	35.44	38.83	46.14	38.62	34.53	27.04	35.89
5	28.08	31.20	33.24	37.23	32.38	30.71	26.51	31.41
10	26.43	27.79	29.11	31.75	29.06	28.07	26.46	28.86
15	25.56	26.60	26.76	28.58	29.06	26.26	26.56	27.55
20	25.14	26.01	25.16	26.79	26.36	25.18	26.63	27.03

The ground truth information for the fundamental frequency at each frame is included in the KEELE database. Therefore, the $F_{true}(l)$ values in (4.18) are known a priori to evaluate.

4.4.2 Performance Comparison

In this subsection, the accumulation based approaches are compared to the conventional methods; ACF [24], CEP [30], WLACF-CEP [33], PEFAC [46], and BaNa [47].

All parameters of the conventional methods are the same as those in the accumulation based approaches, except for the frame length, DFT(IDFT) points, window function of PEFAC and the frame length and DFT (IDFT) points of BaNa, respectively. We used the Hamming window function for the PEFAC and the frame length was set as 90 [ms] according to the suggestion in [46]. The DFT (IDFT) points were 2^{13} which was used in the source code. The source code to implement the PEFAC was collected from [61]. For the BaNa, the frame length was set as 60 [ms] and the DFT (IDFT) points were 2^{16} according to the suggestion in [47]. The source code to implement the BaNa was collected from [57].

Figure 4.5 and Tables 4.1-4.6, respectively, show the average GPE rate on five male and five female speech signals in the KEELE database with different noises. When the SNR is changed from -10 [dB] to infinity [dB] (clean speech case), each plot has been obtained under each SNR condition.

In Fig. 4.5, only the ACF, CEP, and WLACF-CEP are compared. The Rectangular and Hanning windows are denoted as REC and HAN, respectively. Except for the car interior noise and military vehicle noise cases, the ACF provides better performance than the CEP, and a lower GPE rate is obtained with the Rectangu-

Table 4.2: GPE for BaNa with different types of noise under different SNR levels.

	Noise type							
SNR[dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	36.46	59.02	**51.63**	68.45	**62.80**	**36.33**	34.33	57.37
-5	27.44	44.06	**39.07**	55.15	**47.57**	27.19	**28.25**	45.94
0	22.61	32.25	**29.08**	40.54	34.12	22.64	**24.34**	**35.40**
5	19.58	24.42	23.11	29.48	26.14	19.82	**21.31**	**28.32**
10	17.80	21.56	20.04	22.84	21.70	17.90	19.65	23.76
15	16.97	19.23	18.31	19.69	19.16	17.31	18.26	20.27
20	16.59	17.74	17.36	17.70	17.53	16.57	17.52	18.48

Table 4.3: GPE for AACF with different types of noise under different SNR levels.

	Noise type							
SNR[dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	39.50	77.89	65.58	70.96	76.94	44.54	78.30	77.53
-5	24.39	60.13	50.75	57.53	60.48	27.44	60.22	67.91
0	16.40	38.60	34.78	40.34	40.16	17.54	37.33	52.11
5	**12.66**	22.94	22.00	24.78	24.23	**13.26**	22.49	34.26
10	**11.07**	15.60	15.12	16.47	15.97	**11.37**	**14.90**	20.82
15	**10.45**	**12.40**	**12.15**	**12.49**	**12.43**	**10.49**	**11.88**	14.82
20	**10.06**	**10.96**	**10.95**	**10.82**	**10.91**	**10.00**	**10.61**	**12.13**

lar window. This windowing effect comes from the principle described in Section. 4.2. In the train noise case, although the Hanning window leads to a lower GPE than the Rectangular window at low SNRs (<5 [dB]), this might be due to a rapid and strong time variation nature of the train noise. The two noise cases of car interior and military vehicle show a different tendency. The CEP provides better performance than the ACF at low SNRs (<5 [dB]). This could be due to a strong periodical nature of the two noises. See Fig. 4.6 where spectrogram characteristics of all noises used in the experiment are shown. Figure 4.7 shows long-term spectra of all noises. From Figs. 4.6 and 4.7, we can observe that the car interior noise produces a sharp narrow band peak and the military vehicle noise does closed narrow band peaks, respectively. These give a strong periodical nature of the noise at low SNRs. The CEP is known to behave robustly against periodical noises.

In [33], a performance comparison of the WLACF-CEP with ACF based methods was not revealed. As indicated in [33], Fig. 4.5 shows better performance of the WLACF-CEP for a variety of noise being white, pink, train, babble, factory, and HF channel, but the ACF is more robust. On the other hand, for car interior and military vehicle noises, the CEP is still better than the WLACF-CEP

Table 4.4: GPE for APS with different types of noise under different SNR levels.

	Noise type							
SNR[dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	36.93	71.46	62.99	**67.99**	70.67	47.16	73.54	72.70
-5	23.79	51.58	47.69	**53.97**	52.71	29.84	51.94	61.29
0	17.10	32.80	32.12	**37.32**	34.48	19.62	33.24	44.75
5	13.48	21.31	21.82	24.11	22.32	14.91	21.37	28.90
10	12.15	15.82	15.87	17.01	15.93	12.73	15.68	19.52
15	11.38	13.06	13.19	13.32	13.28	11.56	13.09	14.88
20	11.03	11.82	11.87	11.81	11.93	11.09	11.77	12.62

Table 4.5: GPE for C approach with different types of noise under different SNR levels.

	Noise type							
SNR [dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	**35.53**	74.66	63.23	68.73	73.51	41.07	77.87	74.07
-5	**22.67**	54.86	47.71	55.01	55.41	**25.68**	56.65	63.84
0	**16.13**	34.61	31.70	37.63	35.77	**17.50**	35.27	47.70
5	12.86	21.69	**21.07**	**23.65**	22.54	13.52	21.95	30.59
10	11.53	15.59	**14.99**	**16.42**	15.71	11.77	15.61	19.93
15	10.84	12.74	12.54	12.74	12.91	10.98	12.70	14.76
20	10.52	11.45	11.37	11.36	11.51	10.52	11.39	12.29

especially at low SNRs.

We compare the PEFAC, BaNa, and three accumulation based approaches (AACF, APS, and C) in Tables 4.1-4.5, respectively. We show the performance of the ACF in Table 4.6 as the bench mark, which is exactly the same as the ACF with the Rectangular window in Fig. 4.5. The best score in each noise and SNR condition case is highlighted by bold face.

From these Tables, it is observed that the PEFAC and BaNa provide comparatively better performance at low SNRs as emphasized in each original paper [46][47]. However, that is not satisfied in all noise cases. In the babble noise case, the APS provides better score than both the methods at low SNRs. In the white and HF channel noise cases, the C provides better score except for -10 [dB] SNR case of the HF channel noise at low SNRs. At high SNRs, the AACF is very excellent. The AACF provides all best scores for all noise cases at 20 [dB] SNR. It also provides all best scores for noises except for the military vehicle noise at 15 [dB] SNR. At middle SNRs, the C is excellent especially for the train and babble noises, and competitive with the ACF in the pink and factory noises. These results

Table 4.6: GPE for ACF with different types of noise under different SNR levels.

	Noise type							
SNR [dB]	White	Pink	Train	Babble	Factory	HF channel	Car interior	Military vehicle
-10	46.36	66.49	64.97	67.99	68.33	57.82	75.59	75.31
-5	28.79	47.61	50.23	54.33	50.89	38.27	56.24	63.59
0	19.03	**30.76**	33.91	37.84	**33.50**	23.72	37.29	47.01
5	14.36	**20.09**	21.94	24.00	**21.62**	16.49	24.61	30.26
10	12.53	**15.05**	15.66	16.68	**15.70**	13.19	17.29	**19.48**
15	11.42	12.75	12.83	13.18	12.98	12.02	13.74	**14.69**
20	11.27	11.78	11.80	11.86	11.92	11.52	12.21	12.45

suggest that the accumulation based approaches should be employed adequately knowing the information about the noise type and SNR degree. It should be noted here that a complicated and time-consuming post-processing to track the pitch information with time is not included in the accumulation based approaches.

For a comparison of the AACF, APS and C approaches, it should be noted that the APS and C approaches are more effective at low SNRs ($\leq$10 [dB]) in almost all noise cases than the AACF approach in Tables 4.3-4.5. In Tables 4.4 and 4.5, when the GPE values are lower than those in Table 4.3, the locations are marked by yellow color. This result indicates that the time division processing commonly used in the APS and C approaches provides a good effect for suppressing the noise especially in low SNR cases. Also, we can realize that the yellow marked ranges are wider in Table 4.5 by comparing Tables 4.4 and 4.5 carefully. As shown in Table 4.3, the AACF approach is very excellent at high SNRs regardless of noise types. These two observations could explain the combined property of the C approach, that is, a synergistic effect of combining the frequency division processing in the AACF approach with the time division processing in the APS approach appears in the C approach.

We further investigated the relationship between the windowing effect and frame length for pitch detection. In Fig. 4.8, the ACF, AACF, PEFAC, and BaNa methods with the Rectangular and Hanning windows are compared when the frame length is changed from 30 [ms] to 90 [ms]. Here, the typical four noise types of white, pink, babble and car interior are used and the SNR is set to +5 [dB] commonly. Fig. 4.8 shows that the use of the Rectangular window is better than that of the Hanning window regardless of the choice of pitch detector. Also, the GPE rate is increased as the frame length is increased for most of the cases. Only the PEFAC method provides worse results in shorter frame length cases such

as less than 50 [ms]. These results could not change largely the conclusive results of performance comparison in Tables 4.1-4.6.

4.4.3 Processing Time

In Table 4.7, we have compared the processing time per one-second data for each method in the KEELE database. We have tested all methods on a Laptop with Intel (R) Core(TM) i7-8565U, 2 [GHz] clock speed of CPU and 16 [Gigabytes] of memory. For the evaluation, we have used five trials for each method, then calculated the average processing time to obtain reliable measurements. The computational time of the BaNa method is so long because BaNa used the large FFT size for keeping high frequency resolution. The processing time of the PEFAC method provides the second largest processing time because of large FFT size. The AACF and C approaches are commonly shorter processing time than that of PEFAC and BaNa. On the other hand, ACF and APS approaches are almost similar, which are commonly shorter than that of the other methods, since the IFFT operation are directly applied to the conventional power spectrum and accumulated power spectrum, respectively.

Table 4.7: Processing time per second of speech

PEFAC	BaNa	ACF	AACF	APS	C
1.994	26.194	0.473	1.341	0.474	1.342

4.5 Summary

In this chapter, windowing effects have been discussed analytically and experimentally and the usefulness of the Rectangular window in noisy environments has been found. A variety of noise types in practice have been considered and the ACF has been extended to its three accumulation based approaches.

When the ACF is used with the Rectangular window, it provides better GPE rates than the CEP especially at SNRs higher than 5 [dB] regardless of noise types (considered in this paper).

When the ACF is extended to the accumulated ACF approach in which a BPF bank is employed, the accumulation technique provides a further better GPE rate

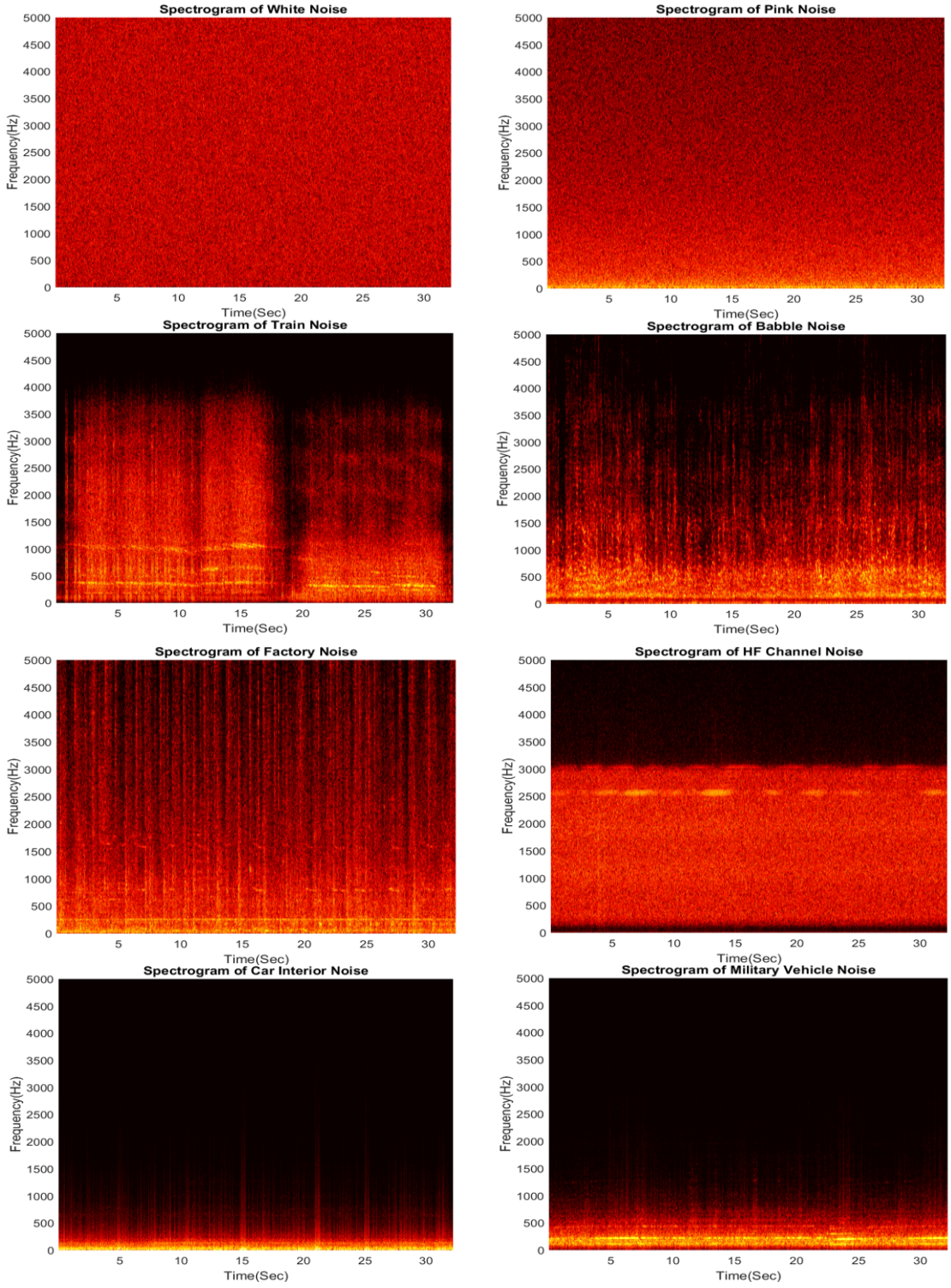

Figure 4.6: Spectrograms for different types of noise

in high SNR cases. When the ACF is extended to the accumulated power spectrum

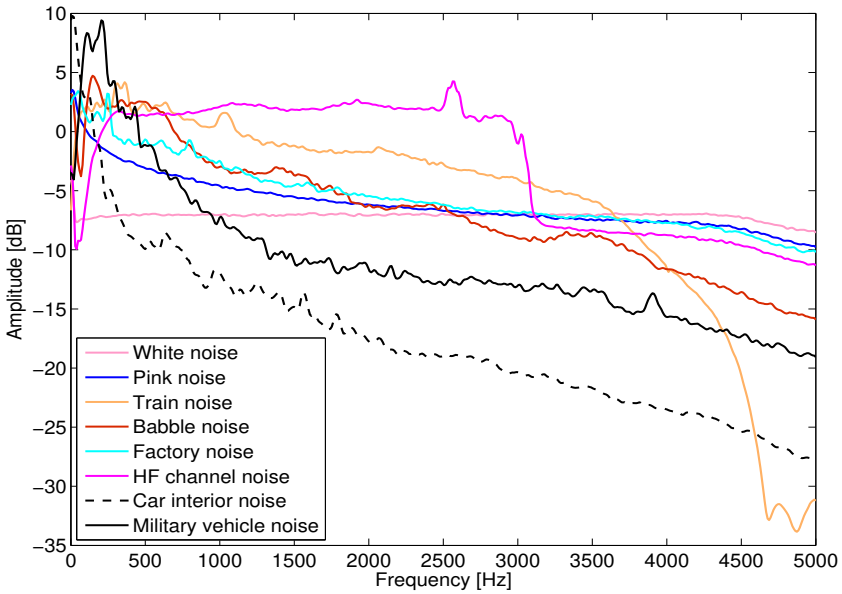

Figure 4.7: Long term spectra of each noise

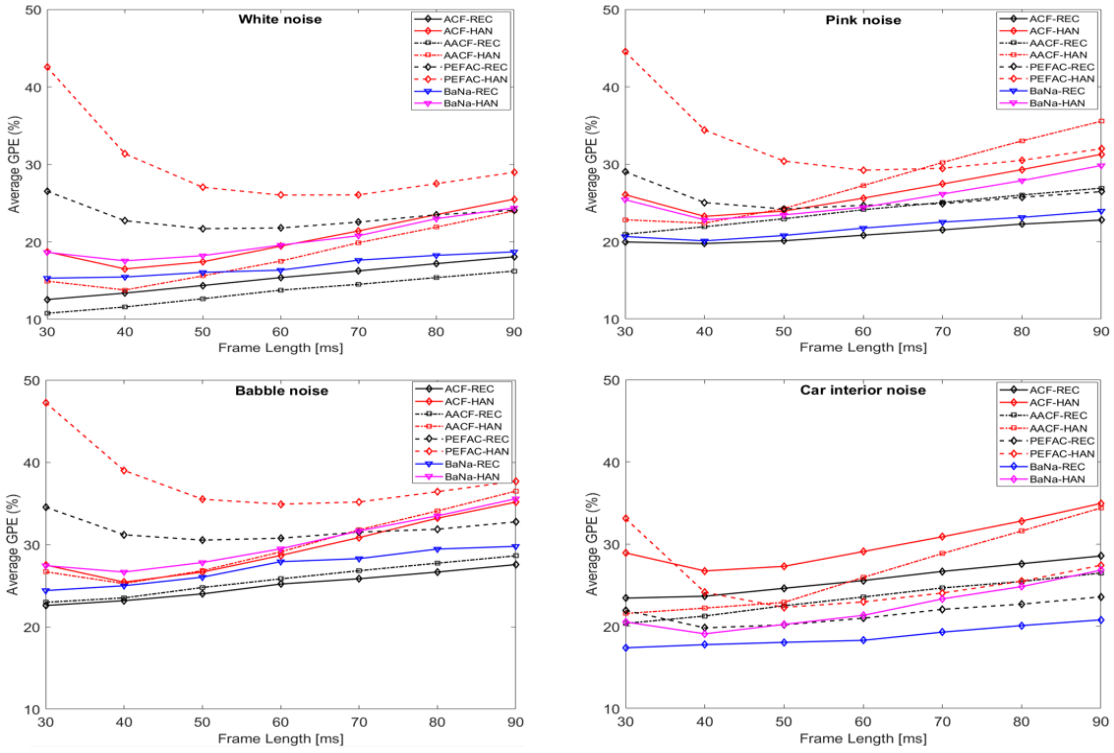

Figure 4.8: Frame length dependency for different types of noise at +5 dB SNR

approach in which shorter subframes are employed, the accumulation technique

becomes robust against babble noise, especially in high noise level cases. For randomly wide-band noises such as white and HF channel, the combination of the accumulated power spectrum and accumulated ACF works effectively in low SNR cases less than 0 [dB]. The combination approach behaves comparatively better in train and babble noises as well. If we have a priori information about the noise type and SNR condition or we can estimate them, then it will be possible to select one adequate pitch detection method among the three proposed accumulation based approaches.

The accumulated approaches disclose the potential to provide more excellent performance than two state-of-the art methods; PEFAC and BaNa. The important point is that a post processing is not involved in the accumulation approaches, which obviously saves the computation time for pitch detection.

Chapter 5

Conclusion and Future Work

This chapter concludes the thesis with a summary of our work. The brief discussion of the future work is also stated in this chapter.

5.1 Summary of the Work

Pitch is an important attribute of human speech, which is originated from the vibration of vocal folds. Pitch period extraction is a key technique to understand most acoustical phenomena in speech communication and plays an important role in speech processing applications such as speech coding, speech recognition, speech enhancement, speech synthesis and so on. The performance of these systems is significantly affected by the accuracy of pitch or fundamental frequency extraction. Reliability and accuracy of pitch extraction algorithms face real challenges, when the speech signal is corrupted by noise.

A few works have been done on the pitch extraction in noisy environments. In noisy environments, the conventional techniques utilize the time consuming post-processing on the harmonics in the frequency domain to detect the more appropriate pitch candidates. These methods also used the large number of DFT (IDFT) size and larger frame length for creating the narrowing pitch peaks. Therefore, these methods are effective at low SNRs in different noise cases which is emphasized in their original paper. However, this is not satisfied at all SNRs in all noise cases. Also, the conventional techniques provide the large computational time which is not suitable for real world applications.

Many pitch extraction methods have been developed in the past because it can be used in real world applications. Nowadays, it has been highly accurate even

in a noisy environment. Therefore, extraction results have come to be demanded. However, since the speech signal is originally time-varying. Thus, pitch extraction accuracy is not satisfied by the effect of noise characteristics. Therefore, the pitch extraction accuracy is further deteriorated. However, we have made a commitment to extract the pitch in noisy environments. The goal of this thesis is to develop some methods to extract the more accurate true pitch peak from the noisy speech signal without any complicated post-processing, which are more convenient and efficient in the real world applications. These proposed pitch extraction methods are performed in batch processing for speech waveforms by dividing into a short time frame. Then, pitch extraction is calculated, and it is performed into the next short time frame by utilizing the batch processing for the speech waveform. In this way, frame processing is accomplished. By this way, the extraction result is obtained for each frame in a short time.

In the first method, we proposed efficient and unique techniques such as FROOT+ and FROOT methods to extract the accurate peak location. Both methods are utilized on fourth-root spectrum. The FROOT+ method reduces the affect of vocal tract characteristics as well as to suppress the non-pitch peaks in the frequency domain, resulting in enhancing the pitch peak in the wide-band noise. On the other hand, the FROOT method is strongly robust against the narrow-band noise. As a result, the pitch extraction accuracy is significantly improved and the computational time is reduced.

After utilizing the first proposed work, we have investigated that, this method used the Hanning window function for generating the framed speech signal. Even, the state-of-the art methods are utilized these window function and also used larger frame length for creating narrowing pitch peaks to extract the accurate pitch candidates. In noisy environments, these window based methods show the pitch detection errors, because wider bandwidth of window function badly behaves due to the corrupted noise.

To improve the pitch extraction accuracy, the second proposed work emphasize to get the narrowing peaks of speech harmonics by utilizing the Rectangular window with standard frame length instead of the Hanning or Hamming window. Based on the Rectangular window function, the three accumulation based approaches show their potentiality for pitch detection without relying on a complicated post-processing. The processing time of the accumulation based approaches have been also reduced significantly.

Real world speech communication and speech processing applications such as

coding, recognition, enhancement, synthesis, require some methods for fast and efficient extraction of the pitch. The first proposed method in this thesis has shown to be capable to extract the pitch very fast, improves the performance of the pitch extraction algorithm and provide excellent extraction accuracy. The second proposed method designs a technique which utilizes the usefulness of Rectangular window and accumulation based approaches to tackle the problem that arises in the case of Hanning/Hamming window based methods. Also, this method has shown better performance and sufficient extraction accuracy which also can be applied in various pitch extraction applications.

5.2 Future Work

It has been shown that the proposed methods can accurately extract pitch even under noisy environments. This important feature of the proposed methods create space for its wide applications in various speech processing tasks. Incorporation of pitch tracking data helps the extraction system to increase the extraction accuracy and provide almost clean speech. Therefore, a noise robust pitch extractor is in a great demand. Two proposed methods in this thesis can be potentially employed to extract more accurate pitch at low SNRs in noisy environments.

In our experiments, we have used eight types of noise which are represented as the real world noise. Real world noises are found very difficult to handle for pitch extraction compared to white noise. In these noise cases, the original papers of the state-of-the art methods are highly concentrated on low SNRs. The conventional methods are not satisfied for all noise cases. From this point of view, our proposed approaches will be highly efficient and effective without any complicated post-processing when they would be employed adequately knowing the information about the noise type and SNR degree. Thus, in future we will extend our research to develop new pitch extraction methods which will be particularly robust against very low SNR cases in every real world noise case. The developed methods could be employed to various applications of speech effectively.

Bibliography

[1] C. E. Shannon, A Mathematical Theory of Communication, Bell System Tech. J., vol. 27, pp. 623-656, 1968.

[2] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice Hall, New York, 1978.

[3] B. S. Atal, Automatic speaker recognition based on pitch contours, J. Acoust. Soc. Am., vol. 52, no. 6, pp. 1687-1697, 1972.

[4] B. S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am., vol. 55, pp. 1304-1312, 1974.

[5] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, Detection of baby voice and its application using speech recognition system and fundamental frequency analysis, Proc. 10th WSEAS Int. Conf. Applied Computer Science, pp. 341-345, 2010.

[6] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, A John Wiley and Sons, Ltd., Publication, London, UK, 2008.

[7] M. Christensen and A. Jakobsson, Multi-pitch Estimation, Morgan and Claypool, USA, 2009.

[8] S. Furui, Research on individuality features in speech waves and automatic speaker recognition techniques, Speech Communication, vol. 5, no. 2, pp. 183-197, 1986.

[9] H. Beigi, Fundamental of Speaker Recognition, Springer, New York, 2011.

[10] A. E. Rosenberg and M. R. Sambur, New techniques for automatic speaker verification, IEEE Trans., Acoust., Speech, and Signal Process., vol. 23, no. 2, pp. 169-176, 1975.

[11] I. R. Titze, Principles of Voice Production, National Center for Voice and Speech, Iowa City, USA, 2000.

[12] A. Tsanas, M. A. Little, P. E. McSharry and L. O. Ramig, Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity, J. Royal Soc. Interface, vol. 8, pp. 842-855, 2011.

[13] C. Llerena, L. Alvarez and D. Ayllon, Pitch detection in pathological voices driven by three tailored classical pitch detection algorithms, Proc. 11th WSEAS Int. Conf. Signal Processing, Computational Geometry and Artificial Vision, pp. 113-118, 2011.

[14] A. S. Spanias, Speech coding: A tutorial review, Proc. of the IEEE, vol. 82, pp. 1541-1582, 1994.

[15] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, 2002.

[16] P.Vary and R. Martin, Digital Speech Transmission, John Wiley Sons Ltd, 2006.

[17] J. L. Flanagan, Speech Analysis, Synthesis, and Perceptions, 2nd ed. New York: Springer-Verlag, 1976.

[18] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, A comparative performance study of several pitch detection algorithms, IEEE Trans. Acoust., Speech, and Signal Process., vol. 24, no. 5, pp. 399-418, 1976.

[19] P. Veprek and M. S. Scordilis, Analysis, enhancement and evaluation of five pitch determination techniques, Speech Communication, vol. 37, pp. 249-270, 2002.

[20] M. M. Sondhi, New methods of pitch extraction, IEEE Trans. Audio Electro. acoust., vol. 16, pp. 262-266, 1968.

[21] D. M. Howard, Peak-picking fundamental period estimation for hearing prostheses, J. Acoust. Soc. Am., vol. 86, pp. 902-910, 1989.

[22] B. Gold and L. R. Rabiner, Parallel processing techniques for estimating pitch periods of speech in the time domain, J. Acoust. Soc. Amer., vol. 46, pp. 442-448, 1969.

[23] N. J. Miller, Pitch detection by data reduction, IEEE Trans. Acoust., Speech, Signal Process., vol. 23, pp. 72-79, 1975.

[24] L. R. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. Acoust., Speech, Signal Process., vol. 25, no. 1, pp. 24-33, 1977.

[25] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, Average magnitude difference function pitch extractor, IEEE Trans. Acoust., Speech, Signal Process., vol. 22, pp. 353-362, 1974.

[26] T. Shimamura and H. Kobayashi, Weighted autocorrelation for pitch extraction of noisy speech, IEEE Trans. Speech and Audio Process., vol. 9, no. 7, pp. 727-730, 2001.

[27] A. Cheveigne and H. Kawahara, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am., vol. 111, no. 4, pp. 1917-1930, 2002.

[28] C. K. Un and S.C. Yang, A pitch extraction algorithm based on LPC inverse filtering and AMDF, IEEE Trans. Acoust., Speech, Signal Process., vol. 25, pp. 526-572, 1977.

[29] A. M. Noll, Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate, in Proc. symp. comput. process., commun. (Brooklyn ,NY), pp. 779-797, 1969.

[30] A. M. Noll, Cepstrum pitch determination, J. Acoust. Soc. Am., vol. 41, no. 2, pp. 293-309, 1967.

[31] S. Ahmadi and A. S. Spanias, Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, IEEE Trans. Speech and Audio Process., vol. 7, no. 3, pp. 333-338, 1999.

[32] H. Kobayashi and T. Shimamura, A modified cepstrum method for pitch extraction, Proc. IEEE Asia-Pacific Int. Conf. Circuits and Systems Microelectronics and Integrating Systems (APCCAS), 1998.

[33] M. A. F. M. R. Hasan, M. S. Rahman, T. Shimamura, Windowless-autocorrelation-based cepstrum method for pitch extraction of noisy speech, J. Signal, Process., pp. 231-239, 2012.

[34] S. Seneff, Real-time harmonic pitch detector, IEEE Trans. Acoust., Speech, Signal Process., vol. 26, pp. 358-365, 1978.

[35] T.V. Sreenivas and P.V.S. Rao, pitch extraction from corrupted harmonics of the power spectrum, J. Acoust. Soc. Am., vol. 65, pp. 223-228, 1979.

[36] M. Lahat, R. J. Niederjohn and D. A. Krubsack, A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech, IEEE Trans. Acoust., Speech, Signal Process., vol. 35, no. 6, pp. 741-750, 1987.

[37] J. Markel, The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio Electro. acoust., vol. 20, no. 5, pp. 367-377, 1972.

[38] N. Kunieda, T. Shimamura and J. Suzuki, Pitch extraction by using auto-correlation function on the log spectrum, IEICE Trans. vol. 3, pp. 435-443, 1997.

[39] W. Hess, Pitch Determination of Speech Signals, Springer-Verlag, 1983.

[40] J. W. Xu and J. C. Principle, A pitch detector based on a generalized correlation function, IEEE Trans. Audio Speech and Lang. Process., vol. 16, no. 8, pp. 1420-1432, 2008.

[41] A. Moreno and J. A. R. Fonollosa, Pitch determination of noisy speech using higher order statistics, Proc. IEEE Int. Conf. Acoust. Speech and Signal Process., (ICASSP), 1992.

[42] C. Shahnaz, W. Zhu and M. O. Ahmad, Pitch estimation based on a harmonic sinusoidal autocorrelation model and a time-domain matching scheme, IEEE Trans. Audio Speech and Lang. Process., vol. 20, no. 1, pp. 322-335, 2012.

[43] S. K. Roy, M. K. I. Molla, K. Hirose and M. K. Hasan, Harmonic modification and data adaptive filtering based approach to robust pitch estimation, Int. J. Speech Tech., vol. 14, no. 4, pp. 339-349, 2011.

[44] D. Talkin, A robust algorithm for pitch tracking (RAPT), in speech coding and synthesis, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier pp. 496-518, 1995.

[45] S. Lin, Robust pitch estimation and tracking for speakers based on sub-band encoding and the generalized labeled multi-Bernoulli filter, IEEE Trans. Speech and Lang. Process., vol. 27, no. 4, pp. 827-841, 2019.

[46] S. Gonzalez and M. Brookes, PEFAC - A pitch estimation algorithm robust to high levels of noise, IEEE/ACM Trans. Audio, Speech and Lang. Process., vol. 22, no. 2, pp. 518-530, 2014.

[47] 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 Trans. Audio Speech and Lang. Process., vol. 22, no. 12, pp. 1833-1848, 2014.

[48] D. J. Hermes, Measurement of pitch by subharmonic summation, J. Accoust. Soc. Am., vol. 83, no. 1, pp. 257-264, 1988.

[49] D. Wang, C. Yu and J. H. L. Hansen, Robust harmonic features for classification-based pitch estimation, IEEE/ACM Trans. Audio Speech and Lang. Process., vol. 25, no. 5, pp. 952-964, 2017.

[50] K. Han and D. Wang, Neural network based pitch tracking in very noisy speech, IEEE/ACM Trans. Audio Speech and Lang. Process., vol. 22, no. 12, pp. 2158-2168, 2014.

[51] S. Motegi and T. Shimamura, Fundamental frequency extraction in narrow band noise, Proc. Spring Conf., Acoust. Society of Jpn, pp. 277-278, 2011.

[52] “20 Countries Language Database,” NTT Advanced Technology Corp., Jpn, 1988.

[53] F. Plante, G. Meyer and W. Ainsworth, A fundamental frequency extraction reference database, Proc. Eurospeech, pp. 837-840, 1995.

[54] A. Varga and H. J Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech Communi., vol. 12, no. 3, pp. 247-251, 1993.

[55] S. Itahashi, Creating speech copora for speech science and technology, IEICE Trans. Funda. of Electroni., Communi, and Computer Scien., vol. E 74, no. 7, pp. 1906-1910, 1991.

[56] L. Sukhostat and Y. Imamverdiyev, A comparative analysis of pitch detection methods under the influence of different noise conditions, J. Voice, vol. 29, no. 4 pp. 410-417, 2015.

[57] Wcng, wireless communication networking group, [Online]. Available: http://www.ece.rochester.edu/projects/wcng/ code.html

[58] D. A. Krubsack, R. J. Niederjohn, An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech, IEEE Trans., Signal Process., vol. 39, no. 2, pp. 319-329, 1991.

[59] S. M. Kay, Modern Spectral Estimation; Theory and Application, Prentice Hall, 1988.

[60] L. R. Rabiner, B. Gold, Theory and Application of Digital Signal Process., Prentice Hall, 1975.

[61] M. Brookes, Voicebox toolkit, [Online]. Available, http://www.ee.ic.ac.uk /hp/staff/dmb /voicebox/voicebox.html

LIST OF PUBLICATIONS

Journal Articles

1. Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura, "Pitch Extraction using Fourth-root Spectrum in Noisy speech", *Journal of Signal Processing*, 17-pages (Accepted), 2020.

2. Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura, "Utilization of Windowing Effect and Accumulated Autocorrelation Function and Power Spectrum for Pitch Detection in Noisy Environments", *IEEJ Trans. Electrical and Electronic Engineering*, 9-pages (Accepted), 2020.

International Conference (Reviewed)

1. Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura, "A Multiple Functions Multiplication Approach for Pitch Extraction of Noisy Speech", *Int., Conf., on Speech Tech., and Human-Com., Dial., (SpeD)*, Oct. 2019.

2. Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura, "Refined Autocorrelation Function for Pitch Detection of Speech", *Int. Workshop on Smart Info-Media Systems in Asia (SISA 2019)*, pp. 72-77, Sep. 2019.

3. Md. Saifur Rahman, Yosuke Sugiura and Tetsuya Shimamura, "Pitch Determination of Noisy Speech Using Cumulant Based Modified Weighted Function ", *Proc. IEEE TENCON*, pp. 1474-1478, Oct. 2018.