NEW TIME-FREQUENCY DOMAIN PITCH ESTIMATION METHODS FOR SPEECH SIGNALS UNDER LOW LEVELS OF SNR

Celia Shahnaz

A thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Concordia University
Montréal, Québec, Canada

September 2009

Your file Votre référence
ISBN: 978-0-494-63361-8
Our file Notre référence
ISBN: 978-0-494-63361-8

Abstract

New Time-Frequency Domain Pitch Estimation Methods for Speech Signals under Low Levels of SNR

Celia Shahnaz, Ph.D.

Concordia University, 2009

Pitch estimation of speech signals is the key to understanding most acoustical phenomena as well as accurately designing many practical systems in speech communication. It is to determine the fundamental frequency or period of a vocal cord vibration causing periodicity in the speech signal. This task becomes very difficult when the speech observations are heavily corrupted by noise. Although a large number of pitch estimation methods have been reported to deal with a noise-free environment, pitch estimation in the presence of noise has been attempted only by a few researchers. As noise generally obscures the periodic structure of the speech waveforms, many existing methods fail to provide accurate pitch estimates when the signal-to-noise ratio (SNR) is very low. The major objective of this research is to develop novel pitch estimation methods capable of handling speech signals in practical situations where only noise-corrupted speech observations are available. With this objective in mind, the estimation task is carried out in two different approaches. In the first approach, the noisy speech observations are directly employed to develop two new time-frequency domain pitch estimation methods. These methods are based on extracting a pitch-harmonic and finding the corresponding harmonic number required for pitch estimation. Considering that voiced speech is the output of a vocal tract system driven by a sequence of pulses separated by the pitch period, in the second approach, instead of using the noisy speech directly for pitch estimation,

an excitation-like signal (ELS) is first generated from the noisy speech or its noise-reduced version. Time-domain as well as time-frequency domain functions of the ELS are then proposed for pitch estimation.

In the first approach, at first, a harmonic cosine autocorrelation (HCAC) model of clean speech in terms of its pitch-harmonics is introduced. In order to extract a pitch-harmonic, we propose an optimization technique based on least-squares fitting of the autocorrelation function (ACF) of the noisy speech to the HCAC model. By exploiting the extracted pitch-harmonic along with the fast Fourier transform (FFT) based power spectrum of noisy speech, we then deduce a harmonic measure and a harmonic-to-noise-power ratio (HNPR) to determine the desired harmonic number of the extracted pitch-harmonic. In the proposed optimization, an initial estimate of the pitch-harmonic is obtained from the maximum peak of the smoothed FFT power spectrum.

In addition to the HCAC model, where the cross-product terms of different harmonics are neglected, we derive a compact yet accurate harmonic sinusoidal autocorrelation (HSAC) model for clean speech signal. The new HSAC model is then used in the least-squares model-fitting optimization technique to extract a pitch-harmonic. A symmetric average magnitude sum function of the speech signal and an impulse-train with its period formed from the extracted pitch-harmonic are proposed to formulate an objective function for determining the desired harmonic number with respect to the pitch-harmonic. Due to the advantageous features of the discrete cosine transform (DCT) over the FFT in the case of real signals, a smoothed DCT power spectrum is adopted to obtain an initial estimate of the pitch-harmonic.

In the second approach, first, we develop a pitch estimation method by using an excitation-like signal (ELS) generated from the noisy speech. To this end, a technique

based on the principle of homomorphic deconvolution is proposed for extracting the vocal-tract system (VTS) parameters from the noisy speech, which are utilized to perform an inverse-filtering of the noisy speech to produce a residual signal (RS). In order to reduce the effect of noise on the RS, a noise-compensation scheme is introduced in the autocorrelation domain. The noise-compensated ACF of the RS is then employed to generate a squared Hilbert envelope (SHE) as the ELS of the voiced speech. With a view to further overcome the adverse effect of noise on the ELS, a new symmetric normalized magnitude difference function of the ELS is proposed for eventual pitch estimation.

Cepstrum has been widely used in speech signal processing but has limited capability of handling noise. One potential solution could be the introduction of a noise reduction block prior to pitch estimation based on the conventional cepstrum, a framework already available in many practical applications, such as mobile communication and hearing aids. Motivated by the advantages of the existing framework and considering the superiority of our ELS to the speech itself in providing clues for pitch information, we develop a cepstrum-based pitch estimation method by using the ELS obtained from the noise-reduced speech. For this purpose, we propose a noise subtraction scheme in frequency domain, which takes into account the possible cross-correlation between speech and noise and has advantages of noise being updated with time and adjusted at each frame. The enhanced speech thus obtained is utilized to extract the vocal-tract system (VTS) parameters via the homomorphic deconvolution technique. A residual signal (RS) is then produced by inverse-filtering the enhanced speech with the extracted VTS parameters. It is found that, unlike the previous ELS-based method, the squared Hilbert envelope (SHE) computed from the RS of the enhanced speech without noise compensation, is sufficient to represent

an ELS. Finally, in order to tackle the undesirable effect of noise of the ELS at a very low SNR and overcome the limitation of the conventional cepstrum in handling different types of noises, a time-frequency domain pseudo cepstrum of the ELS of the enhanced speech, incorporating information of both magnitude and phase spectra of the ELS, is proposed for pitch estimation.

A pitch tracking scheme using a dynamic programming is employed, whenever applicable, to obtain a smoothed pitch contour for practical applications involving severely noise-corrupted speech. In order to study the effectiveness of the proposed pitch estimation methods, extensive simulations are carried out by considering naturally spoken speech signals in the presence of white or multi-talker babble noise at different SNR levels. A comprehensive evaluation of the pitch estimation results using different performance metrics demonstrates the significant superiority of the proposed methods over some of the state-of-the-art methods under low levels of SNR.

Dedication

To my beloved husband Shaikh Anowarul Fattah

Acknowledgments

I would like to express my profound gratitude and indebtedness to my supervisors Dr. M. Omair Ahmad and Dr. Weiping Zhu for their insightful guidance and suggestions during my Ph.D. program. I am grateful to them for providing me the freedom to explore new ideas in a challenging research field. I also want to thank them for spending time with me in improving the presentation of this thesis. The useful suggestions provided by the committee members are deeply appreciated. Especially, I want to thank Dr. M. N. S. Swamy for his comments and encouragement for this research from the very beginning of my Ph.D. program. I feel myself fortunate to receive detailed corrections and constructive suggestions regarding my thesis from Dr. Douglas O'Shaughnessy, a world-renowned speech researcher. His cordial appreciation for my work will act as a source of encouragement to continue innovative research in this field.

I am grateful to Canadian Government for providing me the financial support under the prestigious Commonwealth Scholarship and Fellowship Program administered by the Canadian Bureau for International Education (CBIE), which was crucial for completing this research. I also acknowledge the financial support provided by Concordia University and NSERC, Canada.

I wish to thank Dr. V. D. Ramachandran for his continuous inspiration. I truly acknowledge the friendly cooperation of my colleagues in the Center for Signal Process-

ing and Communications Dr. I. H. Bhuiyan, Dr. Awni Itradat, Dr. Chao Wu, and Dr. Saad Boguezel.

Special note of thanks goes to my husband, Dr. Shaikh Anowarul Fattah, for his constant moral support, caring inspiration, and thoughtful discussions.

Finally, I would like to express my gratefulness to my beloved parents and brother for their endless prayers, patience, love, and constant encouragement. I am also grateful to my in-laws for their foresight and kind cooperation.

Table of Contents

List of Figures

List of Acronyms

ACF	:	Autocorrelation function
AMDF	:	Average magnitude difference function
ASR	:	Automatic speech recognition
CEP	:	Cepstrum
DCT	:	Discrete cosine transform
DLFT	:	Discrete logarithmic Fourier transform
DP	:	Dynamic programming
ELS	:	Excitation-like signal
FFT	:	Fast Fourier transform
FPE	:	Fine pitch error
GC	:	Glottal Closure
GPE	:	Gross pitch error
HCAC	:	Harmonic cosine autocorrelation
HSAC	:	Harmonic sinusoidal autocorrelation
HNM	:	Harmonic number matching
HNPR	:	Harmonic-to-noise power ratio
IDCT	:	Inverse discrete cosine transform

IFFT	:	Inverse fast Fourier transform
LP	:	Linear prediction
LPF	:	Low-pass filter
MSE	:	Mean-squared-error
NRCEP	:	Noise reduced CEP
NRRES	:	Noise reduced RES
PCELS	:	Pseudo-cepstrum of the ELS
PGPE	:	Percentage of GPE
PH	:	Pitch-harmonic
PPROC	:	Parallel processing method
RES	:	Residual
RMSE	:	Root-mean-square-error
RS	:	Residual signal
SHE	:	Squared Hilbert envelope
SIFT	:	Simple inverse filter tracking
SIM	:	SAMSF based impulse-train matching
SAMSF	:	Symmetric average magnitude sum function
SNMDF	:	Symmetric normalized magnitude difference function

SNR	:	Signal-to-noise ratio
STA	:	Short-time analysis
TD	:	Time-domain
VAD	:	Voice activity detector
VTS	:	Vocal tract system
WAUTOC	:	Weighted autocorrelation

Chapter 1

Introduction

1.1 General

1.1.1 Pitch Estimation: Background

Speech signals are fairly stationary when observed over a sufficiently short period of time (typically, $5 \sim 100$ msec). However, over a long period of time (of the order of 1/5 second or more), the speech characteristics change with time to reflect different speech sounds being spoken. According to the speech production model as shown in Fig. 1.1, the production of speech can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The source can be periodic or aperiodic like noise, resulting in voiced speech or unvoiced speech. It is an accepted convention to use the following three-state representation:

1. Silence, where no speech is produced.

2. Voiced speech like "*ee*" in the word "*speech*". The voicing source occurs in the larynx, at the base of the vocal tract, where airflow from the lungs up to the glottis can be interrupted periodically by the vibrating vocal folds. The pulses of air produced by the abduction and adduction of the folds generate a periodic excitation for the vocal tract. Note that a truly periodic signal should

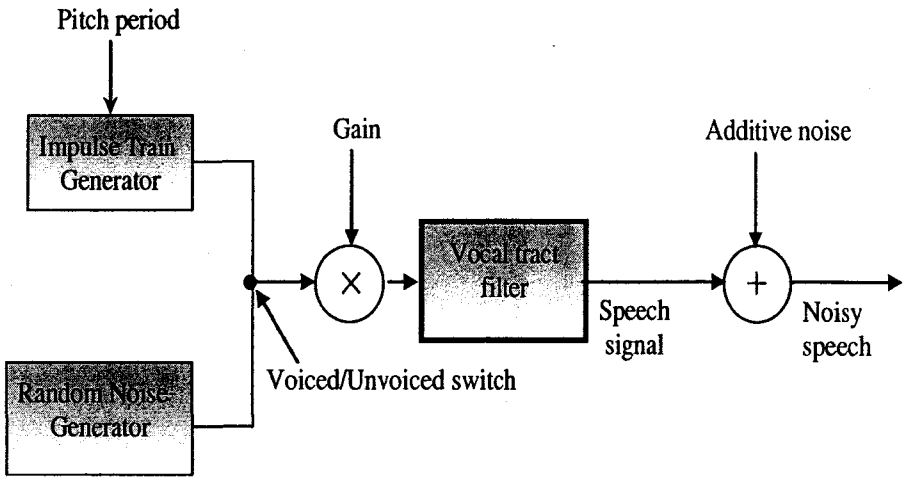


Figure 1.1: Speech production model.

have a discrete-line spectrum, but since the vocal tract changes its shape almost continuously, the output voiced speech is almost periodic.

3. Unvoiced speech like "*p*" in the word "*pat*" in which the vocal folds are not vibrating.

The vocal tract system can be modeled as an acoustic tube with resonances and antiresonances. The resonance frequencies are determined by system parameters representing the formants of the speech signal. For both voiced and unvoiced excitation, the vocal tract, acting as a filter, amplifies certain sound frequencies and energy around formant frequencies, while attenuating energy around antiresonant frequencies between the formants. As a periodic signal, voiced speech has spectra consisting of harmonics of the fundamental frequency of the vocal fold vibration. This frequency, often abbreviated $F_0 = 1/T_0$, is the physical aspect of speech corresponding to the perceived pitch. Thus the term pitch is often used interchangeably with fundamental frequency. Pitch estimation is to aim at determining the fundamental frequency or

period (pitch period) of the vocal fold vibration causing periodicity in the speech signal [1], [2]. For unvoiced speech, as vocal folds are not vibrating, pitch frequency is by convention $F_0 \equiv 0$. The pitch period is dependent on the size and tension of the speaker's vocal folds at any given instant. Since the average size of the vocal folds for men is larger than that for women, the average pitch of an adult male will often be lower than that of a female for a given utterance. The possible pitch range is usually $50 - 250$ Hz and $120 - 500$ Hz for men and women, respectively. Pitch changes in response to stress, intonation, and emotion.

The problem of pitch estimation has received immense interest from different research fields involving speech communication [3]–[13]. Besides providing valuable clues for the excitation source in speech production, the pitch contour of an utterance is useful in speaker recognition [4], [5]. Accurate pitch information is required to reconstruct good-quality speech and to reduce transmission rate in low- and medium-rate voice coders. It is required in almost all speech analysis-synthesis systems [6], such as multiband excitation vocoder [7]. Pitch is also the prime acoustic cue to intonation and stress of speech, and is vital to phoneme identification in tonal languages [8], [9]. Combined with other acoustic features, pitch as a prosody information can be used to improve speech recognition performance [10]. Since pitch variation with time gives the information of accent and intonation, pitch is useful in devices for speech instruction to the hearing impaired learning to speak [11]–[13], and for foreign language training. Particularly in speech enhancement systems, the accuracy of pitch estimation is directly related with the quality of the enhanced speech. Therefore, the pitch estimation has axiomatic importance in advanced audio and speech applications.

1.1.2 Problems in Pitch Estimation

In this subsection, different problems in pitch estimation and their effects on pitch estimation method will be briefly addressed. Accurate and reliable estimation of pitch period from the speech observations alone is often extremely difficult for several reasons.

1. The glottal excitation waveform for voiced speech may not be a perfect train of periodic pulses. Although finding the period of a perfectly periodic waveform is straightforward, measuring the period of a speech signal, which varies both in period and in the detailed structure of the waveform within a period, can be quite a tough task.

2. The difficulty in measuring pitch period is the interaction between the vocal tract and the glottal excitation. Sometimes, the formants of the vocal tract can alter significantly the structure of the glottal excitation signal so that the actual pitch period is hard to determine. Such interactions generally are most deleterious to pitch estimation during rapid movements of the articulators when the formants are also changing rapidly.

3. There is an inherent difficulty in defining the exact beginning and ending of each pitch period during voiced speech segments. Based on the acoustic waveform alone, some features for defining the beginning and ending of the period include the maximum value during the period, the zero crossing prior to the maximum, etc. Peak measurements are sensitive to the formant structure during the pitch period, whereas zero crossings of a waveform are sensitive to the formants, noise, and any dc level in the waveform. The only requirement on such measurements

is that the waveform should be consistent from one period to another in order to be able to define the "exact" location of the beginning and ending of each pitch period. The lack of such consistency can lead to erroneous pitch period estimates.

4. It is extremely hard to determine pitch for low-level voiced speech segments. At voiced-unvoiced boundaries, the expected continuity feature becomes less useful and pitch periods are often irregular.

In the real world, as speech is normally processed in a noisy environment, the obvious inclusion of some background noise is unavoidable. Reliability and accuracy of pitch estimation methods face the real challenge when noise corrupts the speech. Particularly, when the level of noise is low, pitch estimation becomes rather difficult due to the effect of severe noise. At a very low SNR, the overall effect of noise is to obscure the periodic structure of the speech waveform. But as far as real-life applications are concerned, a pitch estimation task has to be performed using only the given noise-corrupted speech observations. Thus, due to the growing demand on speech applications operating in noisy environments, pitch estimation from a very severe noisy observation is an open problem that has drawn many researchers' attention.

1.2 A Review of Existing Pitch Estimation Methods

In this section, we will review some of the typical pitch estimation methods. This literature review not only serves as a necessary background for understanding the state-of-the-art methods, but also supports the motivation behind the research work of the thesis.

Various pitch estimation methods have been developed to overcome the difficulties mentioned in the previous section and a comprehensive review of these methods can be found in [1], [14]–[16]. In general, pitch estimation methods can be classified into two broad categories [1], [16]: time-domain (TD) methods, e.g., [17]–[20] and short-time analysis (STA) based methods, e.g., [2], [21]–[26]. The former attempts to identify one or more features, such as the fundamental harmonic, a quasi-periodic time structure, an alternation of high and low amplitudes, or points of discontinuities in the speech waveform, and then to find pitch markers or epochs in a pitch synchronous manner. It is to be noted that the TD methods operate directly on the speech waveform and thus depend strongly on the shape of the waveform and therefore, they are prone to pitch errors [1], [27]. On the other hand, the STA based methods transform a short-time frame of speech samples into an alternate domain in order to enhance the periodicity information contained in the speech. These methods determine an average fundamental frequency from several contiguous periods over an analysis frame and normally yield more reliable pitch estimates than that achieved by a TD method [1], [27].

The STA for pitch estimation usually involves autocorrelation function (ACF), average magnitude difference function (AMDF), harmonic matching, spectral compression, maximum likelihood, cepstrum and other variations. Among them, the ACF and the AMDF have been commonly employed for pitch estimation [1], [14]. In what follows, we will review some of these methods of pitch estimation for both clean and noisy environments.

1.2.1 Pitch Estimation Methods for Clean Speech

The ACF has peaks or maxima at integral multiples of the pitch period. In general, for a speech frame, a pitch estimation method employing ACF [28]–[31] determines the highest non-zero-lag peak from an exhaustive search of the peaks. There exist several problems associated with the use of the ACF for pitch estimation, such as the choice of window for short-time analysis, selection of suitable frame size, presence of noise, and correlation peaks due to formant structure. To partially eliminate the effects of the second and higher formants, speech is normally low-pass filtered to retain a frequency range of $0 - 900$ Hz [15]. In addition to low-pass filtering, center clipping [32] and infinite peak clipping are used before computing the ACF in [28]. Center clipping sets zeros to low-amplitude speech samples and reduces the magnitude of high-amplitude samples, whereas infinite peak clipping reduces speech to a zero-crossing signal. In [29], properties of a class of nonlinearities applied to the speech signal prior to the ACF analysis are investigated for pitch estimation. It is found that the nonlinearities, such as compressed center clipping, simple center clipping, and the combination of center and peak clipping provide some degree of spectral flattening, thereby enhancing the periodicity peaks in the ACF, and reducing the ACF peaks due to formants. Also, a solution to the problem of choosing an analysis frame size, which adapts to the estimated average pitch of the speakers, was proposed in [29]. In [30], the computed narrowed ACF obtains sharper peaks using multiple frames. In order to further reduce pitch estimation errors in clean speech, in [31] an estimator that is based on the well-known ACF, known as YIN, is proposed. In this method, the error mechanisms of the ACF are analyzed and then a series of modifications, e.g., a difference function formulation, normalization and parabolic interpolation has

been introduced to decrease the error rates.

The AMDF is a variation of the ACF analysis, where instead of correlating the input speech samples at each lag, a difference signal is formed between the shifted and the original speech, and the absolute magnitude of such a difference is summed over the frames. Thus, the AMDF exhibits deep notches or minima at integral multiples of the pitch period. The separation of notches is a direct measure of the pitch period. In general, for a speech frame, a pitch estimation method employing AMDF finds the pitch either directly from the global minimum [33] or from the distribution of the minima of the AMDF [34]. In [35], a binary version of AMDF is proposed, where the AMDF values for a frame of clean speech are first thresholded to a single bit (0 or 1) via a clipping. Then, the pitch period is calculated using the ACF of the resulting single-bit AMDF. Even though the clipping helps in removing the formant structure, the fixed clipping threshold, which is not adaptable to speech intensities of different speakers, may result in an erroneous one-bit AMDF sequence, thus leading to a higher pitch-error. Also, the search of the pitch-peak does not cover the wide range ($50 \sim 500$ Hz) of pitch.

Some STA based pitch estimation methods use the property that if the signal is periodic in the time-domain, the frequency spectrum of the signal consists of a series of impulses at the fundamental frequency and its harmonics [22], [26], [36], [37]. Thus simple measurements can be made on the frequency spectrum of the signal (or a nonlinearly transformed version of it) to estimate signal periodicity. Instead of directly using the fundamental spectral peak, the harmonic structure consisting of spectral peaks at multiples of pitch is more reliable for pitch estimation. Although the frequency of the greatest common divisor of the harmonics often provides a good pitch estimate for female speech with widely spaced harmonics, the reliability of a

pitch estimate for male speech cannot be guaranteed. In [22], a histogram is built in which, for every harmonic frequency present, a counter is incremented for the probable fundamental frequency. The method can be further improved by assigning weights in the histogram, which takes representative amplitude into account to decide the harmonics. In [38], a spectral score function (discrete logarithmic Fourier transform, DLFT) using "template frame" and "cross-frame", spectral correlation functions is developed. The DLFT based pitch estimation methods generally perform better for female speech than for male speech. When pitch is low, the template-frame correlation suffers from missing low harmonics, and the cross-frame correlation suffers from compact spacing of higher order harmonics. This can potentially be improved by using gender-dependent parameters, or by adaptive signal processing, such as a variable frequency range for the DLFT.

Pitch can be estimated using the maximum-likelihood (ML) method [39] and the sinusoidal speech model based approach [24]. In [24], by exploiting a sinusoidal speech model for the input speech waveform, a pitch estimation method is developed by fitting a harmonic set of sine waves to the input data based on a mean-squared-error (MSE) criterion.

Some pitch estimators make use of both time-domain and frequency-domain characteristics of speech signals [21], [23], [40]. Among them, cepstrum (CEP) is a well known pitch estimator for a noise-free environment as proposed in [21], where the cepstrum of a speech frame is computed and the location of the peak cepstral value is used to determine the pitch period. But the CEP approach suffers from practical difficulties due to size, shape, and placement of the window creating the analysis frame. If the pitch period is quite long and the window size is chosen so that only one period or less appears in a given frame, the periodic component of the cepstrum will no

longer consist of a pulse train. Another critical factor is the formant structure of the vocal tract system (VTS). If the VTS is essentially a narrow band filter, the periodic component of the spectrum will be masked by the formant filter and no peaks will occur in the cepstrum [41]. Although various algorithmic modifications have been undertaken in [42] to overcome such difficulties, the pitch estimation performance of high-pitched speakers still remains unsatisfactory.

As an alternative to clipping, some researchers attempted to remove the formants by passing the speech through an inverse filter, whose parameters are derived from the linear prediction (LP) analysis. The inverse filter essentially yields a spectrally flattened signal referred to as the LP residual, which is expected to resemble the excitation signal itself. The ACF or AMDF of the LP residual is then used to determine the pitch period [23], [40]. In [23], a so-called simple inverse filter tracking (SIFT) method is developed, where the pitch period is obtained by interpolating the ACF in the neighbourhood of the peak of the ACF. The SIFT [23] and the LP inverse filtering based AMDF [40] suffers from the less reliable spectral flattening (by inverse filtering) of the speech signal, especially for high-pitched speakers in which often only one harmonic occurs, thus decreasing the pitch estimation accuracy for such speakers.

1.2.2 Pitch Estimation Methods for Noisy Speech

A pitch estimation method that exhibits good performance for clean speech cannot be guaranteed to work well for noisy speech. Although a large number of pitch estimation methods have been reported in the literature for clean speech, pitch estimation from noisy speech has been attempted only by a few researchers. However, many applications require robust pitch estimation from noisy speech, which is an extremely difficult task, since the noise obscures the periodic structure of the speech waveforms

thus causing a significant performance degradation of the pitch estimation methods.

In the TD data reduction method (DARD) [17], the estimation of correct pitch period is troublesome for low level voiced speech corrupted by noise. Also, in the parallel processing method (PPROC) [18], which belongs to the TD pitch estimator, the amplitudes of peaks and valleys of speech are altered due to noise and cannot be detected perfectly. These two TD methods of pitch estimation have somewhat lower resolution due to the sensitivity of waveform peaks, valleys, and zero crossings to formant changes, noise, distortion, etc.

In the STA based CEP method [21], noise affects the cepstrum in the sense that it masks the harmonic structure between formant peaks, obscures part of the periodic component of the log spectrum, and reduces the amplitude of the corresponding cepstral peak and thus makes pitch estimation difficult. In the presence of noise, since the vocal tract cannot be accurately modeled and the estimation accuracy of the parameters of the inverse filter decreases, the discrepancy between the residual signal and the presumed impulse-train excitation increases. As a result, the SIFT [23] and the LP inverse filtering based AMDF [40] suffer from inaccuracies, such as pitch doubling and jitter, and yield poor performance for pitch estimation in noise.

It has been reported in [33] that in the presence of additive noise, since the deep notches of AMDF decrease with increasing time lag, the AMDF based approach for pitch estimation causes pitch doubling errors, which occur at the onset or central portion of a voiced segment. The errors are mainly due to the fact that the global minimum of AMDF is severely affected due to intensity variation and the background noise of the speech signal [43]. In [34], periodicity-aperiodicity measures obtained from an AMDF are utilized to estimate pitch. Modifications are proposed to determine pitch in weakly periodic regions and in some transition regions between

voiced and unvoiced sounds. In order to overcome pitch halving and doubling errors, a hard-threshold-dependent rectification criterion is also suggested but pitch estimation performance is not reported for noise-corrupted speech. It is to be mentioned that most of the pitch estimation methods [26], [44]–[46] estimate pitch from speech observations corrupted by white noise-only.

Among various STA based pitch estimation methods, ACF based approaches are very popular for their simplicity and better performance in noise [14], [44]. Generally, the higher female pitch causes fewer harmonics in the first formant range. Most of the energy of voiced speech is concentrated at these few widely spaced harmonics, which are not easily affected by additive white noise. For low-pitched male speakers, the speech energy is spread over many harmonics, which are more easily affected by noise since they are lower in amplitude for a given total speech energy. That is why the ACF based pitch estimation method in [45] performs better for female speakers than that for male speakers.

Though introduction of center clipping is found to help remove the formant structure for better pitch estimation in clean speech [29], it may hamper pitch estimation in a heavy noisy condition [1]. The clipping methods reported in [29] are implemented in the ACF based method in [45]. But the clipping levels are influenced by the noise, causing more noise components than speech ones to remain after clipping. In terms of spectral flattening, since a high level of white noise implies a flat spectrum, clipping as a spectral flattener reduces the speech components that are not flat. As a result, a reduction of pitch errors up to 50% is seen for low levels of noise, but for high levels of noise (less than 0 dB), the results degraded.

In [46], utilizing the periodicity property of AMDF, the ACF is weighted by the reciprocal of the AMDF in order to emphasize the true pitch-peak for noisy speech.

Since, under a heavy noisy condition, the global maximum of ACF or the global minimum of AMDF may occur at a lag that is a multiple or submultiple of the true pitch period, in the reciprocal AMDF-weighted ACF, the peaks at non-pitch locations may be wrongly emphasized more than those at the true pitch location. This causes inaccurate pitch estimation, especially at a low SNR.

It is worth mentioning that all the methods mentioned above would not perform well for the estimation of pitch from speech that is corrupted by a real-world noise, such as multi-talker babble noise. There are only a few methods [47], [48] that deal with the problem of pitch estimation of real-world noise-corrupted speech. In [47], approximate information of the Glottal Closure (GC) instants has been exploited to extract pitch and performance of this method is evaluated by considering a speech utterance by a female speaker only. It has been confirmed by the authors that their method provides better performance compared to the SIFT method at a low SNR value like 3 dB and shows a graceful degradation at lower values of SNR such as 0 dB. Their results call for a better method to be developed for the accurate determination of the GC instants for using them in the extraction of pitch from a severely degraded speech. In [48], the dominance spectrum as well as ripple-enhanced power spectrum has been utilized for pitch estimation. These two methods have been reported to perform better relative to the YIN method in clean and noisy conditions. It is inferred that methods in [47], [48] perform well at only an SNR level of 0 dB or above.

1.3 Motivation

It is noted that the estimation performance of most of the existing pitch estimation methods, whether the TD pitch estimator or STA based pitch estimator, deteriorate drastically in the presence of noise. The situation becomes worse when the SNR is

very low and only the noise-corrupted speech observations are available. Although numerous pitch estimation methods have been disclosed in the literature to handle clean speech, pitch estimation in noisy environments, especially when dealing with a real world noise, has been attempted only by a few researchers. Moreover, these methods provide satisfactory performance only at moderately low to high levels of SNR. But there has been a growing demand for practical applications in which pitch has to be estimated accurately from noisy speech at very low levels of SNR. Thus, there is an imperative need of making concentrated effort towards the development of efficient pitch estimation methods for both male and female speakers in practical situations where there exists noise and even the SNR is very low.

Considering that voiced speech is the output of a VTS driven by a sequence of pulses separated by the pitch period, our main idea is to extract, from given noise-corrupted speech observations, such a representation of a speech signal where the pitch periodicity is significantly pronounced, thereby facilitating the pitch estimation task. To achieve this target, we may either employ directly the noisy speech observations or generate an excitation-like signal (ELS) from the noisy speech or its noise-reduced version.

If a dominant pitch-harmonic (PH) can be extracted from given noisy speech observations, a periodicity matching technique could be established either in the frequency domain or in the time domain. In the extraction of a PH using a conventional least-squares minimization technique, there is a problem that the true data is not available in practice. Thus, an appropriate model, if it exists, to represent the true data will facilitate the least-squares estimation problem. This idea motivates us to develop noise-robust models in the correlation domain and design least-squares model-fitting based PH extraction techniques. In order to obtain the desired harmonic number

corresponding to the extracted PH, a harmonic number matching technique in the frequency domain, and an impulse-train matching technique in the time domain could be developed. However, in order to handle severe noisy conditions, in frequency domain matching, new harmonic measures are required, while in time domain matching, it is necessary to develop a new noise-robust time-domain function retaining the periodic structure of speech.

In the conventional residual signal (RS) based pitch estimation methods, the RS is first generated using the estimated VTS parameters. These methods have a common shortcoming of neglecting aliasing effects in the autocorrelation domain in the VTS parameter estimation. This causes a serious performance degradation for high-pitched female speakers, especially in a noisy condition. It is to be mentioned that in many real-life applications involving speech or biomedical signals, speaker variability is an important attribute. Therefore, the applications of the existing RS based pitch estimation methods are very much limited. This provides a motivation to develop new time-frequency domain pitch estimation methods without ignoring the aliasing effect in estimation of the VTS parameters in the autocorrelation domain. Irrespective of a different speaker, an excitation-like signal (ELS) could be generated by using the RS obtained accurately from the noisy speech or its noise-reduced version. At the same time, a new noise-compensation scheme could be designed to remove the noise from the RS. To this end, employment of a time-domain function of the ELS would be more desirable to enhance periodicity for reliable pitch estimation under a very low SNR. In order to generate a noise-robust ELS at a very low SNR, one potential alternative could be the introduction of a noise reduction block prior to pitch estimation, a framework already available in many practical applications. A suitably designed noise-reduction scheme would alleviate the necessity of noise compensation in the

RS that could be used to generate an ELS efficiently from the noise-reduced speech. Considering that conventional cepstrum analysis has been very popular in speech signal processing in a noise free environment, a development of the time-frequency domain pseudo-cepstrum of the ELS of the enhanced speech could be very suitable for accurate pitch estimation under a severe noisy condition. Finally, a pitch tracking scheme has to be employed, whenever applicable, to obtain a smooth pitch contour.

1.4 Scope and Organization of the Thesis

The objective of this research is to develop effective methodologies for pitch estimation of speech signals in practical situations where only noise-corrupted speech observations are available. Although pitch estimation from clean speech is a widely studied research, very few algorithms have so far been reported to handle noisy environments. In this research, two different approaches are developed: (1) noisy speech observations are directly employed in pitch estimation and (2) an ELS is employed, which is generated from noisy speech observations or its noise-reduced version. In each of the approaches, new time-frequency domain methods are developed, which are capable of providing accurate pitch estimates even in severe noisy conditions. In the first approach, model-fitting based schemes are first introduced to estimate a PH and then, time or frequency domain methods are proposed to determine the corresponding harmonic number required for pitch estimation. In the second approach, two noise handling schemes are first developed, one for noise-compensation in the residual signal and the other for prior noise reduction from noisy observations, and then the ELS based time or time-frequency domain methods are proposed for pitch estimation. In order to demonstrate the effectiveness of the proposed pitch estimation methods, extensive experimentations are performed under different noisy conditions

for different natural speech signals uttered by a wide range of speakers and results are compared with those from some of the existing methods. The thesis is organized as follows.

In Chapter 2, a harmonic cosine autocorrelation (HCAC) model of clean speech signals is first introduced. By using this model, a least-squares correlation-fitting optimization technique is proposed to extract a PH from the Fast Fourier transform (FFT)-pre-processed noisy speech. Next, exploiting the noise-robust PH estimate, a new frequency domain harmonic measure that manifests the degree of harmonicity of harmonic components in a speech frame is deduced to select a set of harmonic numbers associated with the estimated PH. Finally, a harmonic number matching (HNM) scheme is developed to determine the desired harmonic number for pitch estimation by maximizing a harmonic to noise power ratio (HNPR). In order to obtain a physiologically plausible smoothed pitch contour, a pitch tracking scheme is implemented using dynamic programming (DP).

In Chapter 3, a harmonic sinusoidal autocorrelation (HSAC) model of clean speech in terms of its pitch-harmonics is first derived, where unlike the HCAC model proposed in Chapter 2, cross-product terms of different harmonics are taken into consideration. A least-squares optimization technique utilizing the HSAC model is then presented to extract a pitch-harmonic from the discrete cosine transform (DCT)-pre-processed noisy speech. Next, a symmetric average magnitude sum function (SMSF) of the speech signal, which retains the pitch periodicity even in the severe noisy case, is proposed. In order to obtain the desired harmonic number, an SMSF based impulse-train matching (SIM) scheme is introduced using a periodicity matching between the SMSF and a periodic impulse-train whose period is governed by the noise-robust PH estimate. Finally, a DP based pitch tracking scheme is also performed to obtain a

smoothed pitch contour.

In Chapter 4, a new pitch estimation method is developed using an ELS obtained from the noisy speech observations. First, a scheme for the VTS parameter estimation from noisy speech is proposed based on the principle of homomorphic deconvolution. The estimated VTS parameters are used for an inverse filtering of the noisy speech to generate an RS. Then, a correlation domain noise-compensation scheme is proposed to reduce the effect of noise on the RS and a squared Hilbert envelope (SHE) of the compensated RS is computed as the ELS. Finally, prior to pitch tracking, a new symmetric normalized magnitude difference function (SNMDF) of the ELS is presented and utilized to find the pitch estimate thus overcoming the undesirable effect of noise on the ELS under severe noisy conditions.

In Chapter 5, another ELS based pitch estimation method is proposed using noise-reduced speech observations. Instead of performing noise-compensation in the RS, first, a frequency domain noise subtraction scheme is developed considering the possible cross-correlation between speech and noise, which offers an additional advantage of noise updating from time to time. The enhanced frame thus obtained is then utilized to generate a SHE of the RS, which represents the desired ELS. However, in order to further reduce the adverse effect of noise on the ELS under a severe noisy condition, a time-frequency domain pseudo cepstrum of the ELS of the enhanced speech is proposed, which overcomes the limitation of the conventional cepstrum, which is capable of handling clean speech only.

Finally, some concluding remarks highlighting the contributions of the thesis and suggestions for future work are provided in Chapter 6.

Chapter 2

Pitch Estimation Based on a Harmonic Cosine Autocorrelation Model and Frequency-Domain Matching

2.1 Introduction

Pitch estimation under a heavy noisy condition, especially when only the output noisy speech observations are available, is a very difficult but essential task for many practical applications. In a noisy environment, the performance of conventional pitch estimation methods gets generally degraded due to their inability to retrieve a periodic and harmonic structure of speech obscured by the noise. In this Chapter, a novel time-frequency domain method for the estimation of pitch from speech observations in the presence of heavy noise is presented [49]. The method is developed based on a harmonic cosine autocorrelation (HCAC) model of clean speech expressed in terms of its pitch-harmonics, which is derived from a statistical autocorrelation estimator. In order to overcome the problem of conventional methods that yield poor pitch estimates in severe noisy conditions, we first propose to extract a pitch-harmonic (PH) from the Fast Fourier Transform (FFT)-pre-processed noisy speech by developing and

using a least-squares (LS) autocorrelation-fitting optimization technique that employs the introduced HCAC model [50]. By exploiting the extracted PH along with the FFT based power spectrum of noisy speech, a new harmonic measure that manifests the degree of harmonicity of harmonic components in each frame is deduced to select a set of harmonic numbers associated with the PH [51]. Finally, a frequency domain matching scheme is devised to determine the desired harmonic number based on maximizing the proposed harmonic-to-noise power ratio (HNPR) corresponding to each of the selected harmonic numbers. Since the harmonic measure is governed by a noise-robust estimate of the PH and the HNPR effectively prevents the adverse effect of non-pitch peaks and noise, the proposed harmonic number matching (HNM) scheme offers an ease of acquiring the true harmonic number as well as an accurate pitch estimate. Moreover, a pitch tracking scheme by using dynamic programming (DP) is performed to obtain a physiologically plausible smoothed pitch contour for rendering the suitability of the method in practical applications. To demonstrate the efficacy of the proposed time-frequency domain method, referred to as, HCAC-HNM, extensive simulations are carried out by considering natural speech signals obtained from the *Keele* database in the presence of white or multi-talker babble noise available from the *Noisex92* database and the results are compared with those obtained from some of the state-of-the-art techniques in the literature. The simulation results have confirmed the best pitch estimation performance of the proposed methodology for a wide range of speakers from high to very low SNR levels, such as -10 dB. The superior accuracy and consistency of pitch estimates as provided by the proposed method even in a multi-talker babble noise corruption is a testimony of its suitability in practical applications.

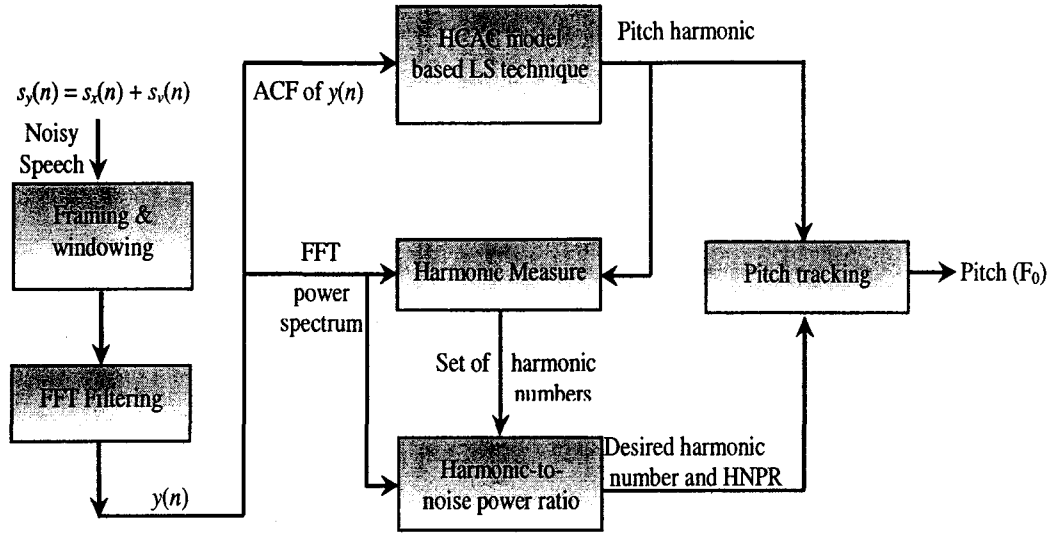
The rest of the chapter is organized as follows. Section 2.2 presents a brief overview

of the proposed method. In Section 2.3, an autocorrelation-domain LS optimization scheme is developed by employing the HCAC model of the clean speech for the estimation of a PH from the FFT pre-processed noisy speech. By utilizing the extracted PH as well as a smoothed power spectrum of noisy speech, a HNM scheme is described in Section 2.4 to determine the harmonic number associated with the PH for pitch estimation. In order to smooth the pitch contour, a DP based pitch tracking is performed in Section 2.5. In Section 2.6, the pitch estimation performance of the proposed method is demonstrated through extensive computer simulations using speech signals corrupted by both white and multi-talker babble noise. Finally, in Section 2.7, the salient features of this investigation are summarized with concluding remarks.

2.2 A Brief Description of the Proposed Method

An overview of the proposed pitch estimation method is shown by a block diagram in Fig. 2.1. In the presence of an additive noise $s_v(n)$, a clean speech signal $s_x(n)$ gets contaminated and produces noisy speech observations $s_y(n)$. The observed noisy speech is segmented by using a sliding window function $w(n)$ into overlapped frames each having a size of N samples. It is well known that voiced speech is dominated by the energy in the first-formant [1]. In order to reduce the detrimental effects of the higher formants and the high frequency noise components in pitch estimation, each windowed noisy frame is low-pass filtered retaining the frequency components up to the upper limit of the first formant band. A windowed low-pass filtered noisy speech frame can be expressed in the time-domain as

$$y(n) = x(n) + v(n) \tag{2.1}$$



FFT: Fast Fourier Transform; HCAC: Harmonic Cosine Autocorrelation; LS: Least Squares; ACF: Autocorrelation Function
HNPR: Harmonic-to-noise-power ratio

Figure 2.1: A block diagram representing the overview of the proposed pitch estimation method.

where $x(n)$ and $v(n)$ represent the windowed low-pass filtered clean speech $s_x(n)$ and noise $s_v(n)$, respectively. Given the pre-processed noisy speech $y(n)$, we develop a novel technique for the extraction of one PH by exploiting the HCAC model of $x(n)$ and a new frequency-domain HNM scheme for determining the harmonic number corresponding to the extracted PH. Thus, with the knowledge of these two quantities, one can easily obtain the value of the pitch on a frame-by-frame basis. Note that in the proposed optimization technique for the PH estimation, an initial estimate of the PH is obtained from the maximum peak of the smoothed FFT power spectrum.

2.3 Extraction of a Pitch-Harmonic using the HCAC Model

In this subsection, we present the proposed method of estimating a PH by fitting the autocorrelation function (ACF) of $y(n)$ with a harmonic cosine autocorrelation (HCAC) model of $x(n)$ using the LS minimization technique.

2.3.1 HCAC Model for Clean Speech

The ACF of a clean speech sequence $s_x(n)$ is given by

$$\phi_{s_x}(m) = E\left[s_x(n)s_x(n+m)\right] \tag{2.2}$$

where m is the discrete lag variable. It is to be mentioned that in speech analysis it is usually assumed that the properties of $s_x(n)$ change relatively slowly with time. This allows examination of a short-time frame of $s_x(n)$ to extract parameters presumed to remain fixed for the duration of the frame. In a voiced frame, the clean speech $x(n)$ can be expressed as a sum of harmonic signals as

$$x(n) = \sum_{p=1}^{\kappa} b_p \cos(n\omega_p + \varphi_p) \tag{2.3}$$

where the parameters b_p, φ_p and ω_p represent, respectively, the amplitude, phase and normalized angular frequency of the p-th harmonic of $x(n)$, and κ is the number of harmonics retained in the first formant band of the pre-processed speech. The normalized angular frequency ω_p of the p-th harmonic is related to the normalized angular pitch frequency ω_0 as

$$\omega_p = p\omega_0 \tag{2.4}$$

Note that $\omega_0 = \dfrac{2\pi F_0}{F_s}$, where F_0 and F_s are, respectively, the pitch frequency and the sampling frequency of $x(n)$ in Hz. Starting from the statistical autocorrelation

estimator, using the representation of clean speech $x(n)$ given by (2.3), a model of the ACF for $x(n)$ can be derived as

$$\phi_{x_{Model}}(m) = \sum_{p=1}^{\kappa} \gamma_p \cos(\omega_p m) \tag{2.5}$$

where $\gamma_p = \frac{b_p^2}{2}$. In deriving (2.5), the contributions from the cross-product terms of different harmonics have been neglected. We refer to (2.5) as the harmonic cosine autocorrelation (HCAC) model of $x(n)$ [50]. It is seen that the HCAC model of clean speech as given by (2.5) is expressed explicitly in terms of the pitch-harmonics, ω_p's of $x(n)$.

2.3.2 HCAC Model Fitting: Least-Squares Optimization

On the other hand, given a speech frame $x(n)$, its ACF can be directly estimated as

$$\phi_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x(n)x(n+|m|), \quad m = 0, \pm 1, \ldots, \pm M, M < N \tag{2.6}$$

Even though in practice the frame length N is finite, the conventional ACF estimator $\phi_x(m)$ in (2.6) offers an efficient way to obtain a reasonably accurate estimate for the true ACF $\phi_{s_x}(m)$ of the speech signal $s_x(n)$ as given by (2.2) [52]. In a voiced frame, the ACF $\phi_x(m)$ of the clean speech $x(n)$ exhibits peaks at lag values that are at or in the neighborhood of pitch period T_0 and at or in the neighborhoods of multiples and submultiples of T_0 with the maximum peak occurring at $m = T_0$ or in its neighborhood. In the presence of noise, i.e., for a pre-processed noisy speech frame $y(n)$, the estimate of ACF $\phi_y(m)$ can be computed using (2.6) as

$$\phi_y(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} y(n)y(n+|m|), \quad m = 0, \pm 1, \ldots, \pm M, M < N. \tag{2.7}$$

Using (2.1),(2.7) can be re-written as

$$\phi_y(m) = \phi_x(m) + \phi_v(m) + \phi_c(m) \tag{2.8}$$

where $\phi_c(m)$ represents the summation of all cross-correlation terms. It is worth mentioning that under a heavy noisy condition, the second and third terms of ACF $\phi_y(m)$ of $y(n)$ in (2.8) cannot be neglected. Consequently, $\phi_y(m)$ may significantly differ from $\phi_x(m)$ at all lags and the peak values in $\phi_y(m)$ may get emphasized or de-emphasized. Yet, the property of $\phi_y(m)$ in retaining the peaks of $\phi_x(m)$ at multiples of the pitch period even under severe noisy conditions motivates us to utilize it for extracting one PH ω_p, which in our method would be sufficient to estimate the pitch.

It is seen that while the HCAC model $\phi_{x_{Model}}(m)$ of $x(n)$, as given by (2.5), does not contain the perceptually less important phase information φ_p, it does preserve the pitch information of $x(n)$ contained in (2.3) in terms of the pitch-harmonics ω_p's. Any one of the κ components in the summation of (2.5), say the p-th one,

$$\phi^{(p)}_{x_{Model}}(m) = \gamma_p \cos(\omega_p m) \tag{2.9}$$

and the associated PH ω_p can be estimated from the M lags of the available noisy ACF $\phi_y(m)$ as computed according to (2.7) by employing an LS fitting technique. For the PH ω_p, the corresponding parameter γ_p is determined such that the total squared error between $\phi_y(m)$ and the component $\phi^{(p)}_{x_{Model}}(m)$ spanned over all the m lags as given by

$$\Theta(\gamma_p) = \sum_{m=1}^{M} \left[\phi_y(m) - \phi^{(p)}_{x_{Model}}(m)\right]^2 \tag{2.10}$$

is minimized. Since we are interested to determine one PH only, it is sufficient to consider one component $\phi^{(p)}_{x_{Model}}(m)$ of the HCAC model in the above autocorrelation-fitting technique. In $\phi_y(m)$, the effect of noise is dominant mainly at $m = 0$. Hence, in order to reduce the noise effect, $m > 0$ is considered. For a given ω_p, one can find the value γ_p that minimizes the function $\Theta(\gamma_p)$ given by (2.10) by equating its

derivative with respect to γ_p to zero, that is,

$$\frac{d\Theta(\gamma_p)}{d\gamma_p} = 0. \tag{2.11}$$

The above equation yields the optimum value of γ_p as given by

$$\gamma_p = \frac{\sum_{m=1}^{M} \phi_y(m) \cos(\omega_p m)}{\sum_{m=1}^{M} \cos^2(\omega_p m)}. \tag{2.12}$$

The LS fitting scheme is based on the assumption of an a *priori* knowledge of the value of a pitch-harmonic ω_p. However, at this stage, this assumption is not fully valid. As a matter of fact, our objective is to determine an accurate value of one of the pitch-harmonics. For this purpose, we proceed as follows. It is well known that in a voiced frame, the power spectrum of the clean speech $x(n)$ exhibits strong peaks at or near pitch-harmonics ω_p's [53]. We analyze the power spectrum of the pre-processed noisy speech frame $y(n)$ based on the discrete Fourier transform (DFT), where the DFT $Y(k)$ of $y(n)$ is given by

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j2\pi nk/N}, \quad 0 \leq k \leq N-1. \tag{2.13}$$

This equation is implemented by the FFT operation [54], where k is the frequency bin index. The frequency points of peak occurrence of the FFT power spectrum $|Y(k)|^2$ of $y(n)$ in the presence of a heavy noise may not be the same as those of $|X(k)|^2$, the FFT power spectrum of $x(n)$. In order to reduce the fluctuations around the strong peaks of $|Y(k)|^2$, it is necessary to smooth it. We accomplish this process as follows:

$$|\tilde{Y}(k)|^2 = o_k |Y(k)|^2 + (1 - o_k) \left[\frac{1}{(2D+1)} \sum_{u=-D}^{D} |Y(k-u)|^2 \right], \tag{2.14}$$

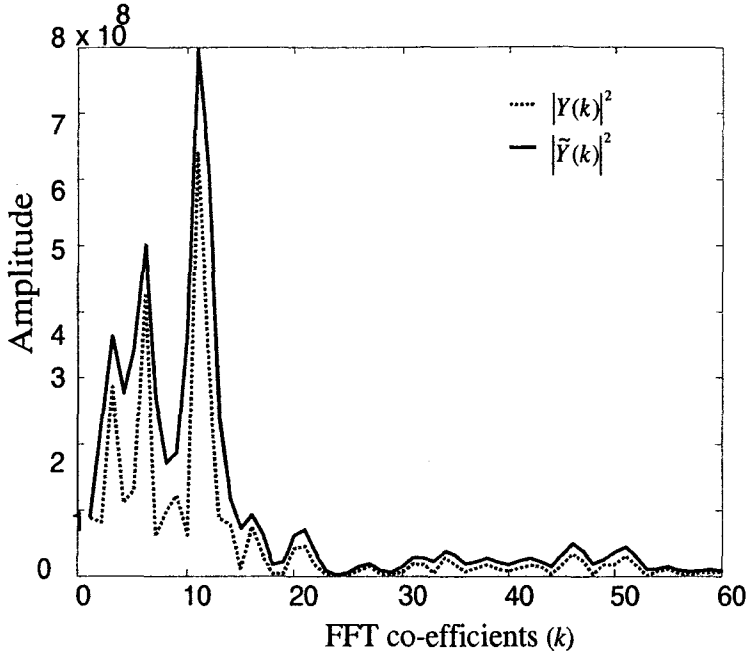


Figure 2.2: FFT power spectrum of $y(n)$ with and without smoothing.

where o_k is a smoothing factor and the length of the smoother has been chosen to be $(2D + 1)$. Fig.2.2 shows the FFT power spectrum of $y(n)$ with and without smoothing. It is evident from this figure that smoothed FFT power spectrum $|\tilde{Y}(k)|^2$ has peaks that are more prominent compared to that of $|Y(k)|^2$. However because of the underlying noise, the frequency point corresponding to the maximum peak of $|\tilde{Y}(k)|^2$ may not represent the exact position of the harmonic ω_p. Thus, this frequency point is taken only as an initial estimate $\omega_p^{(0)}$ for the harmonic ω_p [55].

For this initial estimate $\omega_p^{(0)}$, an initial estimate $\gamma_p^{(0)}(m)$ of the corresponding $\gamma_p(m)$ can be easily obtained using (2.12). With this value of $\gamma_p^{(0)}(m)$, the estimate $\phi_{x_{Model}}^{(p)(0)}(m)$ of the component $\phi_{x_{Model}}^{(p)}(m)$ of the HCAC model can then be computed by using (2.9). Finally, the minimum value of the total squared error $\Theta^{(0)}$ for the initial estimate $\omega_p^{(0)}$ is determined using (2.10). With the initial estimates $\omega_p^{(0)}$, $\gamma_p^{(0)}(m)$,

$\phi^{(p)(0)}_{x_{Model}}(m)$ and $\Theta^{(0)}$ at our disposal, we can now perform an iterative search within a small neighborhood around $\omega_p^{(0)}$ with a reasonable resolution by using (2.12), (2.9), and (2.10) in that order repeatedly. The outcome of this optimization procedure is an accurate estimate ω_{popt} of the pitch-harmonic ω_p that best matches $\phi^{(p)}_{x_{Model}}(m)$ with $\phi_y(m)$.

The PH extraction scheme of the proposed pitch estimation method as described in this Section offers a two-fold advantage. Firstly, the scheme employs the HCAC model of clean speech, which provides a direct relationship between pitch-harmonics and the pitch. Secondly, unlike the approach that relies only on a single lag corresponding to the global maximum of the noisy ACF $\phi_y(m)$ [45], we employ a total of a relatively large number of lags (M) of $\phi_y(m)$ for its LS matching with a noise-free component $\phi^{(p)}_{x_{Model}}(m)$ of the HCAC model. The novelty of the PH extraction scheme lies in its ability to extract the optimum value of the PH ω_{popt} associated with the $\phi^{(p)}_{x_{Model}}(m)$ component of the HCAC model from the available noisy ACF $\phi_y(m)$ through a model-fitting based approach instead of using the noisy ACF $\phi_y(m)$ directly.

2.4 Determination of the Harmonic Number using the HNM Scheme

In this Section, we would like to estimate the pitch F_0 by using the estimated PH, $\omega_{popt} = p_{opt}\omega_0 = p_{opt}(\frac{2\pi F_0}{F_s})$. It is clear that the key to this task is to determine the harmonic number p_{opt} associated with ω_{popt}. To this end, we develop a harmonic number matching (HNM) scheme based on the smoothed FFT power spectrum $|\tilde{Y}(k)|^2$ of $y(n)$, which was already obtained in the previous Section.

2.4.1 Proposed Harmonic Measure

The HNM method begins with determining a range for the harmonic number p, which is an integer. Considering that the pitch frequency of human speech varies typically in the range of 50 Hz to 500 Hz, and recalling that the sampling frequency employed in this study is specified as F_s, the estimated ω_{popt} results in the minimum and maximum of the harmonic number p that can be given by

$$p_{min} = \left\lceil \frac{\omega_{popt}/2\pi}{500Hz/F_s} \right\rceil, p_{max} = \left\lfloor \frac{\omega_{popt}/2\pi}{50Hz/F_s} \right\rfloor \tag{2.15}$$

where $\lceil . \rceil$ and $\lfloor . \rfloor$, respectively, stand for the ceiling and floor operations. Thus, for each choice of p in the range $[p_{min}, p_{max}]$, a potential pitch frequency in terms of the number of FFT coefficients can be written as

$$K_{0,p} = \left\lceil \frac{(\omega_{popt}/2\pi)(N)}{p} \right\rceil \tag{2.16}$$

where N is the size of the FFT operation. Clearly, if $K_{0,p}$ happens to be the true pitch frequency, its harmonics are in general revealed as much stronger peaks in the smoothed FFT power spectrum $|\tilde{Y}(k)|^2$ compared to those corresponding to other FFT coefficients. Let $k_f = [fK_{0,p}]$ be the f-th harmonic frequency of any potential pitch frequency $K_{0,p}$. Then, taking into account the coefficients (amplitudes) of $|\tilde{Y}(k)|^2$ at $K_{0,p}$ and harmonics of $K_{0,p}$, the following harmonic measure associated with the particular harmonic number p is proposed as

$$\Omega_p = \left[\Pi_{f=1}^{\kappa_p} |\tilde{Y}(k_f)|^2 \right]^{\frac{1}{\kappa_p}}, \quad k_f = [fK_{0,p}], \tag{2.17}$$

where κ_p is the number of the harmonics of $K_{0,p}$ retained up to the upper limit of the first formant band in the pre-processed frame $y(n)$. Since $|\tilde{Y}(k)|^2$ exhibits significant peaks at integer multiples of the pitch frequency, it is expected that Ω_p would produce

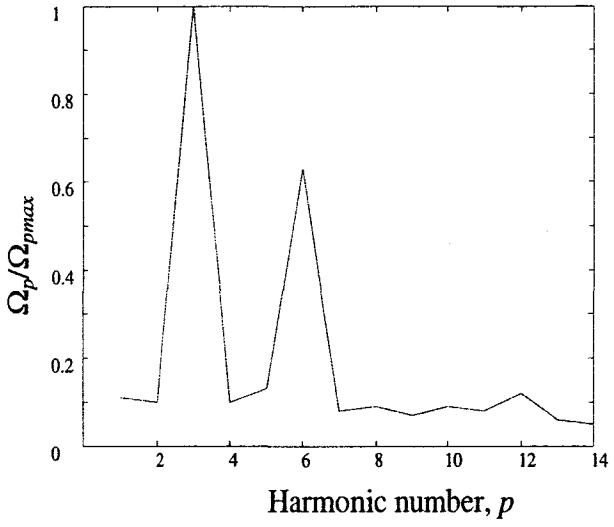


Figure 2.3: A normalized harmonic measure for a reference voiced frame at SNR $= -5$ dB.

the maximum peak at p_{opt}, where the corresponding $K_{0,p}$ is synchronized with the true pitch frequency [56], [57], [58]. By conducting a simple analysis of Ω_p, however, one can find that, for other values of p whose corresponding $K_{0,p}$ is an exact multiple or submultiple of pitch frequency, Ω_p can still give very large values. Note that due to the effect of the voice formant structure as well as noise, the maximum peak of Ω_p, denoted as Ω_{pmax} , does not necessarily occur at p_{opt}. For example, when ω_{popt} is an even harmonic, and odd harmonics are attenuated for some values of p, the maximum of Ω_p occurs at $p = p_{opt}/2$. Fig. 2.3 illustrates a plot of the normalized harmonic measure (Ω_p/Ω_{pmax}) of a reference voiced frame at an SNR of -5 dB, where p_{opt} =6 is found from the known pitch frequency of the reference frame and the PH ω_{popt} obtained for this frame using the LS fitting technique developed in the previous Section. It is clear that Ω_p exhibits prominent peaks for p =3, 6, and 12, while the maximum of Ω_p corresponds to p=3. Therefore, one cannot rely on the maximum peak Ω_{pmax} to determine the true harmonic number p_{opt}. From

our observation and analysis of many such voiced frames, however, it is consistently found that the locations of the prominent peaks of Ω_p, such as p=3, 6, 12 in Fig. 2.3, form a small set of p values including p_{opt}. By using the reduced set of p values, we now propose in the following a second criterion, referred to as the harmonic-to-noise power ratio (HNPR), in the FFT domain in order to determine the true harmonic number p_{opt}.

2.4.2 Proposed Harmonic-to-Noise Power Ratio

Note that, for each value of p in the reduced set, there are κ_p harmonics of $K_{0,p}$, based on which we can define the following HNPR,

$$\mu_p = \frac{\sum_{f=1}^{\kappa_p} [\Im_f - \aleph_f]}{\sum_{f=1}^{\kappa_p} [\aleph_f + \chi_f]} \tag{2.18}$$

where $\Im_f$ and $\aleph_f$ represent, respectively, the overall power and the noise power in the frequency band of the f-th harmonic of $K_{0,p}$, and χ_f represents the noise power in the frequency band between any two consecutive harmonics, where only noise exists. It is clear that each term $[\Im_f - \aleph_f]$ in the numerator of (2.18) quantifies a harmonic power in the corresponding frequency band while each term $[\aleph_f + \chi_f]$ in the denominator represents the total noise power associated with each harmonic. For $k_f = [fK_{0,p}]$, i.e., the f-th harmonic of $K_{0,p}$ with a harmonic bandwidth $B_p = \lceil \frac{K_{0,p}}{2} \rceil$, the overall power in the frequency band $[\lceil k_f - B_p/2 \rceil \sim \lceil k_f + B_p/2 \rceil]$ can be computed as

$$\Im_f = \sum_{k=\lceil k_f - B_p/2 \rceil}^{\lceil k_f + B_p/2 \rceil} |\tilde{Y}(k)|^2 \tag{2.19}$$

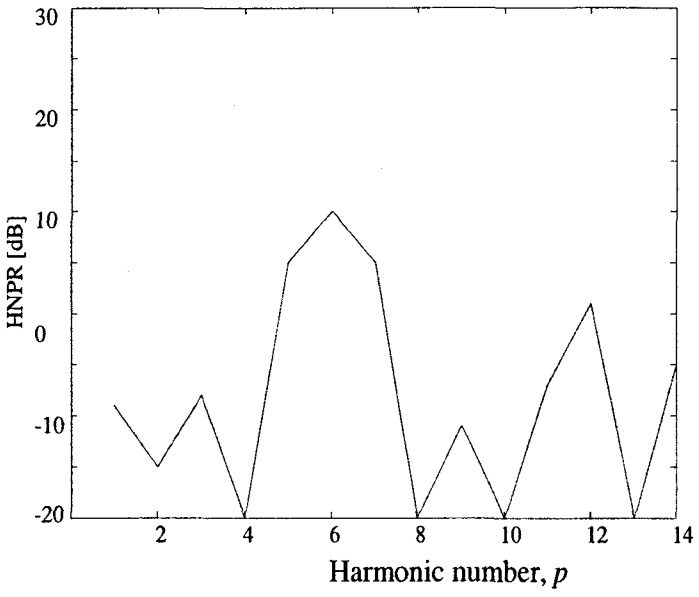


Figure 2.4: The harmonic-to-noise power ratio in the FFT domain for a reference voiced frame at SNR = -10 dB .

Similarly, in the frequency band $[\lceil k_{f-1} + B_p/2 \rceil \sim \lceil k_f - B_p/2 \rceil]$ between any two consecutive harmonics k_{f-1} and k_f of $K_{0,p}$, the noise power χ_f can be given by

$$\chi_f = \sum_{k=\lceil k_{f-1}+B_p/2 \rceil}^{\lceil k_f - B_p/2 \rceil} |\tilde{Y}(k)|^2. \tag{2.20}$$

Clearly, χ_f gives the noise power in each bandwidth $(K_{0,p} - B_p)$. Exploiting this knowledge, the noise power in the frequency band B_p of the f-th harmonic of $K_{0,p}$ can be reasonably assumed to be

$$\aleph_f = \chi_f \left[\frac{B_p}{K_{0,p} - B_p} \right]. \tag{2.21}$$

It is clear from the above discussion that in the computation of μ_p, all the FFT coefficients in the bandwidth of each harmonic of $K_{0,p}$ and those in the bandwidth between any two consecutive harmonics of $K_{0,p}$ have been used. This is the main difference from the calculation of Ω_p, where only the FFT coefficients corresponding

to the multiple of $K_{0,p}$ are employed. Fig. 2.4 shows a plot of HNPR for the reference frame as used in Fig. 2.3 but at an SNR of -10 dB. It is obvious that μ_p reaches its maximum at $p = 6$ as expected, implying that the new criterion μ_p is very noise-robust. For each harmonic number that is lower than the p_{opt} (such as $p = 3$ in Fig. 2.4), it is clear from (2.16) that the corresponding $K_{0,p}$ would be higher than K_{0opt}, the optimum pitch estimate in number of FFT coefficients with respect to p_{opt}. This results in a phenomenon of missing some true pitch-harmonics, thus reducing the number κ_p of harmonics. Also, the bandwidth B_p of each harmonic of $K_{0,p}$ is increased, which in turn elevates the corresponding noise power $\aleph_f$ in B_p. Due to the missing original harmonics, the bandwidth between any two consecutive harmonics of $K_{0,p}$ increases, which boosts the noise power χ_f in the corresponding band. As a consequence of increasing $\aleph_f$ and χ_f, the denominator of (2.18) becomes larger, thus decreasing the value of μ_p. On the contrary, for a harmonic number which is larger than p_{opt} (such as $p = 12$ in Fig. 2.4), the corresponding $K_{0,p}$ is smaller than K_{0opt}, which decreases the harmonic bandwidth B_p and increases the number κ_p of harmonics. In this case, half of these harmonics, which are not the true harmonics, contribute insignificantly to the calculation of harmonic power in the numerator of (2.18). Although the other half of the harmonics is retained at the exact locations of the original harmonics, the power of each harmonic decreases due to the reduced B_p. As a result, the numerator of (2.18) gets a lower value, which in turn reduces the value of μ_p. The significant advantage of μ_p lies in the fact that when p best matches with p_{opt} (such as $p = 6$ in Fig. 2.4), the corresponding $K_{0,p}$ leads towards the true pitch and its harmonic locations in $|\tilde{Y}(k)|^2$, thus yielding the highest value of μ_p. According to our extensive experimentations, the effectiveness of μ_p in matching p_{opt} accurately is validated for most of the voiced frames even at a very low SNR of

-10 dB. Therefore, the value of p corresponding to the largest value of μ_p is regarded as p_{opt}, which leads to the desired estimate $\tilde{F}_0$ of pitch F_0 in Hz for a voiced frame as shown below.

$$\tilde{F}_0 = \left[\frac{K_{0opt}}{N}\right] F_s, \quad K_{0opt} = \left\lceil \frac{(\omega_{popt}/2\pi)(N)}{p_{opt}} \right\rceil \tag{2.22}$$

It is of interest to note that in comparison to the method in [24], which is based on the representation of voiced speech as a sum of harmonic signals, the proposed method requires extracting only one harmonic component and the associated pitch-harmonic, implying that estimation of the unknown parameters related to all other harmonics is not needed. The complete method for pitch estimation whose development started in the Section 2.2 will henceforth be referred to as the HCAC-HNM method.

2.5 HNPR Based Pitch Tracking using Dynamic Programming

A human pitch contour should exhibit a continuous and smooth behavior over time. However, fluctuation in the estimates of pitch over the frames is inevitable, since pitch estimates of the individual frames are obtained independently. In order to achieve an overall smooth pitch contour, it is necessary to consider the pitch information of consecutive frames. In this section, a pitch tracking scheme is presented to reduce the error in the F_0 contour and thus to minimize the fluctuations of the pitch value from frame to frame.

Given a set of potential candidates for the pitch at each frame, a common requirement for pitch (F_0) tracking is to find the optimal pitch path connecting one candidate per frame through the set of pitch candidates over all time frames. In our pitch tracking scheme, the pitch estimate $\tilde{F}_0$ corresponding to K_{0opt} as given in (2.22)

determined by using the HNM scheme, is selected to be one of the members of the set of potential pitch candidates for a frame. For a given frame t, a certain number, say $(J_t - 1)$, of maxima of the HNPR μ_p given by (2.18) yielding the pitch values lying within the pitch range are chosen, and the pitch values corresponding to these maxima are selected as the other possible pitch candidates. Among the J_t number of potential candidates at frame t, $\tilde{F}_0$ is assigned the highest priority to be the correct pitch value for the frame under consideration. The remaining $(J_t - 1)$ candidates are prioritized by sorting them according to decreasing magnitude of the corresponding values of μ_p. We now select the true pitch value for each frame from the set of pitch candidates generated as described above so as to minimize a composite total cost function corresponding to a pitch path consisting of a local cost component and a transition cost component. The first component consists of the summation of the local costs of all the chosen candidates in the path, whereas the second component consists of the sum of the transition costs between the pair of chosen candidates of the neighboring frames in the path. Let the j-th pitch candidate in the t-th frame in terms of FFT co-efficients be represented by $K_{0,t}^j$, which corresponds to the pitch frequency candidate in Hz, $F_{0,t}^j = \left[\frac{K_{0,t}^j}{N}\right] F_s$. The local cost for the candidate $F_{0,t}^j$ is defined as

$$C_{local}(F_{0,t}^j) = -\mu_{p_t^j}, \tag{2.23}$$

where $\mu_{p_t^j}$ represents the HNPR that is computed according to (2.18) for a harmonic number p_t^j corresponding to the candidate $F_{0,t}^j$(or $K_{0,t}^j$) . Obviously, the pitch frequency candidate having a higher score of the HNPR $\mu_{p_t^j}$ will result in a lower local cost $C_{local}(F_{0,t}^j)$. Due to the deviation among the pitch candidates from one frame to

next, a transition cost can be defined as

$$C_{tran}(F^i_{0,t-1}, F^j_{0,t}) = \left| \frac{F^i_{0,t-1}}{F^j_{0,t}} \right|, \tag{2.24}$$

which measures the cost of the pitch path going from the i-th candidate of frame $(t-1)$ to the j-th candidate of frame t. It is seen from this equation that the transition between pitch candidates is favored with a low cost if their values are close. Obviously, a lower transition cost means a high probability of the transition between the two pitch candidates under consideration, thus ensuring a continuity of the pitch contour. Considering that there are O_c consecutive frames, i.e., $t = 1, 2, \ldots, O_c$, the total cost function $C(A)$ corresponding to an arbitrarily chosen trajectory $A = \left\{ F^{j_1}_{0,1}, F^{j_2}_{0,2}, \ldots, F^{j_{O_c}}_{0,O_c} \right\}$ is defined as

$$C(A) = C_{local}(F^{j_1}_{0,1}) + \sum_{t=2}^{O_c} \left[\varpi . C_{tran}(F^{j_{t-1}}_{0,t-1}, F^{j_t}_{0,t}) + C_{local}(F^{j_t}_{0,t}) \right]. \tag{2.25}$$

In (3.39),$F^{j_t}_{0,t}$ means one of the available candidates from all $F^j_{0,t}$ at the t-th frame to realize a path and ϖ is a weighting factor balancing the local and the transition cost. A large value of ϖ in general gives a better continuity of the pitch contour. Now, the problem of pitch tracking is to minimize the total cost function as formulated by (3.39). It is clear from the nature of this formulation that the total cost function can be minimized through its recursive computation in stages. Since DP is best suited to tackle such a problem, we use this technique for solving the problem at hand efficiently [48], [59], [60], [61]. For this purpose, we employ only the first three pitch candidates from the prioritized set for each frame. The choice of three potential candidates supports a reasonable tradeoff between the complexity and performance of the DP in pitch tracking [62].

2.6 Simulation Results

In this Section, extensive simulations are carried out to evaluate the performance of the proposed pitch estimation method in the presence of noise. For this purpose, we consider natural speech signals corrupted by an additive noise. The performance of the proposed HCAC-HNM method in terms of the accuracy and consistency of the pitch estimates is evaluated and compared to some of the state-of-the-art pitch estimation techniques using a common platform.

2.6.1 Simulation Conditions

(a) Database and other details: In our simulation study, we employ the *Keele* database [63] of real speech signals, specially developed in [64] for the purposes of evaluating pitch estimation and tracking systems. This database provides a reference pitch, which is obtained from a simultaneously recorded laryngograph trace, and referred to as the "ground" truth. There are 10 long utterances by 5 male and 5 female mature English speakers with a total duration of 9 minutes. The data sequence in each utterance consists of a phonetically balanced text, "The North Wind Story". The Keele database is of studio quality, sampled at 20 kHz with 16-bit resolution. A speech sequence in the database is divided into overlapped analysis frames, each of size $N = 25.6$ ms at a frame rate of 100 Hz (i.e., a frame shift of 10 ms). Each voiced frame is assigned a specific pitch value, an unvoiced frame is provided with a zero pitch value, and an uncertain frame is filled with a negative value of -1.

In order to imitate a condition for a noisy environment, we add noise with different signal-to-noise ratios (SNRs) to the original clean speech signal giving

$$s_y(n) = s_x(n) + s_v(n). \tag{2.26}$$

In accordance with the state-of-the-art literature, the white noise sequence available in the *Noisex92* database [65] is used to imitate white-noise corrupted speech. As an example of real environmental noise, the multi-talker babble noise sequence available in the *Noisex92* database is utilized. The source of the multi-talker babble noise is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible. The desired SNR is defined as, $SNR = 10\log_{10}[P_{s_x}/P_{s_v}]$, with, $P_{s_x} = \sum_n s_x(n)^2$ and $P_{s_v} = \sum_n s_v(n)^2$. The noisy speech $s_y(n)$ is segmented into N-sample overlapped frames by the application of a window function $w(n)$, yielding a windowed noisy frame,$y'(n) = y_t(n)w(n)$, where the t-th frame is given by

$$y_t(n) = s_y(u_s t + n), n = 0, 1, 2, \ldots, N-1; \quad t = 0, 1, 2, \ldots, O_T - 1. \tag{2.27}$$

To minimize the signal discontinuities at the edges of a frame, each frame $y_t(n)$ is weighted by a $w(n)$ with a frame shift of u_s samples. In (2.27), O_T is the number of the frames in an utterance. In our simulations, we used the same values for the parameters, such as, frame rate and basic frame (or window) size as specified in the *Keele* database. A normalized hamming window function with smooth onset and offset is applied to each 25.6 ms frame. To exclude the silence regions from the speech $s_x(n)$, the frames having an average power of less than one-thirtieth of the initial average power of $s_x(n)$ are removed from the calculation of P_{s_x} and then the average power of $s_x(n)$ is calculated for the remaining frames [48].

To reduce the negative effects of both the additive noise and the formants of the vocal tract, a windowed noisy frame $y'(n)$ is normally low-pass filtered in a preprocessing stage. However, it is difficult to determine a proper cutoff frequency that should be high enough to accommodate the first harmonic for a high-pitch voice and yet

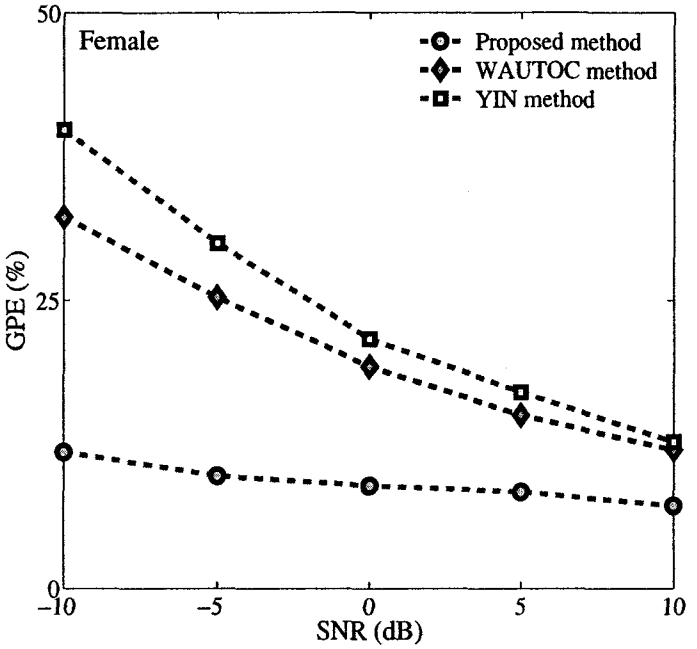


Figure 2.5: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in white noise.

be low enough to reject the second harmonic of a low-pitch voice. In this paper, we propose to use the FFT for the preprocessing of $y'(n)$. The FFT co-efficient of the windowed noisy speech is given by

$$Y'(k) = X'(k) + V'(k), \tag{2.28}$$

where $X'(k)$ and $V'(k)$ represent the FFT coefficients of the windowed clean speech and the windowed noise, respectively. The deleterious impact of noise on speech is relatively small in the first formant range for almost all male and female speakers. Accordingly, the FFT coefficients corresponding to a frequency as high as that of the upper limit of the first formant band should be retained, whereas the rest of the coefficients can be thresholded to zero to effectively eliminate the effects of the

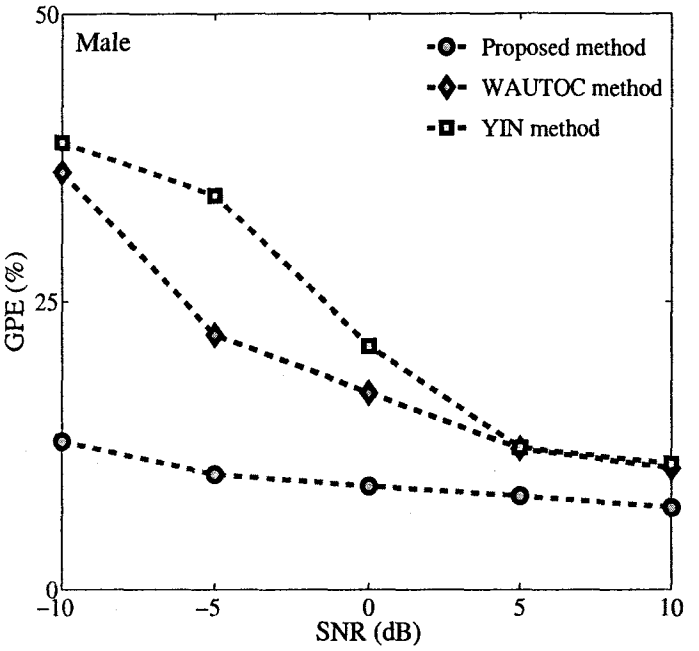


Figure 2.6: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in white noise.

higher formants resulting in a truncated number of FFT coefficients, $Y(k)$. The time-domain frame of preprocessed noisy speech $y(n)$ as given by (2.1) can be obtained through an inverse FFT operation. The pre-processing of each windowed noisy frame $y'(n)$ is performed using N-point FFT and IFFT operations [66]. It should be pointed out that the DCT pre-processing preserves a sufficient number of strong harmonics in $y(n)$, which improves the accuracy of the pitch estimation.

For a frame of the pre-processed noisy speech $y(n)$, in order to determine the optimum value ω_{popt} , we use a search range of 0.1π for ω_p, centered at the initial estimate $\omega_p^{(0)}$ of ω_p obtained from the frequency location corresponding to the maximum peak of the smoothed FFT power spectrum $|\tilde{Y}_f(k)|^2$ of $y(n)$, as given by (2.14). The search resolution for the ω_p is set to be $\Delta\omega_p = 0.01\pi$. The number of ACF lags (M)

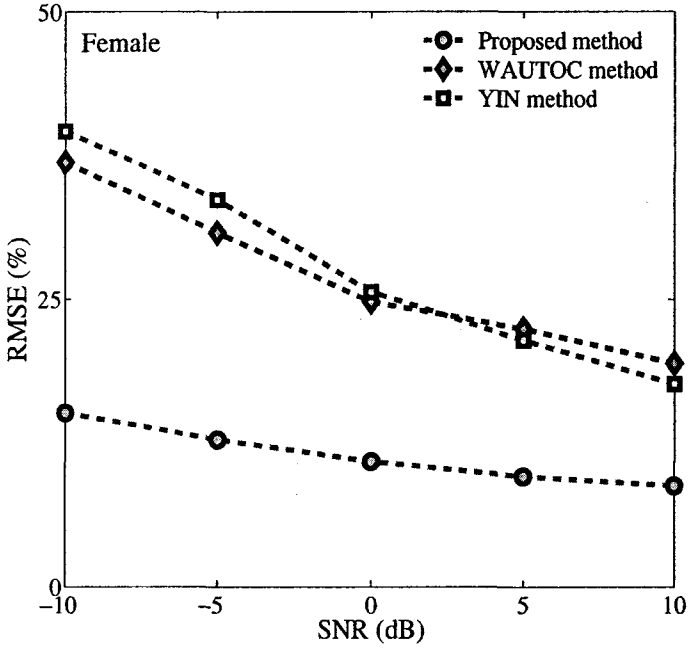


Figure 2.7: RMSE (%) as a function of SNR for female speaker group in white noise.

for the LS problem is selected such that $m_{max} < M < N$ in order to accommodate speakers having the shortest (m_{min} =60 samples or 3 ms) to the longest (m_{max}=400 samples or 20 ms) possible pitch period. The weighting factor ϖ used in (3.39) for pitch tracking is empirically set to 4, since a choice of this value gives the minimum pitch error for many speakers of the database and noise conditions.

(b) Metrics and Comparison Methods Used for Performance Evaluation: For the performance evaluation of the proposed method, criteria considered in our simulation study are 1) the gross pitch error (GPE); 2) the fine pitch error (FPE); and 3) the root-mean-square-error (RMSE). Let $F_0^{ref}(t)$ and $\tilde{F}_0(t)$ represent, respectively, the true and estimated pitch frequencies in Hz of the t-th voiced frame of an utterance obtained from the database. An estimated $\tilde{F}_0(t)$ is classified as "incorrect"

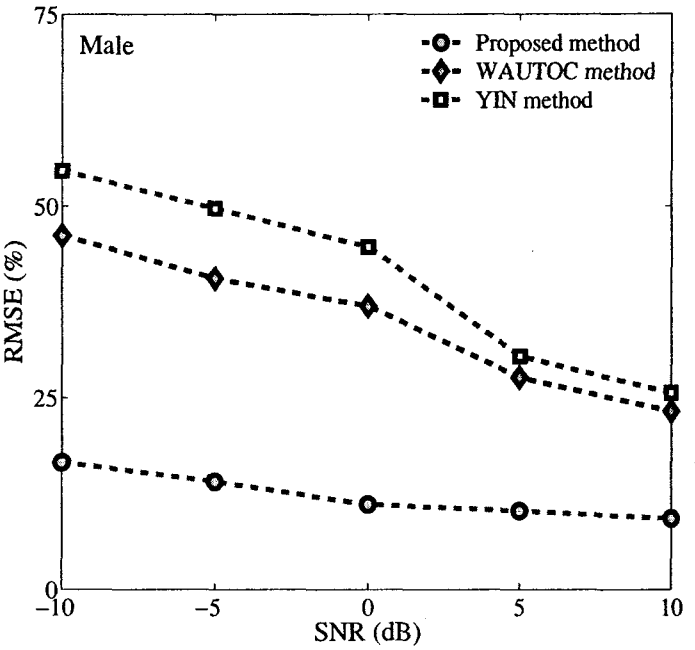


Figure 2.8: RMSE (%) as a function of SNR for male speaker group in white noise.

if it falls outside ±20% of the true pitch value ${F_0}^{ref}(t)$ and is judged to cause GPE, otherwise it is regarded to produce FPE [31], [48]. The percentage of GPE (PGPE) is the ratio of the number of frames (O_{GPE}) yielding GPE to the total number of voiced frames (O_v) multiplied by 100 as given by

$$PGPE = \frac{O_{GPE}}{O_v} \times 100. \tag{2.29}$$

The FPE is calculated over time frames (O_{FPE}) where $\tilde{F}_0(t)$ is estimated correctly, i.e., the error is within ±20%. The mean m_{FPE} and standard deviation σ_{FPE} of FPE are, respectively, defined as

$$m_{FPE} = \frac{1}{O_{FPE}} \sum_{t=1}^{O_{FPE}} \left[\frac{\left| {F_0}^{ref}(t) - \tilde{F}_0(t) \right|}{{F_0}^{ref}(t)} \times 100 \right], \tag{2.30}$$

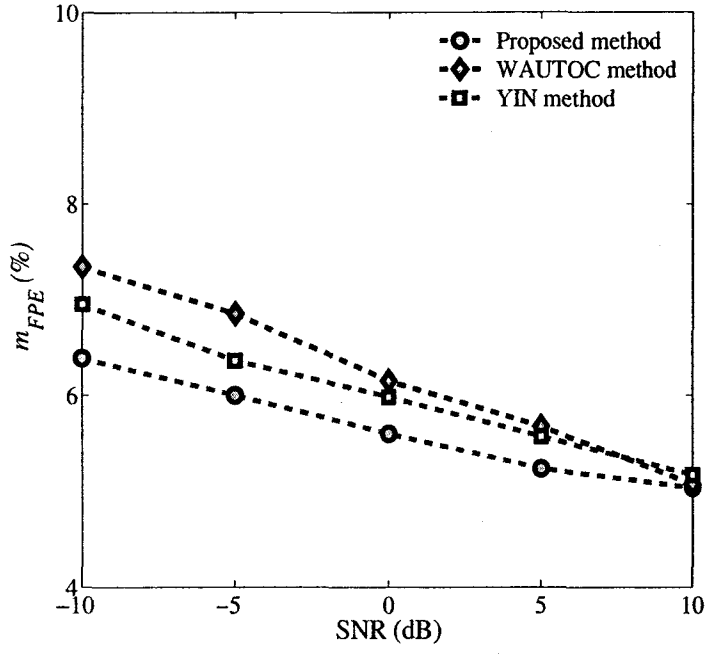


Figure 2.9: m_{FPE} (%) as a function of SNR for all female and male speakers in white noise.

and

$$\sigma_{FPE} = \sqrt{\frac{1}{O_{FPE}-1}\sum_{t=1}^{O_{FPE}}\left[\left(\frac{\left|F_0{}^{ref}(t)-\widetilde{F}_0(t)\right|}{F_0{}^{ref}(t)}\times 100\right)-m_{FPE}\right]^2}. \tag{2.31}$$

As metrics, the PGPE along with mean m_{FPE} and standard deviation σ_{FPE} of FPE provide a good description of the performance of a pitch estimation method. Another metric, the root mean square error (RMSE) as given by

$$RMSE = \sqrt{\frac{\sum_{t=1}^{O_v}\left[\frac{(F_0{}^{ref}(t)-\widetilde{F}_0(t))}{F_0{}^{ref}(t)}\times 100\right]^2}{O_v}}, \tag{2.32}$$

is the measure of error in percentage in the pitch estimates of all the O_v voiced frames in an utterance.

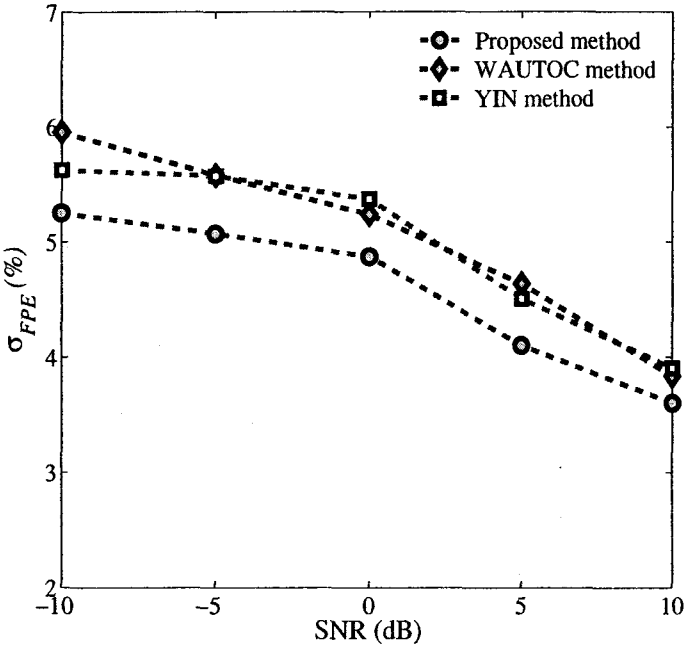


Figure 2.10: σ_{FPE} (%) as a function of SNR for all female and male speakers in white noise.

For the purpose of comparison, we use state-of the-art methods WAUTOC [46] and YIN [31], which have been proposed as improvements to the conventional methods given in [15], [25]. We have implemented the WAUTOC method independently using the parameters specified therein. We have downloaded the YIN software from the author's homepage and run it using the default parameters provided. The performance of the proposed method as well as that of the WAUTOC and YIN methods is evaluated at different levels of SNR in terms of the pitch estimates of the voiced frames based on the voiced/unvoiced labels provided by the *Keele* database.

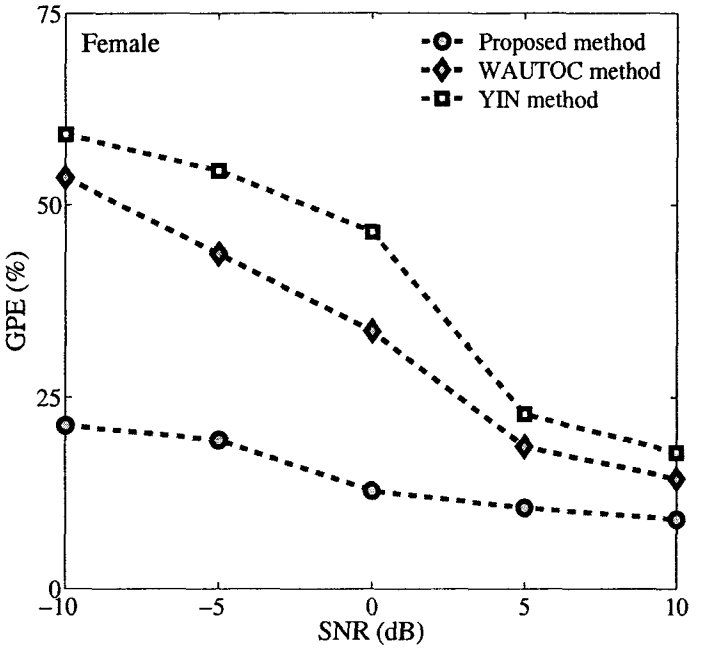


Figure 2.11: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in babble noise.

2.6.2 Simulation Results and Comparisons

(a) Results on white-noise corrupted speech: The pitch estimation performance of the WAUTOC, YIN and the proposed HCAC-HNM methods for the speech signals of the female (5) and male (5) speakers of the database are investigated in the presence of white noise. Figs. 2.5 and 2.6 show the PGPE values as a function of SNR obtained from the three methods for the female and male speaker groups, respectively, where the SNR varies from a very low value of -10 dB to a high value of 10 dB. It is seen from these figures that the proposed HCAC-HNM method exhibits a superior performance with respect to the PGPE metric at all the levels of SNR. In particular, it is evident from these figures that, for the levels of SNR equal to or

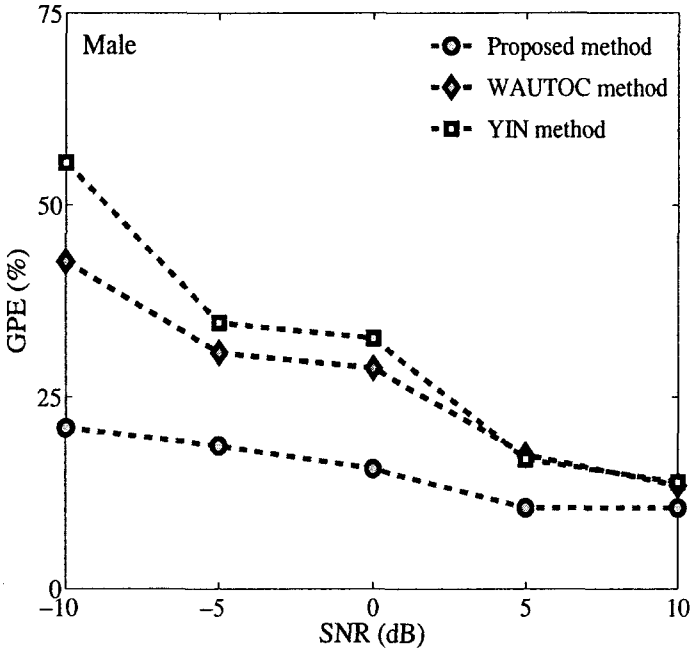


Figure 2.12: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in babble noise.

greater than 0 dB, the PGPE values resulting from the proposed method are very small but the WAUTOC and the YIN methods give much higher values of PGPE in this range. It is also seen from these figures that even at an SNR value as low as -10 dB, when the other two methods produce unsatisfactory results, the proposed method provides acceptable performance.

Figs. 2.7 and 2.8 depict the values of RMSE as a function of SNR obtained by using the three methods for the same female and male speaker groups as above. It is observed from these figures that for the SNR levels of 0 dB or above, the YIN and WAUTOC methods provide similar performances but the RMSE values obtained from the proposed method are much lower. Moreover, the proposed method performs much better for low levels (as low as -10 dB) of SNR.

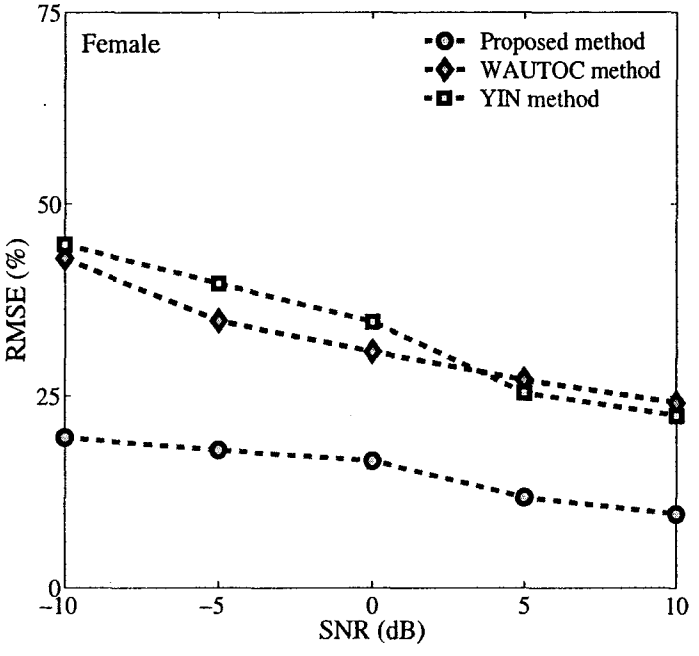


Figure 2.13: RMSE (%) as a function of SNR for female speaker group in babble noise.

Fig. 2.9 and Fig. 2.10 show, respectively, the mean m_{FPE} and standard deviation σ_{FPE} obtained by using the three methods for the set of 10 mixed (5 female plus 5 male) speakers of the database. Clearly, the overall mean and standard deviation values of the fine-pitch errors resulting from the proposed HCAC-HNM method are lower in comparison to that obtained by the other methods in the entire range of the SNR levels considered.

Significantly lower values of the PGPE and RMSE achieved by using the proposed method under a wide range of SNR levels, along with lower values of the overall mean and standard deviation of the fine-pitch errors, indicate its superior ability for an accurate pitch estimation and its high degree of robustness.

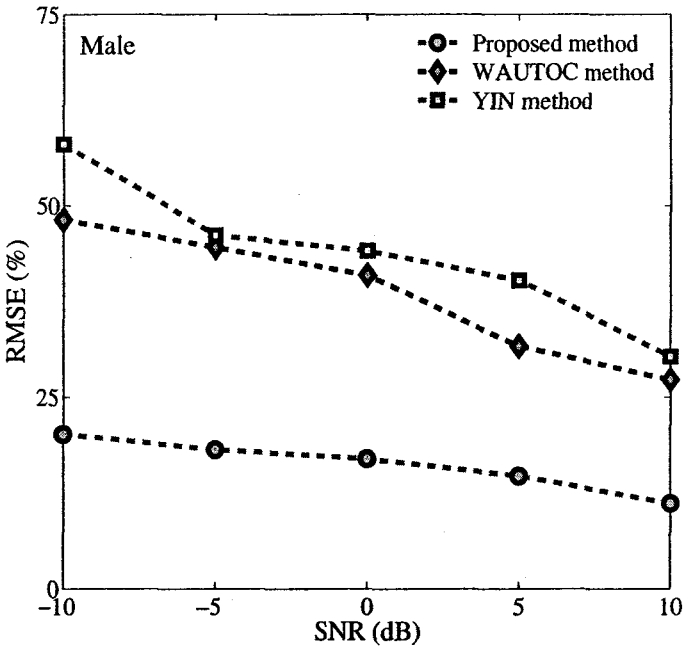


Figure 2.14: RMSE (%) as a function of SNR for male speaker group in babble noise.

(b) Results on Multi-Talker Babble-Noise Corrupted Speech: We now study the robustness of the proposed HCAC-HNM and the other two methods in the presence of babble noise, which is a non-Gaussian non-stationary colored noise. The pitch estimation results obtained from the babble-noise-corrupted speech in terms of the PGPE, RMSE, mean m_{FPE}, and standard deviation σ_{FPE} for each of the methods are portrayed in Fig. 2.11 through Fig. 2.16. It can be seen that the performance of all the three methods degrades in the presence of babble noise compared to that in the white noise due to presence of the harmonic components in the babble noise itself. However, the proposed HCAC-HNM method retains its superiority with respect to all the four performance metrics at all the levels of SNRs for the same male and female speaker groups as the ones considered for the white-noise corrupted speech.

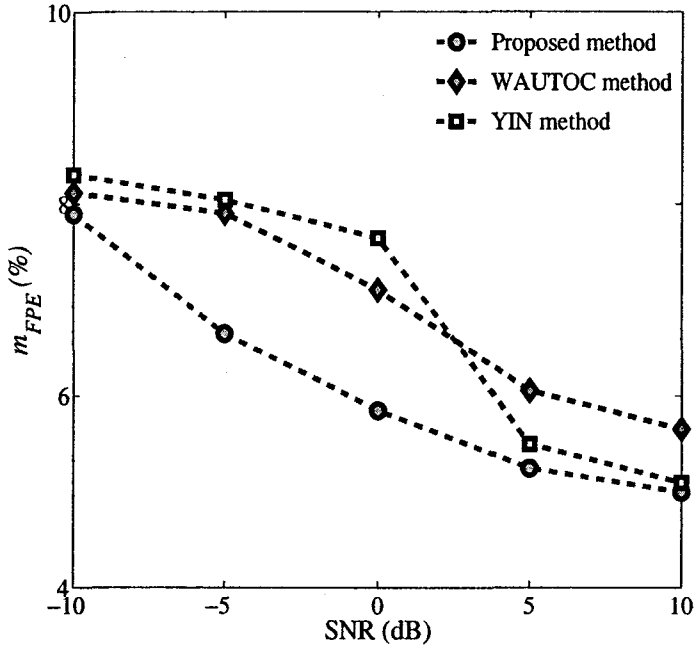


Figure 2.15: m_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

Figs. 2.11 and 2.12 provide plots for the PGPE values as a function of SNR obtained from the three methods for the female and male speaker groups. It is seen that in contrast to the white noise case, it is only the proposed method that maintains satisfactory performance at 0 dB or above. Also, the PGPE values achieved by the proposed method still remain acceptable for low levels of SNR (as low as -10 dB).

The RMSE resulting from using the three pitch estimation methods for the female and male speaker groups is shown in Figs. 2.13 and 2.14, respectively. As seen, the proposed HCAC-HNM method continues to provide quite good results even at a very low SNR of -10 dB, whereas the performance of the other two methods deteriorates significantly at SNR levels below 10 dB.

As an illustration of the accuracy of the proposed pitch estimation method, the

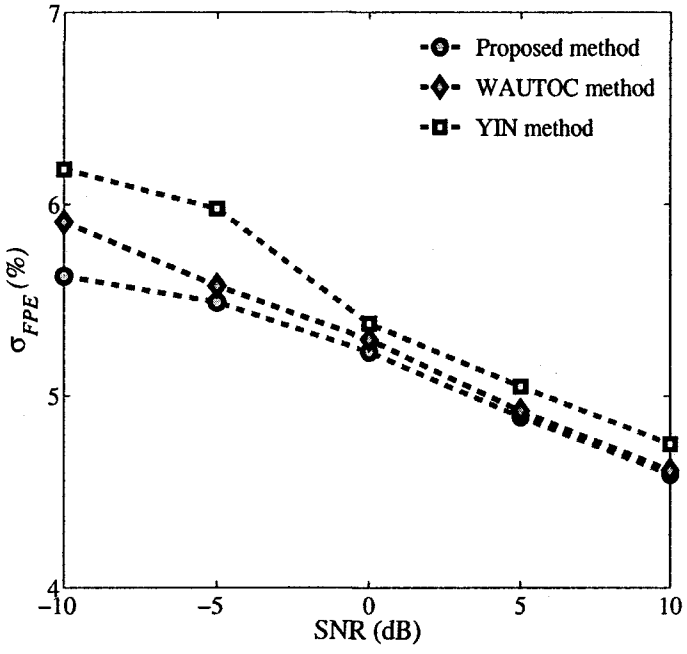


Figure 2.16: σ_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

mean m_{FPE} and standard deviation σ_{FPE} obtained from the three methods as a function of SNR are plotted in Figs. 2.15 and 2.16, respectively, for the same set of 10 mixed (5 female plus 5 male) speakers as the one used in Figs. 2.9 and 2.10. As expected, the estimation accuracy of the proposed HCAC-HNM method is reduced in comparison to that of the case of the white noise corruption, but its performance still remains considerably better than that provided by the other two methods for a wide range of SNR from high (10 dB) to very low (−10 dB).

It has been observed from the experimental results of the proposed method in the presence of white or Babble noise that, the errors between the true and estimated pitch values are randomly distributed over the time among different frames. However, as expected, the magnitude of an error in transition region and weakly voiced frames is

found higher in comparison to that obtained in the strongly voiced frames. Moreover, it has been found that the errors do not exhibit a tendency of being biased towards positive or negative value.

Here, we evaluate the net effect of pitch tracking by using DP on the pitch estimation performance of the proposed HCAC-HNM method. It is to be mentioned that, in the WAUTOC method, the Lagrange's method of three-point interpolation operation around the detected maximum peak of the weighted autocorrelation function is used for improving the accuracy of pitch extraction [46]. In the YIN method, parabolic interpolation as well as a procedure, which is reminiscent of median smoothing or dynamic programming is performed to obtain a best local pitch estimate [31]. Fig. 2.17 shows a reference pitch contour accompanied by the spectrogram of a 1.4-second excerpt of the clean speech of a male speaker from the reference database. In the same figure, the pitch contours from other pitch estimation methods are also overlaid on the spectrograms of the white noise-corrupted speech. It is seen from the figure that in contrast to the other methods, the pitch contour resulting from the proposed method is comparatively smoother even at SNR of -10 dB. Similarly, Fig. 2.18 illustrates a comparison of the pitch contours resulting from the three methods for female speech corrupted by babble noise at SNR $= -10$ dB. From Fig. 2.18, it is noted that the proposed method yields a smoother contour even in the presence of babble noise. The pitch contours obtained from the three methods and shown in Figs. 2.17 and 2.18 have convincingly demonstrated that the double and half-pitch errors in the proposed method can be significantly reduced by the use of the proposed pitch tracking scheme.

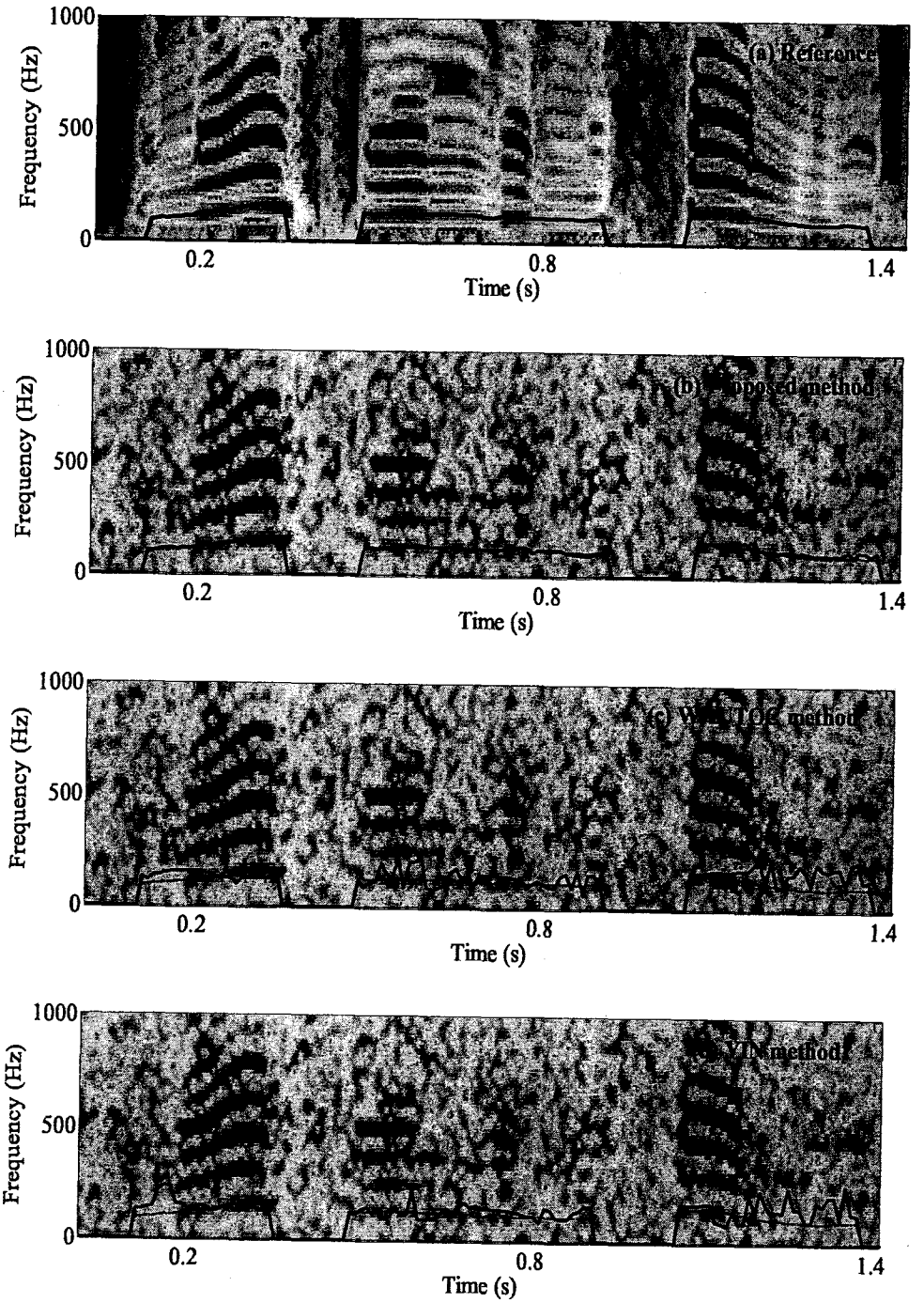


Figure 2.17: Pitch contours of different methods at SNR = -10 dB in white noise.

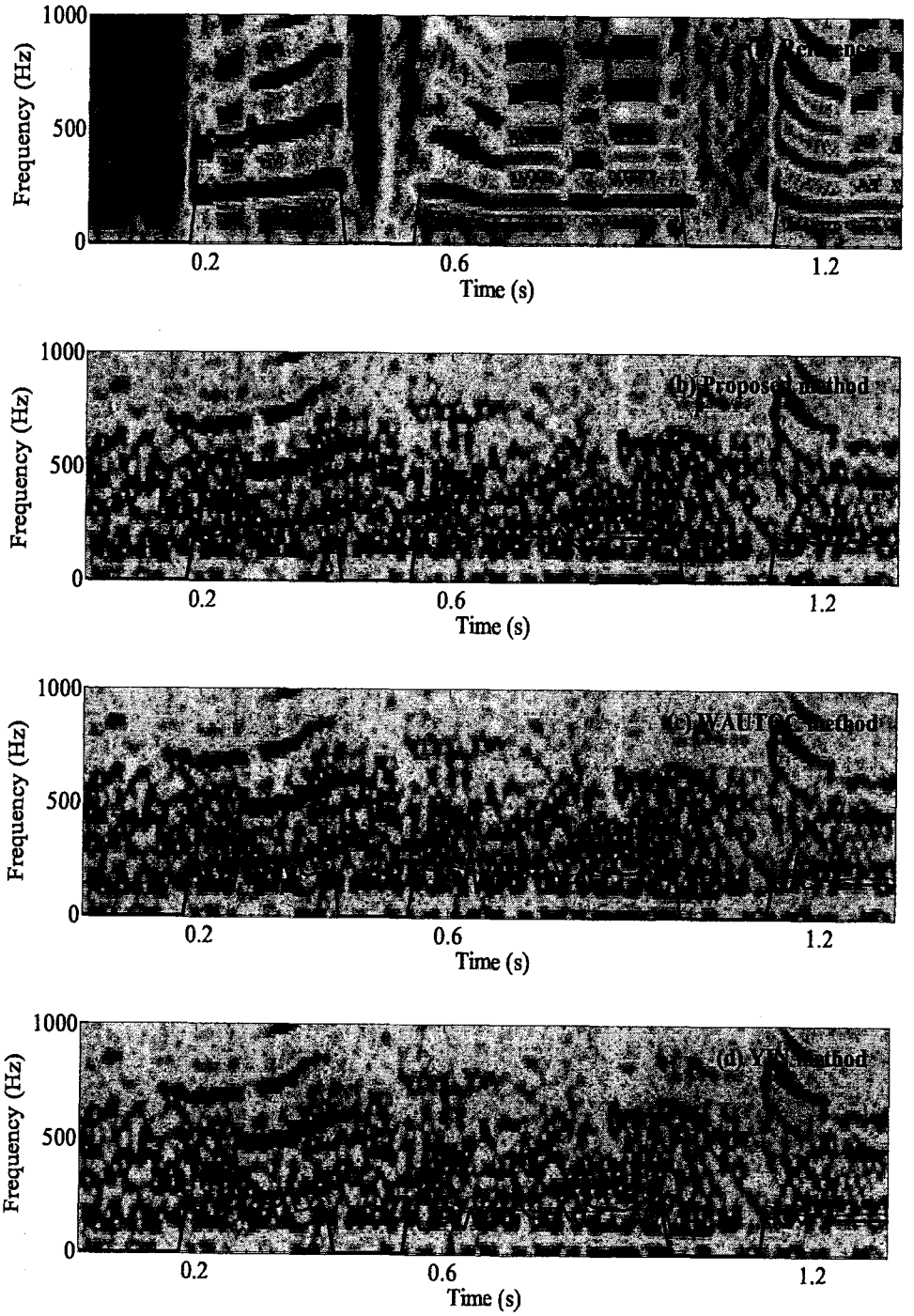


Figure 2.18: Pitch contours of different methods at SNR $= -10$ dB in babble noise.

Thus, the effect of pitch tracking using the DP technique applied to the proposed HCAC-HNM method renders its suitability for practical applications involving severely noise-corrupted speech.

2.7 Conclusion

In this Chapter, an effective methodology for the estimation of pitch from the observed speech signal in the presence of heavy additive noise has been presented. The proposed method is based on an autocorrelation model of clean speech, called the HCAC model, originally derived from a statistical autocorrelation estimator. The significant feature of this model is that it is expressed in terms of the pitch-harmonics of clean speech.

A least-squares autocorrelation-fitting optimization technique that employs the HCAC model has been presented for the estimation of a PH. The proposed optimization scheme has an advantage in that it provides the flexibility of incorporating some a priori knowledge of the PH, if available, in the process of pitch estimation. Since the presence of noise makes it difficult to determine an appropriate harmonic number, a new harmonic measure, which is based on an FFT based power spectrum of the noisy speech and governed by a noise-robust estimate of the PH, has been proposed to provide a possible set of harmonic numbers corresponding to the PH. Then, in order to obtain the true harmonic number desired for pitch estimation, a harmonic-to-noise power ratio (HPNR) for each harmonic number in the derived set has been formulated and maximized by employing a HNM scheme. A DP based pitch tracking scheme has also been presented to yield a smoothed pitch contour.

Extensive simulations have been carried out to demonstrate the performance of the proposed time-frequency domain method, referred to as HCAC-HNM. It has been shown that the method outperforms some of the state-of-the-art methods and is able

to estimate pitch with sufficient accuracy and consistency for both female and male speakers in noisy environments at very low levels of SNR. Moreover, the proposed method is shown to perform much better than some existing methods for multi-talker babble noise-corrupted speech signals; therefore, it is inferred that the HCAC-HNM method is readily suitable for real-life applications.

Chapter 3

Pitch Estimation Based on a Harmonic Sinusoidal Autocorrelation Model and Time-Domain Matching

3.1 Introduction

Considering that, in the derivation of the HCAC model in the previous chapter, the cross-product terms of different harmonics are neglected, in this chapter, a time-frequency domain pitch estimation method for speech observations under severe noisy conditions is presented based on an exact autocorrelation model of clean speech [67]. Starting from a conventional autocorrelation estimator, the key contribution here lies in the derivation of a simple yet accurate harmonic sinusoidal autocorrelation (HSAC) model of the clean speech in terms of its pitch-harmonics. We first propose to extract a pitch-harmonic (PH) from the discrete cosine transform (DCT)-preprocessed noisy speech by developing and using an autocorrelation domain least-squares (LS) fitting optimization technique that employs the proposed HSAC model. Then, an impulse-train with its period formed from the extracted PH and a symmetric average magnitude sum function (SAMSF) of the speech signal with its periodicity similar

to the pitch period are proposed to formulate an objective function. The maximization of the objective function through the matching of the periodicity of the impulse train with that of the SAMSF results in the desired harmonic number for pitch estimation. The advantage of the SAMSF based impulse-train matching (SIM) scheme is two-fold, namely, the period of the impulse train is governed by the noise-robust estimate of the PH; and several maxima, instead of a single global maximum, of the SAMSF are utilized in the matching. Thus, the "non-pitch-maxima" of the SAMSF are expectedly disregarded, making it easier to acquire the true harmonic number associated with the PH and therefore to obtain an accurate pitch estimate. Moreover, a dynamic programming based pitch tracking scheme is employed to obtain a physiologically plausible smoothed pitch contour as far as practical applications are concerned. To demonstrate the effectiveness of the proposed time-frequency domain method, referred to as, HSAC-SIM, an extensive simulation study with comparisons to some of the state-of-the-art techniques in the literature is conducted by considering natural speech signals of female and male speakers obtained from the *Keele* database in the presence of both white and multi-talker babble noises adopted from the *Noisex92* database. It is shown that the proposed method is capable of estimating pitch accurately and consistently for an SNR level as low as -10 dB. The superior pitch estimation performance of the proposed method for a multi-talker babble noise environment also is evidence of its suitability in practical applications.

The rest of the chapter is organized as follows. In Section 3.2, a brief overview of the proposed method is presented. We first show the derivation of the HSAC model of clean speech in Section 3.3. Then, a scheme for PH estimation from the DCT pre-processed noisy speech is described that utilizes the LS autocorrelation model-fitting optimization in conjunction with the proposed HSAC model. By exploiting the

extracted PH, an SAMSF based impulse-train matching (SIM) scheme is developed in Section 3.4 to determine the harmonic number with respect to the PH for pitch estimation. A dynamic programming (DP) based pitch tracking scheme employed to smooth the pitch contour is presented in Section 3.5. In Section 3.6, the pitch estimation performance of the proposed method is illustrated through detailed computer simulations using speech signals corrupted by both white and multi-talker babble noise. Finally, in Section 3.7, some of the distinctive features of this investigation reinforced by the experimental results are summarized.

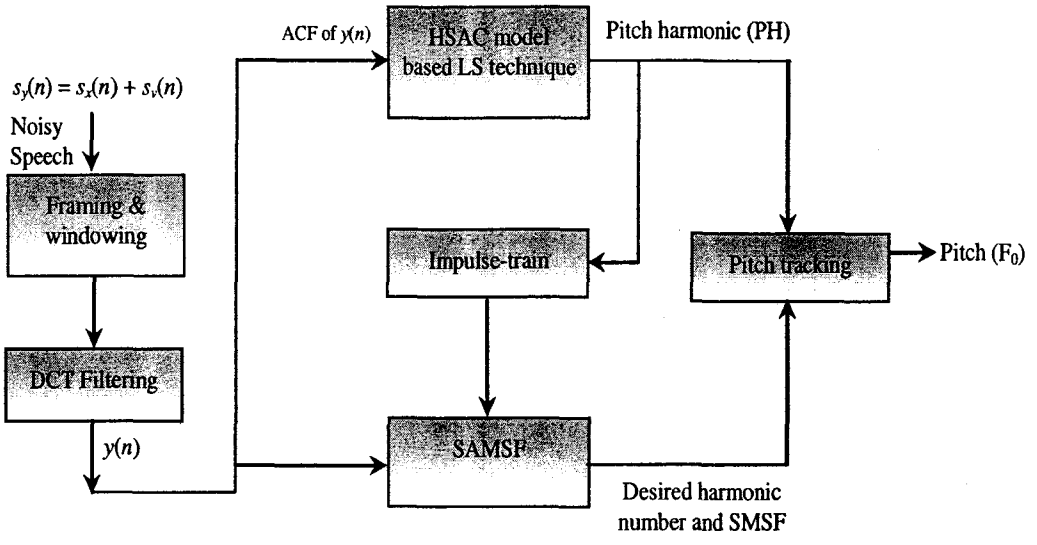
3.2 A Brief Description of the Proposed Method

An overview of the proposed pitch estimation method is shown by a block diagram in Fig. 3.1. Similar to the previous Chapter, each windowed frame of the observed noisy speech $s_y(n)$ is low-pass filtered to retain the frequency components up to the upper limit of the first formant band. A windowed low-pass filtered noisy speech frame can be expressed in the time-domain as

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

where $x(n)$ and $v(n)$ represent the windowed low-pass filtered clean speech $s_x(n)$ and noise $s_v(n)$, respectively. First, we derive an HSAC model for the pre-processed clean speech $x(n)$. Next, given the pre-processed noisy speech frame $y(n)$, we develop a novel technique for the extraction of one PH by exploiting the HSAC model and a new time-domain SIM scheme for determining the harmonic number corresponding to the extracted PH.

It is known that the DCT is far superior to the DFT for the transformation of real signals. For a real signal, the DFT gives a complex spectrum and leaves nearly



DCT: Discrete Cosine Transform, SAMSF: Symmetric Average Magnitude Sum Function

Figure 3.1: A block diagram representing the overview of the proposed pitch estimation method.

one-half the data unused. In contrast, the DCT generates a real spectrum of real signals and thereby makes the computation of redundant data unnecessary. As the DCT is derived from the DFT, all the desirable properties of DFT are preserved and fast algorithms for its computation also exist. The DCT offers an added advantage that it has a strong energy compaction property, which is helpful to obtain noise immunity. Due to such attractive features of the DCT over the DFT, a smoothed DCT power spectrum is exploited to obtain an initial estimate of the PH [68].

3.3 A Pitch-Harmonic Extraction using the HSAC Model

In this section, starting from the conventional autocorrelation estimator, we develop a simple yet accurate HSAC model of $x(n)$. The new model, represented in terms of

pitch-harmonics, is then employed to extract a PH, which is used for pitch estimation from the pre-processed noisy speech $y(n)$.

3.3.1 HSAC Model for Clean Speech

As described in Chapter 2, the clean speech $x(n)$ in a voiced frame is expressed as

$$x(n) = \sum_{p=1}^{\kappa} b_p \cos(n\omega_p + \varphi_p), \tag{3.2}$$

where the parameters b_p, φ_p and ω_p represent, respectively, the amplitude, phase and normalized angular frequency of the p-th harmonic of $x(n)$, and κ is the number of harmonics retained in the first formant band of the pre-processed speech. The normalized angular frequency ω_p of the p-th harmonic is related to the normalized angular pitch frequency ω_0 as

$$\omega_p = p\omega_0. \tag{3.3}$$

Note that $\omega_0 = \dfrac{2\pi F_0}{F_s}$, where F_0 and F_s are, respectively, the pitch frequency and the sampling frequency of $x(n)$ in Hz. Using the Euler formula, each real term $b_p \cos(n\omega_p + \varphi_p)$ of $x(n)$ can be expressed as the sum of two complex exponentials and thus $x(n)$ given by (3.2) can be rewritten as

$$x(n) = \sum_{p=1}^{\kappa} \frac{b_p}{2} \left[e^{j(n\omega_p + \varphi_p)} + e^{-j(n\omega_p + \varphi_p)} \right]. \tag{3.4}$$

By letting $\tau_p = \dfrac{b_p}{2} e^{j\phi_p}$ and $\zeta_p = e^{j\omega_p}$, we can express $x(n)$ as

$$x(n) = \sum_{p=1}^{\kappa} \left[\tau_p (\zeta_p)^n + \tau_p^* (\zeta_p^*)^n \right]. \tag{3.5}$$

Since $x(n)$ is real, it is seen from its representation in (3.5) that for a harmonic number p, the complex exponential $(\zeta_p)^n$ and the associated complex constant coefficient τ_p appears in complex conjugate pairs. Given a clean speech $x(n)$ in a frame,

the autocorrelation function (ACF) of $x(n)$ in a frame can be estimated conventionally as

$$\phi_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x(n)x(n+|m|), \quad m = 0, \pm 1, \ldots, \pm M, M < N, \tag{3.6}$$

where m is the discrete lag variable. Even though the frame length N is finite, the conventional ACF estimator $\phi_x(m)$ defined in (3.6) offers an efficient way to obtain a reasonably accurate estimate for the ACF of the speech signal $s_x(n)$. Substituting (3.2) into (3.6), $\phi_x(m)$ can be written as

$$\begin{aligned} \phi_x(m) &= \frac{1}{N} \sum_{n=0}^{N-1-|m|} \left[\sum_{p=1}^{\kappa} b_p \cos(n\omega_p + \varphi_p) \sum_{q=1}^{\kappa} b_q \cos\{(n+m)\omega_q + \varphi_q\} \right], \\ m &= 0, \pm 1, \ldots, \pm M, M < N. \end{aligned} \tag{3.7}$$

By substituting the expression for $x(n)$ given by (3.4) into (3.7), we get

$$\begin{aligned} \phi_x(m) &= \frac{1}{N} \sum_{n=0}^{N-1-|m|} \sum_{p=1}^{\kappa} \sum_{q=1}^{\kappa} \frac{b_p b_q}{4} T_{n_{pq}}, \\ T_{n_{pq}} &= \left[e^{j(n\omega_p + \varphi_p)} + e^{-j(n\omega_p + \varphi_p)} \right] \left[e^{j(n\omega_q + \omega_q m + \varphi_q)} + e^{-j(n\omega_q + \omega_q m + \varphi_q)} \right] \end{aligned} \tag{3.8}$$

By letting $\tau_q = \frac{b_q}{2} e^{j\phi_q}$ and $\zeta_q = e^{j\omega_q}$, and recalling that $\tau_p = \frac{b_p}{2} e^{j\phi_p}$ and $\zeta_p = e^{j\omega_p}$, (3.8) can be rewritten as

$$\phi_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} \sum_{p=1}^{\kappa} \sum_{q=1}^{\kappa} [\Pi_1 + \Pi_2 + \Pi_1^* + \Pi_2^*], \quad m = 0, \pm 1, \ldots, \pm M, M < N \tag{3.9}$$

where

$$\Pi_1 = \tau_p \tau_q (\zeta_p \zeta_q)^n \zeta_q^m, \quad \Pi_2 = \tau_p^* \tau_q (\zeta_p^* \zeta_q)^n \zeta_q^m. \tag{3.10}$$

By interchanging the sums in (3.9) and grouping the terms containing ζ_q^m and $(\zeta_q^m)^*$, the expression for the ACF of $x(n)$ can finally be written as

$$\phi_x(m) = \sum_{q=1}^{\kappa} \left[\lambda_q \zeta_q^m + \lambda_q^* (\zeta_q^m)^* \right], \quad m = 0, \pm 1, \ldots, \pm M, M < N, \tag{3.11}$$

where λ_q can be shown to be given by

$$\lambda_q = \frac{1}{N}\left[\tau_q \sum_{p=1}^{\kappa} \tau_p \sum_{n=0}^{N-1-|m|} (\zeta_p \zeta_q)^n + \tau_q \sum_{p=1}^{\kappa} \tau_p^* \sum_{n=0}^{N-1-|m|} (\zeta_p^* \zeta_q)^n\right]. \tag{3.12}$$

Comparing (3.11) and (3.5), it is seen that the expressions for the ACF $\phi_x(m)$ of the signal and the signal $x(n)$ itself have the same complex exponential format with the difference being that for $\phi_x(m)$ the constant co-efficient factor associated with the term ζ_q^m is λ_q , whereas for $x(n)$ the factor associated with the term ζ_p^n is τ_p. Thus, the ACF $\phi_x(m)$ preserves the frequency characteristics of $x(n)$. Letting, $\lambda_q = P_q e^{j\theta_q}$, and recalling that $\zeta_q = e^{j\omega_q}$, (3.11) can be expressed in the following form as

$$\begin{aligned}
\phi_x(m) &= \sum_{q=1}^{\kappa} \left[P_q e^{j(\omega_q m + \theta_q)} + P_q e^{-j(\omega_q m + \theta_q)}\right], \quad m = 0, \pm 1, \ldots, \pm M, M < N \\
&= \sum_{q=1}^{\kappa} 2P_q \cos(\omega_q m + \theta_q) \\
&= \sum_{q=1}^{\kappa} \left[\alpha_q \cos(\omega_q m) + \beta_q \sin(\omega_q m)\right],
\end{aligned} \tag{3.13}$$

where $\alpha_q = 2P_q \cos\theta_q$ and $\beta_q = 2P_q \sin\theta_q$ are constants depending only on λ_q . In deriving (3.13), the contributions from the cross-product terms of different harmonics have not been neglected. We refer to (3.13) as the harmonic sinusoidal autocorrelation (HSAC) model of $x(n)$. It is seen that the above model is expressed explicitly in terms of the pitch-harmonics, ω_q's of $x(n)$.

3.3.2 HSAC Model Fitting: Least-Squares Optimization

In this subsection, we present the proposed technique of extracting a PH by fitting the ACF of $y(n)$ with the HSAC model of $x(n)$ derived in the previous subsection using the LS minimization technique [51].

In the presence of noise, for a pre-processed noisy speech frame $y(n)$, an estimate of ACF $\phi_y(m)$ can be computed using (3.6) as

$$\phi_y(m) = \sum_{n=0}^{N-1-|m|} y(n)y(n+m), \quad m = 0, \pm 1, \ldots, \pm M, M < N. \tag{3.14}$$

Using (3.1), (3.14) can be re-written as

$$\phi_y(m) = \phi_x(m) + \phi_v(m) + \phi_c(m), \tag{3.15}$$

where $\phi_c(m)$ represents the summation of all the cross-correlation terms. The property of $\phi_y(m)$ in retaining the peaks of $\phi_x(m)$ at multiples of pitch period even under severe noisy conditions motivates us to utilize it for extracting one PH ω_q, which in our method would be sufficient to estimate the pitch.

It is seen that while the HSAC model of $x(n)$, as given by (3.13), does not contain the perceptually less important phase information, it does preserve the pitch information of $x(n)$ contained in (3.2) in terms of the pitch-harmonics ω_q's. Any one of the κ components in the summation of (3.13), say the q-th one,

$$\phi_x^{(q)}(m) = \alpha_q \cos(\omega_q m) + \beta_q \sin(\omega_q m) \tag{3.16}$$

and the associated PH ω_q can be estimated from the M lags of the available noisy ACF $\phi_y(m)$ as computed according to (3.14) by employing a LS fitting technique. For the PH ω_q, the corresponding parameters α_q and β_q are determined such that the total squared error between $\phi_y(m)$ and the component $\phi_x^{(q)}(m)$ spanned over all the M lags as given by

$$\Lambda(\alpha_q, \beta_q) = \sum_{m=1}^{M} \left[\phi_y(m) - \phi_x^{(q)}(m)\right]^2 \tag{3.17}$$

is minimized in the least-squares sense. Since we are interested to determine only one PH ω_q, it is sufficient to consider one component $\phi_x^{(q)}(m)$ of the HSAC model in the

above autocorrelation-fitting technique. In (3.17), the zero-th lag of $\phi_y(m)$ has been excluded, since at this lag the effect of noise is maximum. For a given ω_q, one can uniquely find the optimum values of α_q and β_q that minimize the function $\Lambda(\alpha_q, \beta_q)$ given by (3.17) by equating the partial derivatives of $\Lambda(\alpha_q, \beta_q)$ with respect to α_q and β_q to zero, that is,

$$\frac{\partial \Lambda(\alpha_q, \beta_q)}{\partial \alpha_q} = 0, \quad \frac{\partial \Lambda(\alpha_q, \beta_q)}{\partial \beta_q} = 0. \tag{3.18}$$

The above equation yields the following system of linear equations:

$$\alpha_q \sum_{m=1}^{M} \cos^2(\omega_q m) + \beta_q \sum_{m=1}^{M} \sin(\omega_q m)\cos(\omega_q m) = \sum_{m=1}^{M} \phi_y(m)\cos(\omega_q m) \tag{3.19}$$

$$\alpha_q \sum_{m=1}^{M} \cos(\omega_q m)\sin(\omega_q m) + \beta_q \sum_{m=1}^{M} \sin^2(\omega_q m) = \sum_{m=1}^{M} \phi_y(m)\sin(\omega_q m). \tag{3.20}$$

The LS fitting scheme is based on the assumption of an *a priori* knowledge of the value of a pitch-harmonic ω_q. However, at this stage, this assumption is not fully valid. As a matter of fact, our objective is to determine an accurate value of one of the pitch-harmonics ω_q's. For this purpose, we follow a similar way that is adopted to extract the PH by employing the HCAC model-fitting based optimization technique in Chapter 2.

It is well known that, in a voiced frame, the power spectrum of the clean speech $x(n)$ exhibits strong peaks at or near pitch-harmonics, ω_q's [54]. Considering the advantageous features of DCT in the case of real signals, such as $x(n)$ as described in Section 3.2, we analyze the power spectrum of the pre-processed noisy speech frame $y(n)$ based on the DCT, where the DCT $Y(k)$ of $y(n)$ is given by [69]

$$Y(k) = c_d(k) \sum_{n=1}^{N} y(n) \cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right]. \tag{3.21}$$

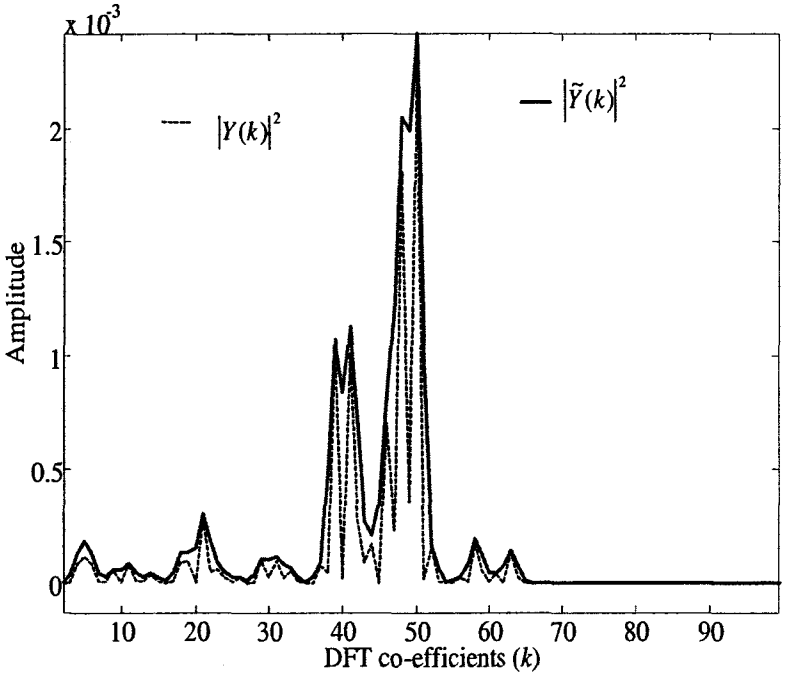


Figure 3.2: DCT power spectrum of $y(n)$ with and without smoothing.

In this equation, k is the frequency bin index and the co-efficient $c_d(k)$ is given by [70]

$$c_d(k) = \begin{cases} \sqrt{\frac{1}{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \leq k \leq N \end{cases} . \tag{3.22}$$

The frequency points of peak occurrence of the DCT power spectrum $|Y(k)|^2$ of $y(n)$ in the presence of a heavy noise may not be the same as those of $|X(k)|^2$, the DCT power spectrum of $x(n)$. In order to reduce the fluctuations around the strong peaks of $|Y(k)|^2$, it is necessary to smooth it. We accomplish this process as follows:

$$|\tilde{Y}(k)|^2 = o_k|Y(k)|^2 + (1 - o_k)\left[\frac{1}{(2D+1)}\sum_{u=-D}^{D}|Y(k-u)|^2\right], \tag{3.23}$$

where o_k is a smoothing factor and the length of the smoother has been chosen to be $(2D+1)$. Fig.3.2 shows the DCT power spectrum of $y(n)$ with and without smoothing.

It is seen from this figure that smoothed DCT power spectrum $|\tilde{Y}(k)|^2$ has peaks that are more prominent compared to that of $|Y(k)|^2$. However because of the underlying noise, the frequency point corresponding to the maximum peak of $|\tilde{Y}(k)|^2$ may not represent the exact position of a pitch-harmonic ω_q. Thus, this frequency point is taken only as an initial estimate $\omega_q^{(0)}$ for the pitch-harmonic ω_q.

For this initial estimate $\omega_q^{(0)}$, initial estimates $\alpha_q^{(0)}$ and $\beta_q^{(0)}$ of α_q and β_q can be easily obtained by solving the linear system given by (3.19) and (3.20). With these values of $\alpha_q^{(0)}$ and $\beta_q^{(0)}$, the estimate $\phi_x^{(q)(0)}(m)$ of the component $\phi_x^{(q)}(m)$ of the HSAC model can then be computed by using (3.16). Finally, the minimum value of the total squared error $\Lambda^{(0)}$ corresponding to the initial estimate $\omega_q^{(0)}$ is determined using (3.17). With the initial estimates $\omega_q^{(0)}$, $\alpha_q^{(0)}$, $\beta_q^{(0)}$, $\phi_x^{(q)(0)}(m)$ and $\Lambda^{(0)}$ at our disposal, we can now perform an iterative search for ω_q within a small neighborhood around $\omega_q^{(0)}$ with a reasonable resolution by using (3.19), (3.20), (3.16), and (3.17) in that order repeatedly. The outcome of this optimization procedure is an accurate estimate ω_{qopt} of the pitch-harmonic ω_q that best matches $\phi_x^{(q)}(m)$ with $\phi_y(m)$.

The PH extraction scheme of the proposed pitch estimation method as described in this subsection offers a two-fold advantage. Firstly, the scheme employs the HSAC model of clean speech that provides a direct relationship between pitch-harmonics and the pitch. Secondly, unlike the approach which relies only on a single lag corresponding to the global maximum of the noisy ACF $\phi_y(m)$ [45], we employ relatively a large number (M) of lags of $\phi_y(m)$ for its LS matching with the noise-free component $\phi_x^{(q)}(m)$ of the HSAC model. The novelty of the PH estimation scheme lies in its ability to extract the optimum value of the PH ω_{qopt} associated with the $\phi_x^{(q)}(m)$ component of the HSAC model from the available noisy ACF $\phi_y(m)$ through a model-fitting based approach instead of using the noisy ACF $\phi_y(m)$ directly.

3.4 Determination of Harmonic Number using the SIM Scheme

In this Section, we present a scheme for determining a harmonic number so that this number along with the PH obtained in the previous Section can be used to find an estimate of the pitch, since the pitch-harmonic ω_{qopt} and the corresponding harmonic number q_{opt} is related as $\omega_{qopt} = q_{opt}\omega_0$. For this purpose, we proceed as follows.

3.4.1 Proposed Symmetric Average Magnitude Sum Function

Corresponding to a voiced speech signal, we propose a new time domain average magnitude sum function as

$$\xi_x(m) = \sum_{n=0}^{Q-1} |x(l) + x(n)|, \quad m \in [0, 1, \ldots, (Q-1)] \tag{3.24}$$

where

$$l = mod(n + m, Q) \tag{3.25}$$

and Q is the number of speech samples employed to compute $\xi_x(m)$ for every lag m. In the evaluation of (3.24), we choose $Q = N + m_{max}$, where m_{max} is the maximum possible pitch period of human speech. Since pitch usually changes slowly with time, such a choice of Q ensures a duration in which the pitch parameter would not change significantly and at the same time at least two pitch periods would be covered to provide periodicity information for a wide range of speakers. By analyzing the terms on the right side of (3.24), we observe that $\xi_x(m)$ has the following properties :

1. At $m = 0$, for each term $|x(l) + x(n)|$ of the summation of $\xi_x(m)$, we have $l = mod(n, Q) = n$ as $n \in [0 : (Q-1)]$ and $x(l) = x(n)$. Hence, $\xi_x(m)$ gives its maximum value at $m = 0$.

2. According to (3.25), by designing l in each term $|x(l)+x(n)|$ of $\xi_x(m)$ as the modulo of $(n+m)$ with respect to Q, it can be shown that for $m \in [1 : \{\frac{Q}{2}-1\}]$, $\xi_x(m) = \xi_x(Q-m)$. Thus $\xi_x(m)$ has an even symmetry about $m_s = \frac{Q}{2} = \frac{(N+m_{max})}{2}$. As such, it is sufficient to compute $\xi_x(m)$ for $0 \leq m \leq m_s$ only. In view of this symmetry property, $\xi_x(m)$ as defined by (3.24) is referred to as the symmetric average magnitude sum function (SAMSF) of $x(n)$.

3. We know that, for a voiced speech $x(n)$ with pitch period T_0, $x(n)$ is similar with $x(n+\rho T_0)$, where ρ is a nonnegative integer, and that the vocal tract response decays exponentially with n within the pitch period T_0. Using these facts and since, $x(l) = x(mod(n+m, Q)) = x(n+m)$ for $(n+m) < Q$, it can be shown that at $m = \rho T_0$,$Q' = (Q-\rho T_0)$ of the Q terms in (3.24) for which $(n+\rho T_0) < Q$, $x(l) = x(n)$. On the other hand, for $\rho T_0 < m < (\rho+1)T_0$, in each of the Q terms, $x(l) \neq x(n)$. As a result, a term $|x(l)+x(n)|$ for $m \neq \rho T_0$ has a smaller magnitude compared to the corresponding term when $m = \rho T_0$. Thus, we have $\xi_x(\rho T_0) > \xi_x(\rho T_0 + m_b)$, where $0 \leq (\rho T_0 + m_b) \leq m_s, 0 < m_b < T_0$. Also, the function $\xi_x(m)$ as defined by (3.14) exhibits a local maxima at $m = \rho T_0$, $\rho = 0, 1, \ldots, (\left\lfloor \frac{m_s}{T_0} \right\rfloor)$ indicating that $\xi_x(m)$ possesses a periodicity similar to that of $x(n)$, that is, $\xi_x(m+\rho T_0) \approx \xi_x(m)$.

4. In the calculation of $\xi_x(\rho T_0)$, the contribution of $Q' = (Q-\rho T_0)$ terms is much larger than those of the remaining terms, where $x(l) \neq x(n)$. This results in the local maximum of $\xi_x(m)$ to decrease when m increases with increasing multiples of T_0, that is, $\xi_x(\rho T_0) > \xi_x((\rho+1)T_0)$.

5. It is seen from (3.6) that a total of $(N-1-|m|)$ terms to be added for computing an ACF value reduces as the lag variable m increases. On the other hand, for

the computation of an $\xi_x(m)$ value as given by (3.24), the number of terms $Q = (N + m_{max})$ is independent of m. As a result, the maxima of $\xi_x(m)$, in contrast to that of ACF, decays at a much slower rate.

The properties of SAMSF discussed above concern clean speech signals. However, in practical situations, one has to deal with speech signals corrupted by noise. Given Q samples of a pre-processed noisy speech signal $y(n)$, its SAMSF $\xi_y(m)$ can be expressed as [71]

$$\xi_y(m) = \sum_{n=0}^{Q-1} |y(l) + y(n)|, \quad m \in [0, 1, \ldots, m_s]. \tag{3.26}$$

Using the expression of $y(n)$ given by (3.1) in the above equation, we have

$$\xi_y(m) = \sum_{n=0}^{Q-1} |\,[x(l) + x(n)] + [v(l) + v(n)]\,|, \quad m \in [0, 1, \ldots, m_s], \tag{3.27}$$

from which we get the following inequality

$$\xi_y(m) \leq \sum_{n=0}^{Q-1} |x(l) + x(n)| + \sum_{n=0}^{Q-1} |v(l) + v(n)| = \xi_x(m) + \xi_v(m), \tag{3.28}$$

where $\xi_x(m)$ is as given by (3.24) and

$$\xi_v(m) = \sum_{n=0}^{Q-1} |v(l) + v(n)|, \quad m \in [0, 1, \ldots, m_s] \tag{3.29}$$

is the SAMSF of the pre-processed noise. By introducing an equality term $\varepsilon(m) \geq 0$ in (3.28), an expression for $\xi_y(m)$ can be rewritten as

$$\xi_y(m) = \xi_x(m) + \xi_v(m) - \varepsilon(m) = \xi_x(m) + \Gamma(m), \quad m \in [0, 1, \ldots, m_s], \tag{3.30}$$

where $\Gamma(m) = \xi_v(m) - \varepsilon(m)$ represents the error introduced in the presence of noise and it vanishes when $v(n) \equiv 0$. Since $\xi_v(m) \geq 0$ and $\varepsilon(m) \geq 0$, $\Gamma(m)$ is smaller than $\xi_v(m)$. We will now conduct experiments to examine and illustrate the behavior of

$\xi_y(m)$ in comparison to that of $\xi_x(m)$. We use voiced frames of different strengths corrupted by different types and levels of noises as a platform of this experiment. In Fig. 3.3, we show examples of the typical behavior of $\xi_y(m)$ observed from our experiment. Plots in Figs. 3.3(a) and (b) correspond to a strongly voiced frame at SNR $= -5$ dB and those in Figs. 3.3(c) and (d) correspond to a weakly voiced frame at SNR $= -10$ dB. The type of noise embedding the speech is white in the case of Figs. 3.3(a) and (c), whereas it is a multi-talker Babble noise in the case of Figs. 3.3(b) and (d). It is evident from these figures that peaks of $\xi_x(m)$ at multiples of the pitch period are well-preserved and remain prominent in $\xi_y(m)$ whether the frame under consideration is strongly or weakly voiced, or whether the corruption is due to the presence of a white or a Babble noise regardless of its level. This property of $\xi_y(m)$ in retaining pitch-harmonic peaks even under severe noisy conditions motivates us to utilize several maxima instead of only a single maximum of $\xi_y(m)$ for the determination of the harmonic number corresponding to the PH extracted in the previous subsection. In what follows, by exploiting $\xi_y(m)$ as a template function, a scheme for matching the periodicity of $\xi_y(m)$ and that of a periodic impulse-train is developed to acquire the desired harmonic number for pitch estimation.

3.4.2 Proposed Impulse-Train

First, a submultiple T_{qopt} (an integer) of the pitch period T_0 can be obtained as

$$T_{qopt} = \left\lceil \frac{2\pi}{\omega_{qopt}} \right\rceil \tag{3.31}$$

where ω_{qopt} is the optimum value of the extracted PH and $\omega_q = q\omega_0$.

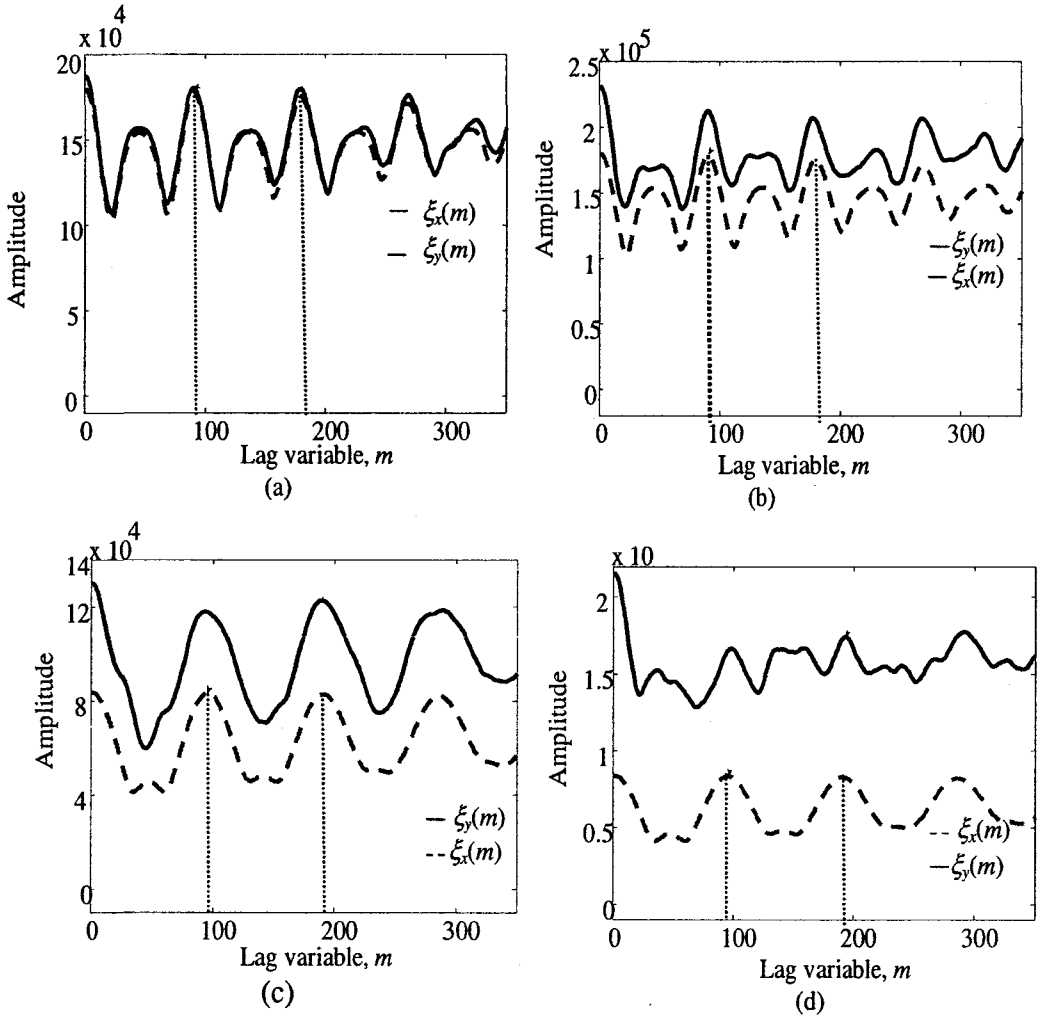


Figure 3.3: Plots for $\xi_y(m)$ and $\xi_x(m)$ for typical voiced frames considering different levels and types of noise and strengths of voiced frames. A strongly voiced frame at SNR $= -5$ dB under (a) white noise and (b) Babble noise. A weakly voiced frame at SNR $= -10$ dB under (c) white noise and (d) Babble noise.

An impulse train $I(m,q)$ with a variable period given by $T_I = qT_{qopt}$ can be formulated as [50]

$$I(m,q) = \sum_{t_I=0}^{I_T-1} \delta(m - t_I T_I), \quad m \in [0,1,\ldots,m_s] \tag{3.32}$$

where $\delta(m)$ is the Kronecker delta function and I_T the number of unit impulses used in the impulse train. By changing the value of q, the period of T_I the impulse train $I(m,q)$ can be varied. We can now find the optimum value of q for which the period of the impulse train matches the true periodicity of the SAMSF $\xi_y(m)$. For this purpose, we define the following objective function:

$$\eta(q) = \sum_{m=0}^{m_s} I(m,q)\xi_y(m) \tag{3.33}$$

where the impulse train $I(m,q)$ is weighted by the SAMSF $\xi_y(m)$ given in (3.26). Since $\xi_y(m)$ exhibits maxima at integral multiples of T_0 , the objective function $\eta(q)$ is maximized when a value of q is chosen for which the period T_I of the impulse train gets synchronized with the true pitch period T_0 [72]. This optimum value of q can be written as

$$q_{opt} = arg \ \max_q [\eta(q)] . \tag{3.34}$$

The parameter q is an integer assuming a value for which $m_{min} < (I_T - 1)T_I \leq m_s$, where m_{min} is the minimum possible pitch period of human speech. The method of determining the optimum harmonic number q_{opt} corresponding to ω_{qopt} is referred to as SAMSF-weighted impulse-train matching (SIM) scheme.

Thus for a voiced frame, q_{opt} leads to the desired estimate of the pitch period as

$$\tilde{T_0} = q_{opt} T_{qopt} \tag{3.35}$$

where T_{qopt} is obtained from (3.31). An estimate of pitch frequency (F_0) in Hz can

be determined as

$$\tilde{F}_0 = \frac{F_s}{\tilde{T}_0}. \tag{3.36}$$

It is of interest to note that in comparison to the method in [24], which is based on the representation of voiced speech as a sum of harmonic signals, the proposed method requires extracting only one harmonic component and the associated pitch-harmonic, implying that estimation of the unknown parameters related to all other harmonics is not needed. The complete method for pitch estimation whose development started in the Section 3.2 will henceforth be referred to as the HSAC-SIM method.

3.5 SAMSF Based Pitch Tracking using Dynamic Programming

Usually, the pitch (F_0) of a speech signal shows a smooth behavior over time that can be exploited to optimize the overall F_0 contour after its estimation at each time frame. Given a set of potential candidates for the pitch at each frame, a common requirement for F_0 tracking is to find the optimal pitch path connecting one candidate per frame through the set of pitch candidates over all time frames. In this Section, a pitch tracking scheme proposed in Chapter 2 can easily be employed to reduce the error in the F_0 contour. It is important to keep in mind the obvious changes required for candidate generation and cost computation due to the different approach of the development of the HSAC-SIM method.

In the pitch tracking scheme of this Chapter, the pitch period $\tilde{T}_0 = q_{opt}T_{qopt}$ as given in (3.35), determined by using the SIM scheme, is selected to be one of the members of the set of potential pitch candidates for a frame. For a given frame t, a certain number, say $(J_t - 1)$, of local maxima of the SAMSF $\xi_y(m)$ given by (3.26) lying within the pitch period range but excluding the one located at or nearest to

the lag $m = \tilde{T}_0$ are chosen, and the corresponding locations of these maxima are selected as the other possible pitch candidates. Among the J_t number of potential candidates at frame t, $\tilde{T}_0$ is assigned the highest priority to be the correct pitch value for the frame under consideration. The remaining $(J_t - 1)$ candidates are prioritized by sorting them according to decreasing magnitude of the corresponding values of $\xi_y(m)$. By letting $F_{0,t}^j$ represent the j-th pitch frequency candidate in the t-th frame, the local cost for the candidate $F_{0,t}^j$ is defined as

$$C_{local}(F_{0,t}^j) = -\eta(F_{0,t}^j), \tag{3.37}$$

where $\eta(F_{0,t}^j) = \sum_{m=0}^{m_s} I(m, F_{0,t}^j)\xi_y(m)$ is the objective function corresponding to $F_{0,t}^j$, as formulated in (3.33). It is noted that $I(m, F_{0,t}^j)$ refers to the impulse train given by (3.32) with its period, $T_I = \left\lceil \frac{F_s}{F_{0,t}^j} \right\rceil$. Obviously, the pitch frequency candidate having a higher score of the objective function $\eta(F_{0,t}^j)$ will result in a lower local cost $C_{local}(F_{0,t}^j)$. Due to the deviation among the pitch candidates from one frame to next, a transition cost is computed as

$$C_{tran}(F_{0,t-1}^i, F_{0,t}^j) = \left| \log \frac{F_{0,t-1}^i}{F_{0,t}^j} \right| \tag{3.38}$$

which measures the cost of the pitch path going from the i-th candidate of frame $(t-1)$ to the j-th candidate of frame t. Considering that there are O_c consecutive frames, i.e., $t = 1, 2, \ldots, O_c$, the total cost function $C(A)$ corresponding to an arbitrarily chosen trajectory $A = \left\{F_{0,1}^{j_1}, F_{0,2}^{j_2}, \ldots, F_{0,O_c}^{j_{O_c}}\right\}$ is defined as

$$C(A) = C_{local}(F_{0,1}^{j_1}) + \sum_{t=2}^{O_c} \left[\varpi . C_{tran}(F_{0,t-1}^{j_{t-1}}, F_{0,t}^{j_t}) + C_{local}(F_{0,t}^{j_t})\right]. \tag{3.39}$$

In (3.39), $F_{0,t}^{j_t}$ means one of the available candidates from all $F_{0,t}^j$ at the t-th frame to realize a path and ϖ is a weighting factor balancing the local and transition costs. A

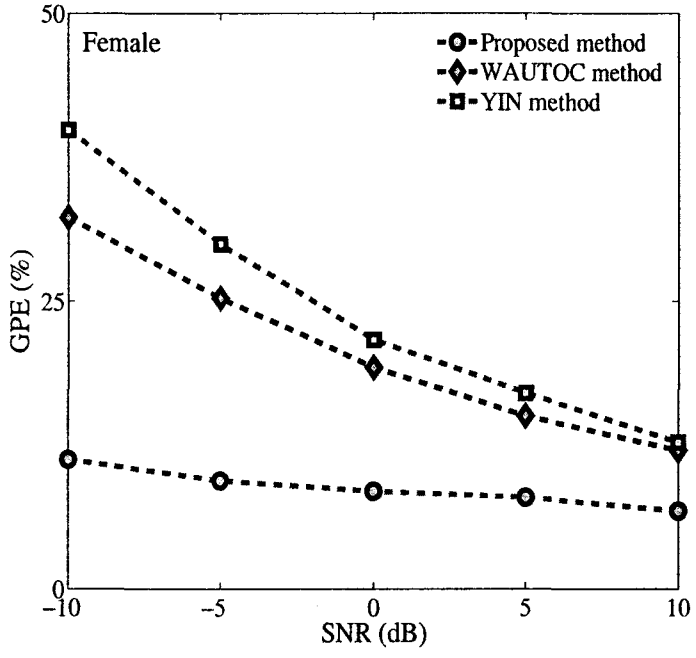


Figure 3.4: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in white noise.

large value of ϖ in general gives a better continuity of the pitch contour. The task of pitch tracking is now to minimize the total cost function, which can be performed through the recursive computation of the total cost function in stages. Due to the suitability of DP in handling such a problem, we employ DP for efficiently minimizing the total cost function for pitch tracking [48], [59], [60], [61]. For this purpose, we use only first three pitch candidates from the prioritized set for each frame [62].

3.6 Simulation Results

In this Section, we perform a number of simulations for the estimation of pitch in the presence of noise. We consider naturally spoken speech signals degraded by additive noise for the purpose of simulations. Next, some simulation results on pitch estimation

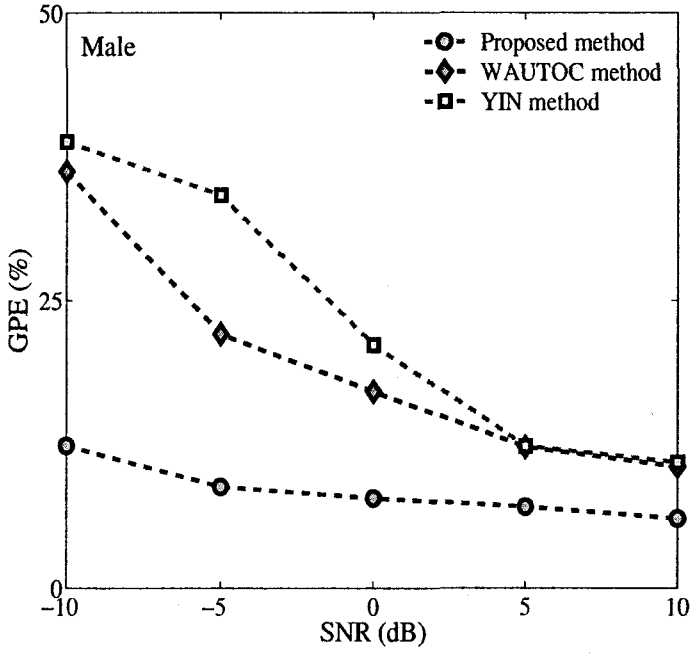


Figure 3.5: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in white noise.

are presented and the estimation performance of the proposed HSAC-SIM method is compared with that of some of the existing pitch estimation methods.

3.6.1 Simulation Conditions

(a) Database and other details: In our simulation study, we employ the *Keele* database [63], [64] of real speech signals, which provides a reference pitch referred to as the "ground" truth. The *Keele* database is of studio quality, sampled at 20 kHz with 16-bit resolution, contains a data sequence "The North Wind Story" uttered by 5 male and 5 female mature English speakers with a total duration of 9 minutes. Here, the speech is segmented into analysis frames, each of size $N = 25.6$ ms at a frame rate of 100 Hz (i.e., a frame shift of 10 ms). Each voiced frame is assigned a

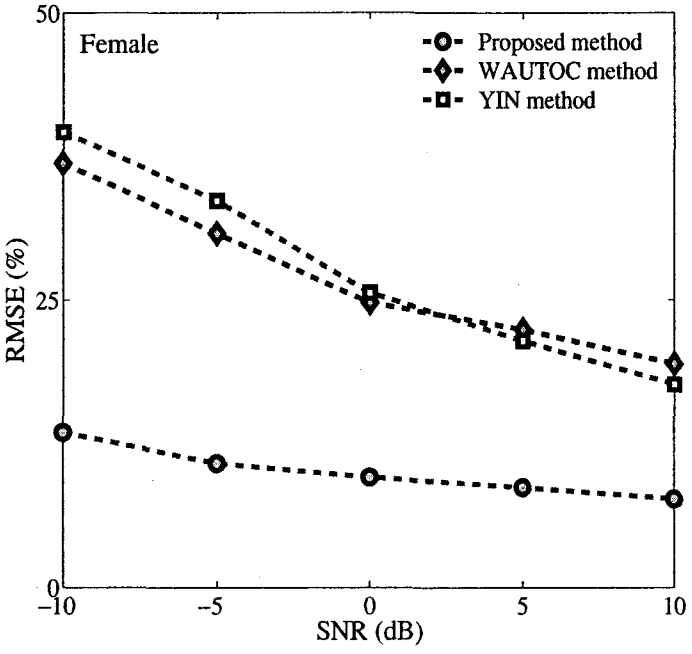


Figure 3.6: RMSE (%) as a function of SNR for female speaker group in white noise.

specific pitch value, an unvoiced frame is provided with a zero pitch value, and an uncertain frame is filled with a negative value of -1.

In order to imitate a condition for a noisy environment, the noisy speech $s_y(n)$ is generated according to (2.26) by adding white or multi-talker babble noise to the original clean speech $s_x(n)$, where the *Noisex92* database [65] is used to obtain the noises. Given the noisy speech $s_y(n)$ and applying a sliding N-sample normalized Hamming window $w(n)$, we obtain a windowed noisy frame, $y'(n) = y_t(n)w(n)$. In our simulations, we used the same values for the parameters, such as frame rate and basic frame (or window) size as specified in the *Keele* database.

To reduce the negative effects of both the additive noise and the formants of the vocal tract, a windowed noisy frame $y'(n)$, in this chapter, we propose to use the DCT

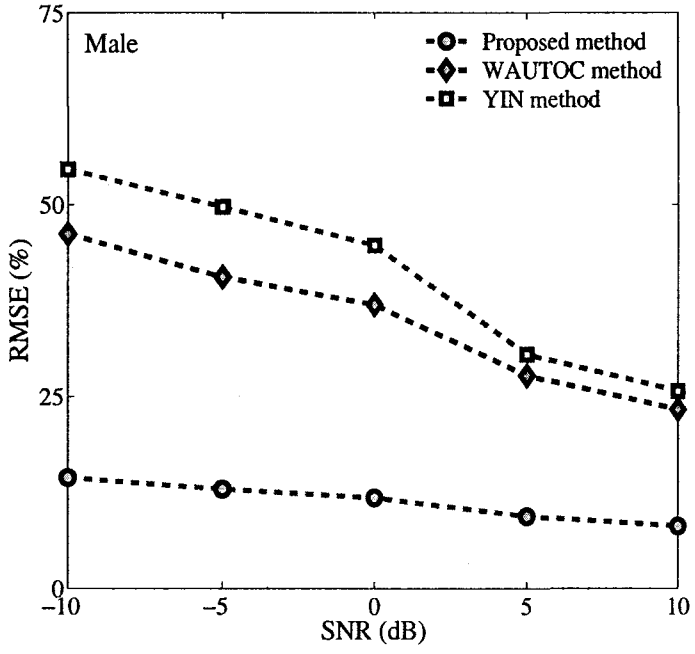


Figure 3.7: RMSE (%) as a function of SNR for male speaker group in white noise.

for the preprocessing of $y'(n)$. The DCT co-efficient of the windowed noisy speech is given by

$$Y'(k) = X'(k) + V'(k) \tag{3.40}$$

where $X'(k)$ and $V'(k)$ represent the DCT coefficients of the windowed clean speech and the windowed noise, respectively. The deleterious impact of noise on speech is relatively small in the first formant range for almost all male and female speakers. Moreover, due to the strong energy compaction property of DCT, most of the speech information tends to be concentrated in a few low-frequency DCT coefficients. Accordingly, the DCT coefficients corresponding to the frequency as high as that of the upper limit of the first formant band should be retained, whereas the rest of the coefficients can be thresholded to zero to effectively eliminate the effects of the

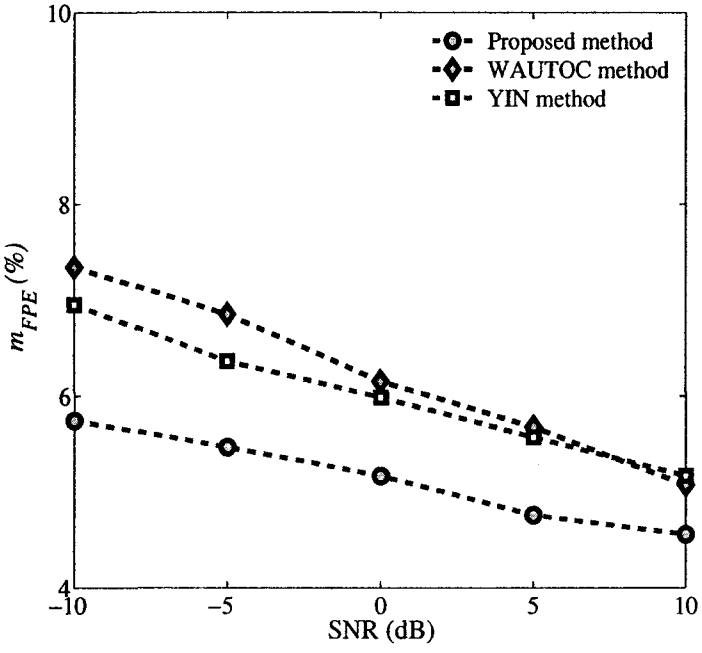


Figure 3.8: m_{FPE} (%) as a function of SNR for all female and male speakers in white noise.

higher formants, resulting in a truncated number of DCT coefficients, $Y(k)$. The time-domain frame of preprocessed noisy speech $y(n)$ as given by (3.1) can be obtained through an inverse DCT operation. The pre-processing of each windowed noisy frame $y'(n)$ is performed using N-point DCT and IDCT operations [73]. It should be pointed out that the DCT pre-processing preserves sufficient number of strong harmonics in $y(n)$, which improves the accuracy of the pitch estimation.

In a frame $y(n)$, in order to determine the optimum value ω_{qopt} of a PH, the search resolution and range for ω_q are kept the same as that used in Chapter 2 for extracting a PH. Unlike in Chapter 2, an initial estimate $\omega_q{}^{(0)}$ of ω_q is obtained in this Chapter from the frequency location corresponding to the maximum peak of the smoothed DCT power spectrum $|\tilde{Y}(k)|^2$ of $y(n)$, as given by (3.23). The number of

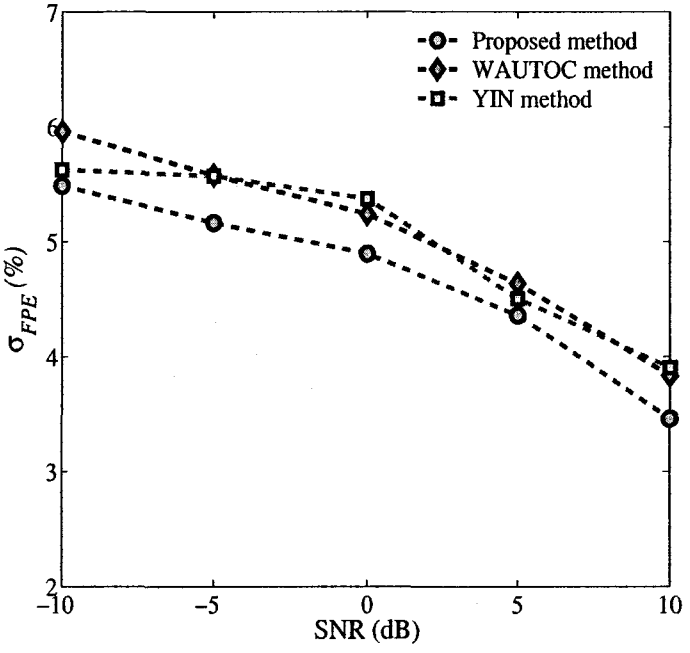


Figure 3.9: σ_{FPE} (%) as a function of SNR for all female and male speakers in white noise .

ACF lags (M) for the LS problem is selected such that $m_{max} < M < N$ in order to accommodate speakers having the shortest ($m_{min} = 60$ samples or 3 ms) to the longest ($m_{max} = 400$ samples or 20 ms) possible pitch period. In order to find the optimum harmonic number q_{opt} corresponding to ω_{qopt} by employing the SIM scheme, the number of unit impulses ($I_T \geq 2$) of the impulse train must be kept constant for a particular speaker depending on the value of m_{max}.

(b) Metrics and Comparison Methods Used for Performance Evaluation: The metrics used for the performance evaluation are 1) the percentage gross pitch error (PGPE); 2) the mean of fine pitch error (m_{FPE}); 3) the standard deviation of fine pitch error (σ_{FPE}); and the 4) the root-mean-square-error (RMSE) as defined before in Chapter 2. The performance of the proposed HSAC-SIM method is evalu-

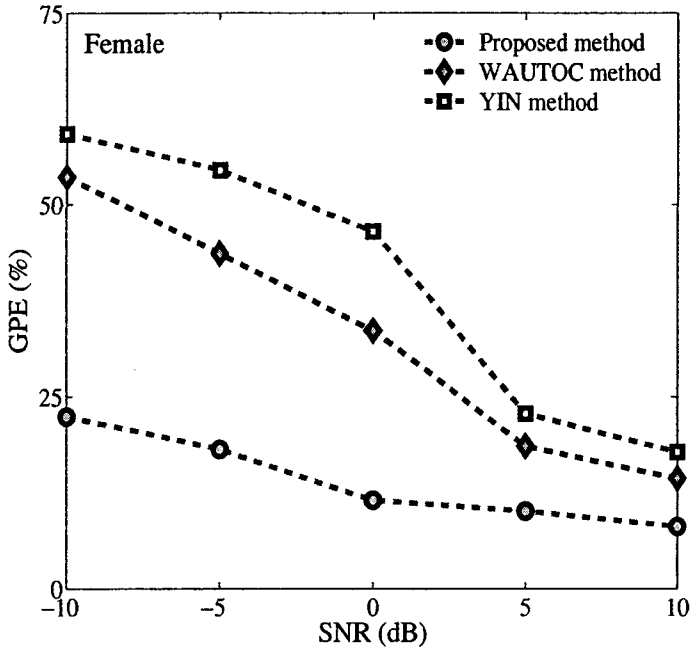


Figure 3.10: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in babble noise.

ated at different levels of SNR in terms of the pitch estimates of the voiced frames based on the voiced/unvoiced labels provided by the Keele database and compared with that of the WAUTOC [46] and YIN [31] methods considered for the purpose of comparison in Chapter 2. For the independent implementation of the WAUTOC method, we have used the parameters specified by the authors while the YIN software is downloaded from the author's homepage and is run using the default parameters provided.

3.6.2 Results and Comparisons

(a) Results on white-noise corrupted speech: The PGPE values as a function of SNR obtained from the WAUTOC, YIN and the proposed HSAC-SIM methods

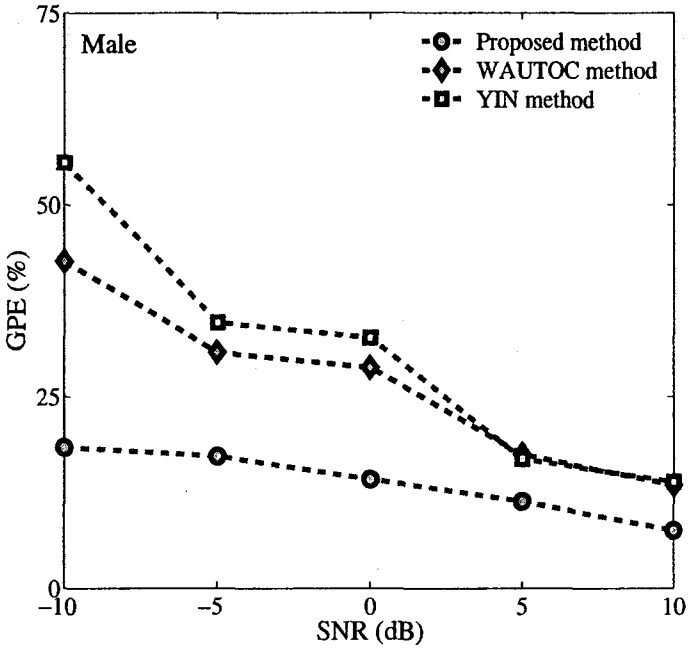


Figure 3.11: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in babble noise.

are shown in Figs. 3.4 and 3.5 for white-noise corrupted speech signals of the female (5) and male (5) speaker groups, respectively, where the SNR varies from a very low value of -10 dB to a high value of 10 dB. It is seen from these figures that at SNR $= 0$ dB or higher, although the other methods provide acceptable performances, the estimation accuracy achieved by the proposed method is much higher. It is also seen from these figures that even at SNR $= -10$ dB, when the other two methods produce unsatisfactory results, the proposed method successfully estimates the pitch with sufficient accuracy.

Figs.3.6 and 3.7 present the variation of RMSE values with respect to the level of SNR for all three methods using the same female and male speaker groups as above. It is observed from these figures that the YIN and WAUTOC methods give much

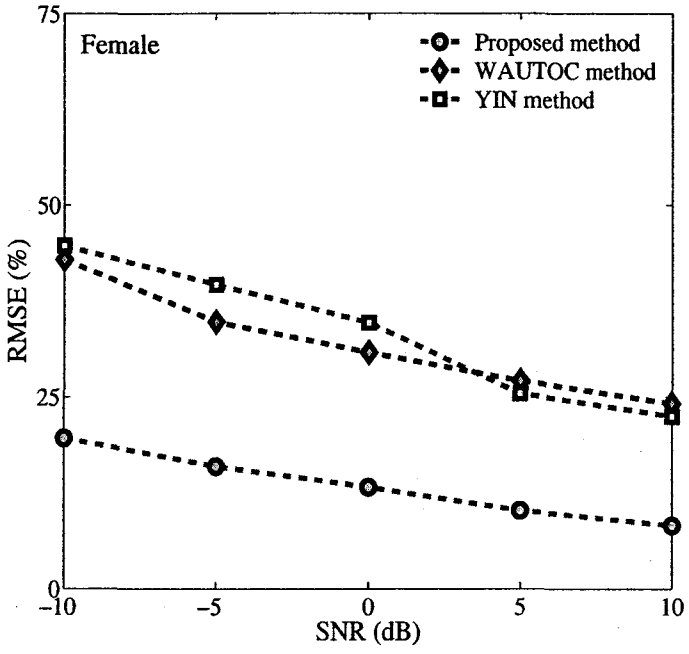


Figure 3.12: RMSE (%) as a function of SNR for female speaker group in babble noise.

higher RMSE values at the levels of SNR, such as 0 dB or above, and they show a poor performance when the SNR is low. However, the proposed method performs significantly better even at a low level of SNR (as low as -10 dB).

In Figs. 3.8 and 3.9, the mean m_{FPE} and standard deviation σ_{FPE} resulting from the three methods are plotted for the set of 10 mixed (5 female plus 5 male) speakers of the database. Clearly, the overall mean and standard deviation values of the fine-pitch errors obtained by using the proposed HSAC-SIM method are lower in the entire range of the SNR levels considered, indicating high estimation accuracy in comparison to that achieved by the other methods.

Significantly smaller PGPE and RMSE achieved by using the proposed method under a wide range of SNR levels, along with lower overall mean and standard de-

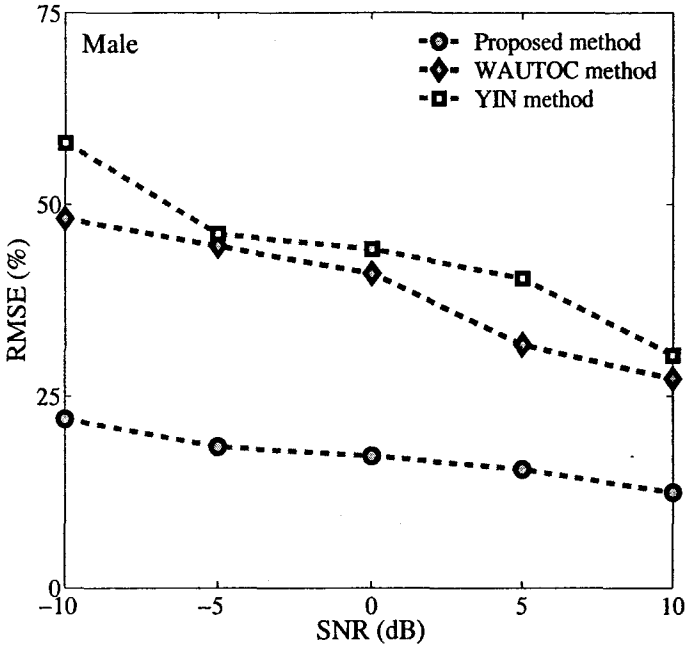


Figure 3.13: RMSE (%) as a function of SNR for male speaker group in babble noise.

viation of the fine-pitch errors, indicate its high degree of estimation accuracy and robustness.

(b) Results on Multi-Talker Babble-Noise Corrupted Speech: We now examine the robustness of the proposed HSAC-SIM and the other two methods in the presence of a real-life babble noise. The pitch estimation results in terms of the PGPE, RMSE, mean m_{FPE}, and standard deviation σ_{FPE} for each of the methods are depicted in Fig. 3.10 through Fig. 3.15. As expected, the performance of all the three methods degrades in the presence of babble noise compared to that in the white noise. However, an important observation that can be made from these figures is that the proposed HSAC-SIM method remains superior with respect to all the four performance metrics at all the levels of SNRs for the same male and female speaker

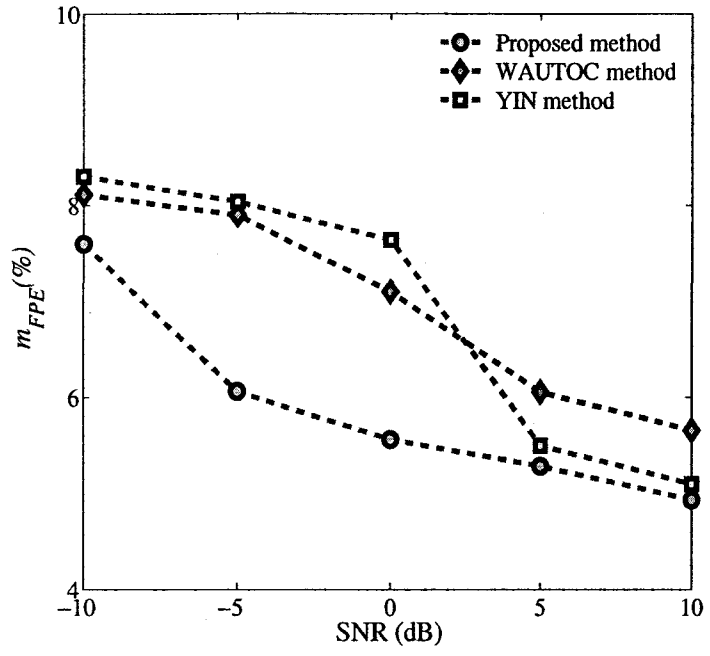


Figure 3.14: m_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

groups as the ones considered for the white-noise corrupted speech.

Figs. 3.10 and 3.11 portray the PGPE values as a function of SNR obtained from the three methods for the female and male speaker groups, respectively. It is seen that, in contrast to the white noise case, it is only the proposed method that continues to provide satisfactory satisfactory performance at SNR levels, such as 0 dB or above. Also, the proposed method still remains capable of estimating pitch with acceptable accuracy at a very low level of SNR of -5 dB or even lower than that.

Figs. 3.12 and 3.13 provide plots for the RMSE resulting from using the three pitch estimation methods for the female and male speaker groups, respectively. As seen, the proposed HSAC-SIM method remains significantly better even for the low levels of SNR, such as -10 dB. It may be pointed out that the RMSE values of the

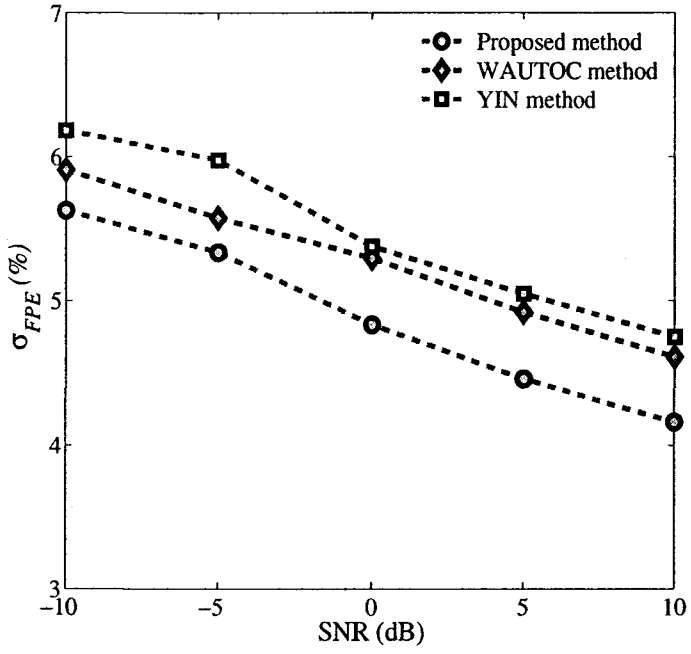


Figure 3.15: σ_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

other two methods degrade much at an SNR level below 10 dB.

The mean m_{FPE} and standard deviation σ_{FPE} obtained from the three methods as a function of SNR are plotted in Figs. 3.14 and 3.15, respectively, for the same set of 10 mixed (5 female plus 5 male) speakers as the one used in Figs. 3.8 and 3.9. These figures illustrate that the estimation accuracy of the proposed HSAC-SIM method is reduced expectedly in comparison to the white noise case, but it still provides quite better results compared to that provided by the other two methods for a wide range of SNR.

In order to evaluate the net effect of pitch tracking by using DP on the pitch estimation performance of the proposed HSAC-SIM method, we finally plot a reference pitch contour for a 1.4-second excerpt of the clean speech of a male speaker from

the reference database and the pitch contours from other pitch estimation methods in Fig.3.16. The reference pitch contour is accompanied by the spectrogram of clean speech and the other pitch contours are overlaid on the spectrograms of the white noise-corrupted speech. It is visible from the figure that in contrast to the other methods, the proposed method yields a comparatively smoother pitch contour even at an SNR of -10 dB. Similarly, Fig. 3.17 shows a comparison of the pitch contours resulting from the three methods for female speech corrupted by babble noise at SNR $= -10$ dB. From Fig. 3.17, it is clear that the proposed method is able to give a smoother contour even in the presence of babble noise. The pitch contours in Figs.3.16 and 3.17 obtained from the three methods have convincingly illustrated that the HSAC-SIM method is capable of significantly reducing the double and half-pitch errors by the use of the proposed pitch tracking scheme. Thus, we infer that the proposed method is suitable for real-life applications involving noise-corrupted speech with a very low SNR.

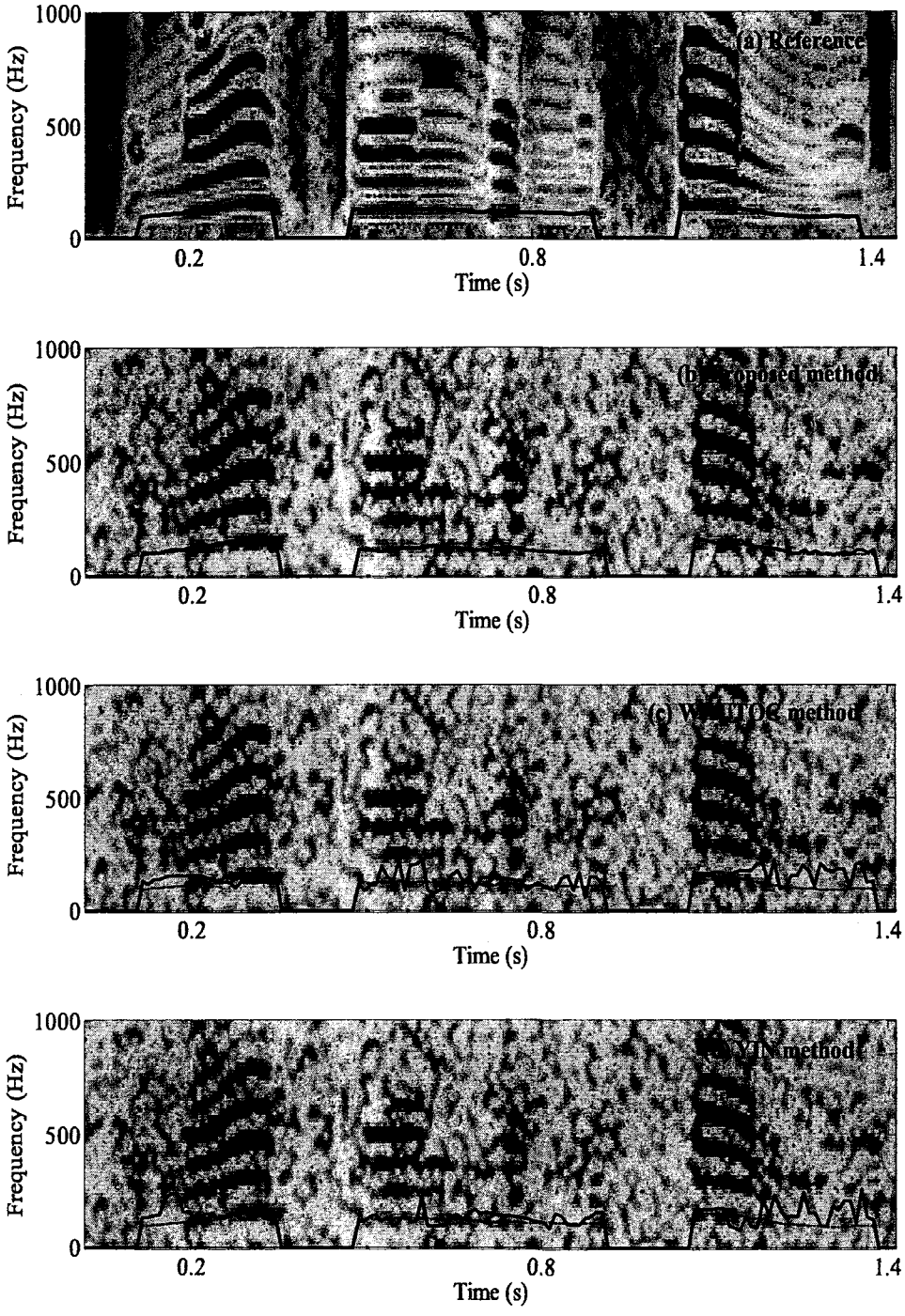


Figure 3.16: Pitch contours of different methods at SNR = -10 dB in white noise.

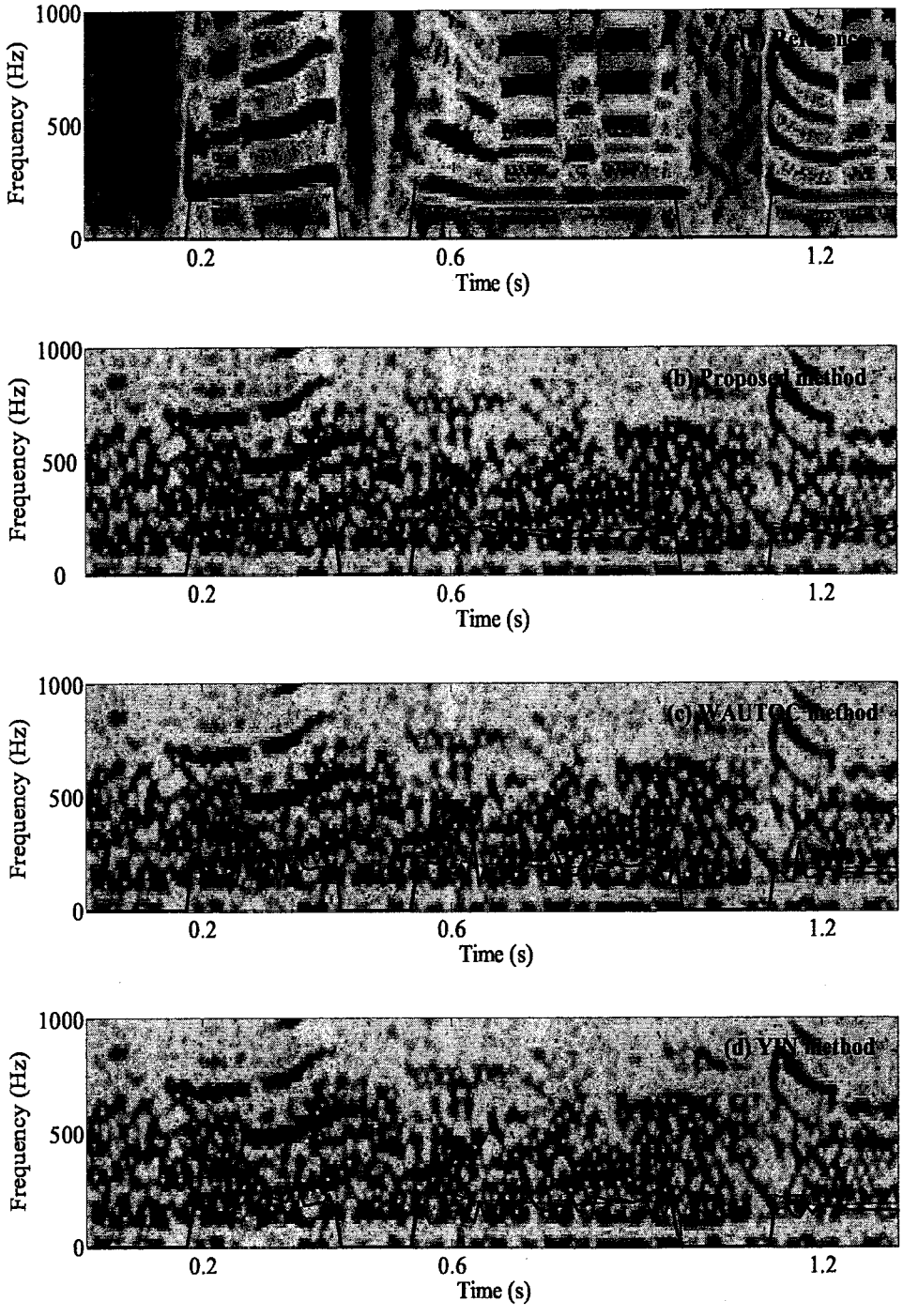


Figure 3.17: Pitch contours of different methods at SNR = -10 dB in babble noise.

3.7 Conclusion

In this chapter, a new method for the estimation of pitch from noise-corrupted speech observations has been presented. The proposed method is based on developing a compact but accurate autocorrelation model of clean speech, called the HSAC model. Similar to the HCAC model proposed in Chapter 2, it is also expressed in terms of pitch-harmonics of clean speech but, unlike the HCAC model, the HSAC model is derived from the conventional autocorrelation estimator and the cross-product terms of different harmonics are taken into consideration in its derivation.

A least-squares autocorrelation-fitting optimization technique employing the HSAC model has been presented for a more accurate estimation of PH from the noisy speech. The proposed optimization technique can incorporate some *a priori* knowledge of the PH, if available, to facilitate the process of pitch estimation. Here, for the purpose of obtaining such a priori knowledge of the PH, the DCT is employed in contrast to the HCAC model based optimization in which the FFT is used. Since the determination of the harmonic number associated with the PH becomes difficult in a noisy condition, an impulse-train with its period governed by a PH estimate and a new SAMSF, which is able to retain its periodicity peaks at integer multiples of the pitch period as well as prevent the decay of its peaks at a much slower rate in contrast to that of the conventional ACF even in the presence of a severe noise, have been introduced to formulate an objective function. Then, in order to achieve the desired harmonic number for accurate pitch estimation, the objective function has been maximized by matching the periodicity of the impulse-train and that of the SAMSF through a proposed SIM technique. Finally, a pitch tracking scheme using DP has been performed to obtain a smoothed pitch contour.

From an intensive simulation study on naturally spoken speech signals, it has been shown that the proposed time-frequency domain method, referred to as HSAC-SIM, is able to estimate pitch with sufficient accuracy and consistency for a wide range of speakers in the presence of white or multi-talker babble noise with very low levels of SNR as compared to some of the state-of-the-art techniques in the literature. The better efficacy of the HSAC-SIM method in a multi-talker babble noise scenario supports its applicability in real-life applications.

Chapter 4

Pitch Estimation Based on a Magnitude Difference Function of an Excitation-Like Signal Obtained From Noisy Speech

4.1 Introduction

In the widely used working model for speech production, the vocal-tract system (VTS) is driven by an excitation signal. In general, the VTS can be treated as an acoustic resonator whose resonance frequencies are determined by the system parameters representing the formants of the speech signal. For voiced speech, the excitation signal can be modeled as a train of impulses that are separated by the pitch period. In this chapter, we consider the voiced speech as an output of the VTS driven by an impulse-train excitation. The significant excitation of the VTS within one pitch period takes place at the instant of glottal closure (GC) [74]. One may determine the characteristic of voice excitation, i.e., the pitch period by careful analysis of the speech signal with the help of GC instants [75]. Based on this concept, the ideal solution to the problem of pitch period estimation is to extract or generate the excitation signal consisting of pulses. However, generation of the excitation signal from the observed

output speech data is a difficult task, especially when the speech observations are heavily noise-corrupted.

An error or residual signal (RS) of speech has been commonly employed for the pitch period estimation. Since the peak of the RS is supposed to indicate a GC instant [76], [77], some characteristics of the excitation signal can be better observed in the RS than the speech signal itself. Hence, some pitch estimation methods [23], [47] have been developed in which the RS or the processed RS is used for computing the conventional autocorrelation function (ACF). Such methods involve the inverse filtering of the speech signal by using the VTS parameters. If the VTS parameters are accurately identified, the RS should resemble the excitation signal containing pitch information. One approach to derive the RS from the given speech is to exploit the VTS parameters obtained from the ACF based linear prediction (LP) analysis [78], [79]. These methods of VTS parameter identification however ignore the effect of increased overlapping in the ACF of speech due to aliasing between the ACF of the vocal-tract impulse response and that of excitation pulses, especially for high-pitch female speakers [41], [80]. Moreover, due to the phase angles of the formants at the GC instants, peaks of either polarity occur in the LP residual around the GC instants. Hence the direct use of the LP residual to detect the GC instants is difficult. Note that the LP based methods of RS extraction are capable of handling only the noise-free speech. In noisy environments, due to an inaccurate estimate of the VTS parameters, the RS fail to deliver the distinguishable peaks at the GC instants [81], [82] [83], [84]. As a result, even the ACF computed from the RS or the processed RS becomes ineffective in representing the GC instants and the performance of pitch estimation deteriorates significantly under low levels of SNR. As such, our target is to generate an excitation-like signal (ELS) for pitch estimation by identifying the VTS parameters

while overcoming the overlapping problem without reducing the accuracy much in the presence of noise. At the same time, we expect the bipolar fluctuations of the RS at the GC instants to be removed from the ELS. Instead of using a time-domain function of the noisy speech [85], [86], [87], a time-domain function of the ELS thus generated from the noisy speech would be more desirable for reliable pitch estimation under a very low SNR.

In this chapter, a new approach using an ELS obtained from the noisy speech is proposed for the estimation of pitch from speech signals under low levels of SNR [88]. The first task is to extract the VTS parameters from the noisy speech based on the principle of time-frequency domain homomorphic deconvolution. By utilizing the extracted VTS parameters, an inverse filtering of the noisy speech is then performed to yield a RS. Next, we propose to employ a noise-compensated autocorrelation function (ACF) of the RS to generate a squared Hilbert envelope (SHE) representing an ELS of the voiced speech. In order to overcome the undesirable effect of noise on the ELS at a very low SNR, a new symmetric normalized magnitude difference function (SNMDF) of the ELS is proposed, which is eventually used to estimate the pitch on a frame-by-frame basis. Finally, a dynamic programming based pitch tracking scheme using the SNMDF values from the neighboring frames is presented to obtain a smooth pitch contour. In order to demonstrate the effectiveness of the proposed method, extensive simulations are performed by considering naturally spoken speech signals obtained from the *Keele* database in the presence of additive white or multi-talker babble noise adopted from the *Noisex92* database and the results are compared with those obtained from some of the existing methods. It is shown that the proposed time-frequency domain method, referred to as SNMDF-ELS, is capable of estimating the pitch accurately and consistently for both female and male speaker groups at

an SNR level as low as -10 dB, a level at which most of the existing methods fail to provide accurate estimation. The suitability of the proposed method in practical applications is illustrated through quite a satisfactory pitch estimation when the speech observations are heavily corrupted by a real-world multi-talker babble noise also.

The rest of the chapter is organized as follows. In Section 4.2, we present a brief overview of the proposed pitch estimation method through a block diagram. In Section 4.3, we first describe the principle of Homomorphic Deconvolution (HD) in the ACF domain for the identification of VTS parameters. Then a scheme for the identification of the VTS parameters using HD in the presence of noise is presented. Section 4.4 presents the proposed methodology to generate an ELS under noisy conditions. In this section, first, by utilizing the inverse VTS, an RS is produced and a scheme of noise compensation on the RS in the ACF domain is introduced. Based on the Hilbert Transform of the noise-compensated ACF of the RS, the generation of a SHE as the ELS is then described. In Section 4.5, we introduce an SNMDF of the ELS and demonstrate its effectiveness for pitch estimation and pitch tracking under very low SNRs. The performance of the proposed method is illustrated in Section 4.6 through extensive computer simulations using both white and multi-talker babble noise-corrupted speech signals. Finally, in Section 4.7, the prominent features of this investigation are summarized with some concluding remarks.

4.2 A Brief Description of the Proposed Method

An overview of the proposed pitch estimation method is presented through a block diagram in Fig. 4.1. A pre-processed frame of the observed noisy speech $s_y(n)$ can

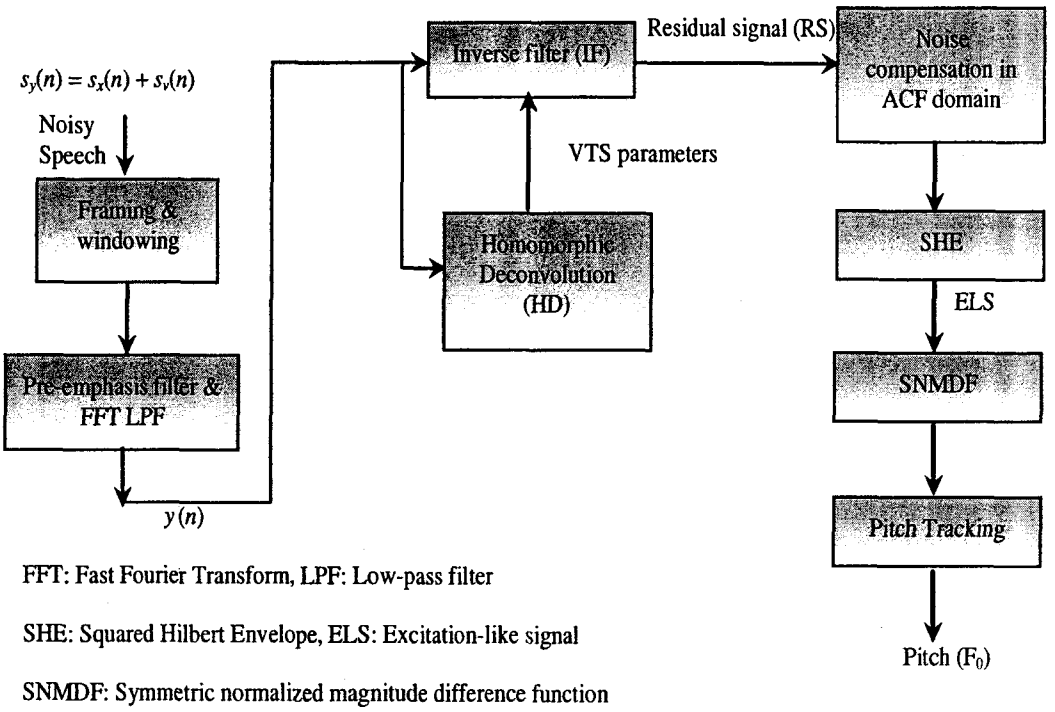


Figure 4.1: A block diagram representing the overview of the proposed pitch estimation method.

be expressed in the time-domain as

$$y(n) = x(n) + v(n) \tag{4.1}$$

where $x(n)$ and $v(n)$ represent the pre-processed versions of clean speech $s_x(n)$ and additive noise $s_v(n)$, respectively. Note that in Chapters 2 and 3, each windowed noisy speech frame of $s_y(n)$ was low-pass filtered to retain only the first formant (e.g., the $0 - 900$ Hz range) in the pre-processed noisy speech frame $y(n)$ and then one pitch-harmonic (PH) and the corresponding harmonic number were determined to estimate the pitch. In this chapter, the pre-processing involves the windowing operation first and then the windowed noisy speech frame is pre-emphasized. In addition to signal pre-emphasis, a low-pass filtering is also performed to preserve a bandwidth

up to 5 kHz so as to exclude only the superfluous high-frequency content that is not of our interest. Such a pre-processing assumes we retain $4-5$ formants, which facilitates the identification of the VTS parameters required for the ELS generation. Unlike the previous two chapters, instead of using $y(n)$ directly for computing a time-domain function for pitch estimation, we first attempt to pre-whiten $y(n)$ by filtering it through an inverse of the VTS to generate an ELS of the voiced speech and then compute a time-domain function of the obtained ELS for pitch estimation.

4.3 Vocal-Tract System Identification by Homomorphic Deconvolution

In order to identify the VTS parameters required for producing the RS of speech, we propose to employ a technique via the principle of homomorphic deconvolution (HD) in the ACF domain. The proposed HD based technique overcomes the aliasing problem of the conventional LP based methods of VTS parameter identification.

4.3.1 Homomorphic Deconvolution (HD) in the Correlation Domain

For a frame of speech signal, the human vocal tract system (VTS) is characterized by a p-th order autoregressive (AR) model. The impulse response of the VTS is given by

$$g(n) = -\sum_{k=1}^{p} a_k g(n-k) + \delta(n), \tag{4.2}$$

where a_k represents the AR parameters of the VTS and $\delta(n)$ is an impulse. By multiplying (4.2) with $g(n-m)$ and summing over n, the ACF $\phi_g(m)$ of $g(n)$ can be written as

$$\phi_g(m) = -\sum_{k=1}^{p} a_k \phi_g(m-k), \qquad 0 < m \leq p \tag{4.3}$$

where m is the discrete lag variable and $\phi_g(m)$ obeys a recursive relation that relates the autocorrelation values to the a_k parameters of the VTS. For a causal sequence $g(n)$, $\phi_g(-m) = \phi_g(m)$ and for $m = 1, \ldots, p$, (4.3) comprises p linear equations that can be written in a matrix notation as

$$\mathbf{\Phi}_g \mathbf{a} = -\phi_g \tag{4.4}$$

where $\mathbf{\Phi}_g$ is the autocorrelation matrix with dimension $p \times p$ containing elements $\phi_g(m,k) = \phi_g(m-k)$, $\mathbf{a}$ is a column vector of length p containing a_k parameters of the VTS, and $\phi_\mathbf{g}$ is also a column vector of length p containing autocorrelations $\phi_g(m)$. If we are given a set of autocorrelation coefficients $\phi_g(1), \phi_g(2), \ldots \phi_g(p)$, p number of a_k parameters can be obtained by solving (4.4), which is referred to as the normal equations [2]. However in reality for natural speech signals, $g(n)$ as well as its ACF $\phi_g(n)$ are not available, only the VTS output, i.e., speech sequence $s_x(n)$, is available in a noise-free environment.

A frame of $s_x(n)$ is allowable to extract the VTS parameters presumed to remain fixed within it. The voiced speech $x(n)$ in a frame can be modeled in time domain as the following convolution operation,

$$x(n) = e(n) * g(n) = \sum_{t_I=0}^{(I_T-1)} g(n - t_I T_0), \tag{4.5}$$

where $e(n)$ indicates a frame of an impulse-train excitation driving the VTS with impulse response $g(n)$. If I_T number of impulses (i.e., $(I_T - 1)$ periods) are accommodated inside the frame $e(n)$ of the excitation, then

$$e(n) = \sum_{t_I=0}^{(I_T-1)} \delta(n - t_I T_0). \tag{4.6}$$

In (4.5) and (4.6), T_0 is the pitch period representing the separation between excitation pulses. Thus the resulting voiced speech $x_(n)$ consists of recurring replicas of $g_(n)$ as

seen in (4.5). The ACF $\phi_x(m)$ of $x(n)$ is conventionally estimated as

$$\phi_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x(n)x(n+|m|), m = 0, \pm 1, \pm 2, \ldots, \pm M, M < N. \tag{4.7}$$

In order to realize the effectiveness of the ACF $\phi_x(m)$ in identifying the VTS parameters a_k , it can be expressed as

$$\phi_x(m) = \phi_e(m) * \phi_g(m) = \sum_{t_I=0}^{(I_T-1)} \phi_g(n - t_I T_0), \quad \forall m, \tag{4.8}$$

where $\phi_e(m)$ is the ACF of $e(n)$, which is periodic with pitch period T_0. It is seen from (4.8) that $\phi_x(m)$ represents an aliased version of $\phi_g(m)$, as a copy of $\phi_g(m)$ is repeated periodically with a periodicity equivalent to T_0. Depending on T_0, $\phi_g(m)$ overlaps and alters the underlying $\phi_x(m)$. It is observed that for low-pitch male speakers with a larger T_0, $g(n)$ damps out within T_0 causing an insignificant overlapping and in any period, $\phi_x(m) \approx \phi_g(m - t_I T_0)$. Thus, by using the lags of $\phi_x(m)$ lying within a period, the VTS parameters can be estimated from (4.4). On the other hand, for high-pitch female speakers with a shorter T_0, $g(n)$ cannot decay completely within T_0, causing a significant overlapping. Since, within a period, the lags of $\phi_x(m)$ differ considerably from that of $\phi_g(m)$, an accurate estimate of the VTS parameters cannot be guaranteed in such a case [41]. Irrespective of the speaker's T_0, which is unknown beforehand, in order to extract the true solutions of the VTS parameters a_k using (4.4), an estimate of $\phi_g(m)$ from $\phi_x(m)$ has to be obtained.

One possible way to handle this problem is to separate the effect of $\phi_e(m)$ from $\phi_x(m)$ and thus obtain an estimate of $\phi_g(m)$ that is nearly free from aliasing. Cepstrum analysis has widely been used for homomorphic deconvolution or separation of signals that have been combined through a nonlinear operation, like convolution [89]. As seen from (4.8), since $\phi_x(m)$ is produced because of the convolution of $\phi_g(m)$ and

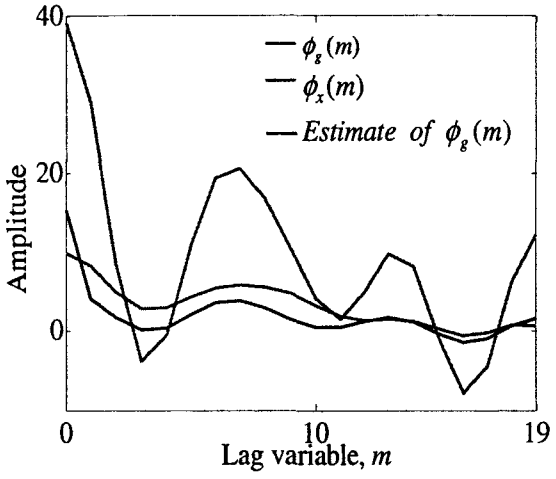


Figure 4.2: An estimate of the ACF $\phi_g(m)$ of the vocal tract impulse obtained from clean speech.

$\phi_e(m)$, we employ homomorphic deconvolution of $\phi_x(m)$ through cepstrum analysis in order to obtain an estimate of $\phi_g(m)$ from $\phi_x(m)$. For an N-point real sequence $f(n)$, in general, the cepstrum of $f(n)$ can be defined as [90]

$$c_f(n) = T^{-1}[\log|T[f(n)]|] \tag{4.9}$$

where $T[\cdot]$ and $T^{-1}[\cdot]$, respectively, represent a transform and its inverse operator. In most applications, T refers to the discrete Fourier transform and as usual a natural logarithm is used in (4.9).

By substituting the expression of $\phi_x(m)$ as given by (4.8) in (4.9), the real cepstrum $c_{\phi_x}(n)$ of $\phi_x(m)$ can be expressed as

$$c_{\phi_x}(n) = c_{\phi_g}(n) + c_{\phi_e}(n). \tag{4.10}$$

Generally, $c_{\phi_g}(n)$, the cepstrum of the VTS impulse response, is the quickly decaying portion of $c_{\phi_x}(n)$ and it is represented within the low quefrency zone. The cepstrum $c_{\phi_e}(n)$ of the frame of periodic impulse-train excitation exhibits significant values at

multiples of the pitch period T_0 in $c_{\phi_x}(n)$. Since $c_{\phi_g}(n)$ and $c_{\phi_e}(n)$ are well separated in the quefrency domain, the effect of $c_{\phi_e}(n)$ on $c_{\phi_x}(n)$ can be removed by using a low-time lifter [41]. The output of such a liftering operation in the quefrency domain is referred to as liftered $c_{\phi_x}(n)$, which is close to $c_{\phi_g}(n)$. By using the inverse cepstrum operation on the liftered $c_{\phi_x}(n)$, an estimate of $\phi_g(m)$ can be obtained and the use of such an estimate in the normal equations given by (4.4) can provide better identification results of the VTS parameters a_k. In Fig. 4.2, $\phi_x(m)$, $\phi_g(m)$ and an estimate of $\phi_g(m)$ for a voiced speech frame are plotted. This figure shows that, since the excitation component $c_{\phi_e}(n)$ has been eliminated via the HD technique, the resulting estimate of $\phi_g(m)$ is neatly separated from $\phi_x(m)$ and it appears to be a good approximation of $\phi_g(m)$.

4.3.2 Identification of VTS using HD in the Presence of Noise

In the presence of noise, the ACF of a noisy speech frame $y(n)$ can be computed as

$$\phi_y(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} y(n)y(n+|m|), m = 0, \pm 1, \pm 2, \ldots, \pm M, M < N. \tag{4.11}$$

Since $y(n) = x(n) + v(n)$, (4.11) can be re-written as

$$\phi_y(m) = \phi_x(m) + \phi_n(m) \tag{4.12}$$

where $\phi_n(m) = \phi_v(m) + \phi_c(m)$, $\phi_v(m)$ is the ACF of $v(n)$ and $\phi_c(m)$ represents the sum of cross-correlation terms. It is worth mentioning that under a heavy noisy condition, $\phi_n(m)$ in (4.12) cannot be neglected. In order to compute the cepstrum of $\phi_y(m)$, we apply a logarithm operation on the discrete Fourier transform of $\phi_y(m)$ as given by (4.12) and obtain

$$\log(\Phi_y(k)) = \log\left[\Phi_x(k)\left(1 + \frac{\Phi_n(k)}{\Phi_x(k)}\right)\right] = \log[\Phi_x(k)\Phi_w(k)] = \log[\Phi_x(k)] + \log[\Phi_w(k)]; \tag{4.13}$$

here $\Phi_y(k)$, $\Phi_x(k)$ and $\Phi_n(k)$ are the DFTs of $\phi_y(m)$, $\phi_x(m)$ and $\phi_n(m)$, respectively. By taking the inverse DFT of $\log(\Phi_y(k))$, the real cepstrum $c_{\phi_y}(m)$ of $\phi_y(m)$ can be expressed as

$$c_{\phi_y}(m) = c_{\phi_x}(m) + c_{\phi_w}(m) = c_{\phi_g}(m) + c_{\phi_e}(m) + c_{\phi_w}(m). \tag{4.14}$$

The term $c_{\phi_w}(m)$, arising because of the noise, determines how the noise affects $c_{\phi_y}(m)$. Based on the analysis of many stationary and non-stationary noise signals, it has been observed that, autocorrelation sequence of a noise signal exhibits the largest absolute value at the zero lag location [91]. One way of reducing the effect of $c_{\phi_w}(m)$ on $c_{\phi_y}(m)$ is to lessen the effect of noise from the noisy ACF $\phi_y(m)$. The noise effect is reduced by excluding the zero lag (or replacing by a smaller value) while computing $c_{\phi_y}(m)$ from $\phi_y(m)$ [92]. Next, a liftering operation is performed in a similar fashion as described in the noise-free case, which not only removes the effect of $c_{\phi_e}(m)$ on $c_{\phi_y}(m)$, but also further reduces the effect of $c_{\phi_w}(m)$ from the high quefrency portion of $c_{\phi_y}(m)$. Thus, the noise-reduced liftered cepstrum denoted by

$$\widetilde{c}_{\phi_y}(n) \approx c_{\phi_g}(n) + \widetilde{c}_{\phi_w}(n) \tag{4.15}$$

is employed in the inverse cepstrum operation to obtain an estimate $\widetilde{\phi}_g(m)$ of $\phi_g(m)$. Such an estimate $\widetilde{\phi}_g(m)$ is used in (4.4) to identify the VTS parameters. Since in the presence of noise, lower lags of $\widetilde{\phi}_g(m)$ generally become more corrupted than that of the higher lags, a few lower lags of $\widetilde{\phi}_g(m)$ are avoided in the computation of the VTS parameters. Utilizing the autocorrelation coefficients $\widetilde{\phi}_g(p+1), \widetilde{\phi}_g(p+2), \ldots, \widetilde{\phi}_g(p+S)$ in (4.4) yields a set of linear equations, which can be represented in the following

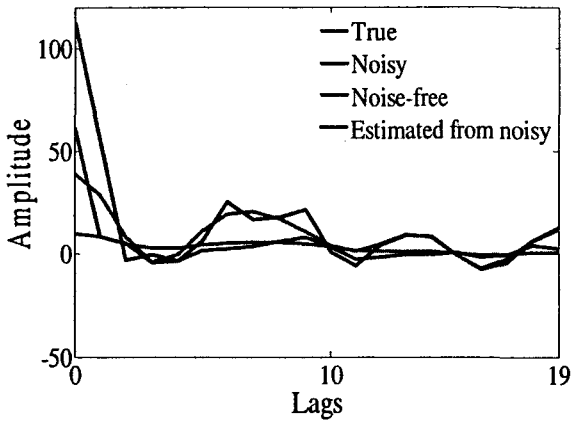


Figure 4.3: An estimate of ACF $\phi_g(m)$ of the vocal tract impulse response in the presence of noise.

matrix form

$$\begin{bmatrix} \widetilde{\phi}_g(p) & \widetilde{\phi}_g(p-1) & \dots & \widetilde{\phi}_g(1) \\ \widetilde{\phi}_g(p+1) & \widetilde{\phi}_g(p) & \dots & \widetilde{\phi}_g(2) \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{\phi}_g(p+S-1) & \dots & \dots & \widetilde{\phi}_g(S) \end{bmatrix} \begin{bmatrix} \widetilde{a}_1 \\ \widetilde{a}_2 \\ \vdots \\ \widetilde{a}_p \end{bmatrix} = - \begin{bmatrix} \widetilde{\phi}_g(p+1) \\ \widetilde{\phi}_g(p+2) \\ \vdots \\ \widetilde{\phi}_g(p+S) \end{bmatrix} \tag{4.16}$$

where S governs the number of equations to be used in (4.16). An estimate $\widetilde{a}_k$ of the VTS parameters can easily be obtained from the least-squares solution of (4.16) [93]. In Fig. 4.3, $\phi_x(m)$, $\phi_g(m)$, $\phi_y(m)$ and $\widetilde{\phi}_g(m)$ for a voiced speech frame are plotted. The figure depicts that by performing cepstrum deconvolution of $\phi_y(m)$; the effect of excitation components as well as some degree of additive noise is reduced, thus yielding $\widetilde{\phi}_g(m)$ close to $\phi_g(m)$. Moreover, the use of higher lags of $\widetilde{\phi}_g(m)$ in (4.16) would provide reasonably better estimates of the VTS parameters $\widetilde{a}_k$, even in a variety of different noisy environments.

4.4 Generation of an Excitation-like Signal via Inverse VTS and Hilbert Transform

For voiced speech, the significant excitation of the VTS occurs at the instant of GC at which closure of vocal folds takes place within a pitch period. In order to estimate the pitch period, we want to use the GC instants that are expected to be better detected from the peaks of an ELS than that of the speech itself. We propose to generate the ELS by exploiting the Hilbert Transform of the ACF of an RS which can be produced as an output of inverse VTS.

4.4.1 Residual Signal as an Output of Inverse VTS

If we perform an inverse-filtering of the speech $x(n)$ in a frame, where $x(n)$ is let to pass through an inverse VTS filter $A(z)$ [1] given by

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \tag{4.17}$$

the output of the inverse VTS filter is referred to as the error or RS:

$$\Re(n) = T^{-1}[A(z)X(z)] = x(n) + \sum_{k=1}^{p} a_k x(n-k), \tag{4.18}$$

with T^{-1} representing the inverse operator of a z transform T, where $T[x(n)] = X(z)$. If the VTS parameters can be accurately identified from $x(n)$, the RS is supposed to closely resemble the excitation signal that produces $x(n)$. In voiced speech, since the strength of the excitation signal containing the pitch information is higher around the GC instants, large errors or peaks are associated with the GC instants in the RS. It is found that the effect of phase angles of the formants corresponding to the VTS parameters a_k is to introduce bipolar fluctuations around the GC instants [74].

In the presence of noise, when the noise-corrupted speech $y(n)$ in a frame is passed through an inverse VTS system $\widetilde{A}(z)$ with the parameters $\widetilde{a}_k$ already identified from

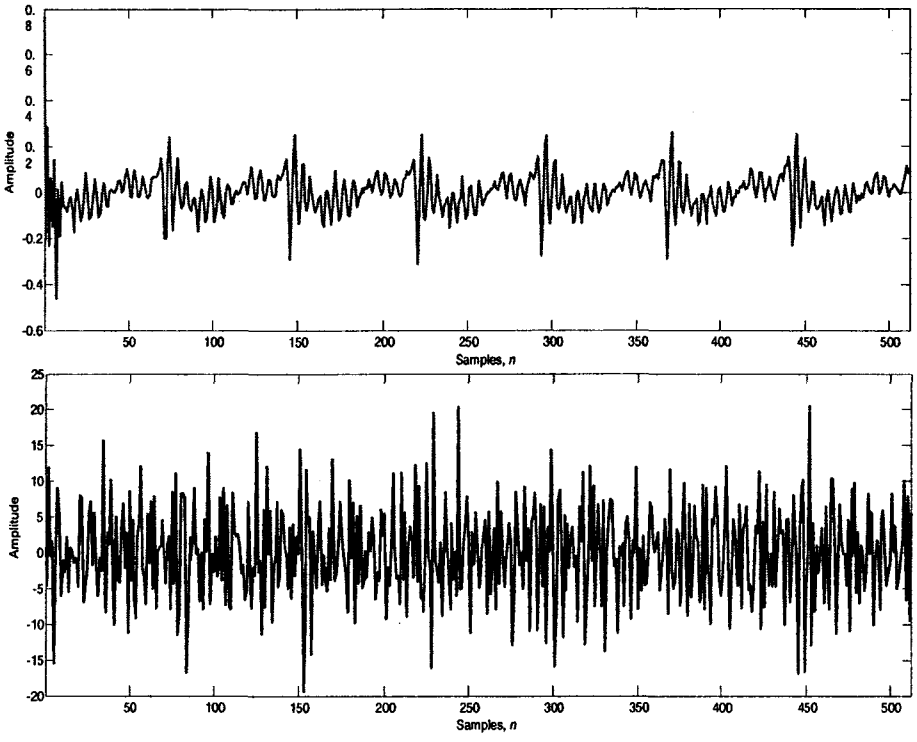


Figure 4.4: (a) Residual signal $\Re(n)$ obtained from clean speech $x(n)$ (b) Residual signal $\widetilde{\Re}(n)$ obtained from noisy speech $y(n)$.

$y(n)$, the resulting RS $\widetilde{\Re}(n)$ is given by

$$\widetilde{\Re}(n) = T^{-1}[\widetilde{A}(z)Y(z)] = y(n) + \sum_{k=1}^{p} \widetilde{a}_k y(n-k), \tag{4.19}$$

where $T[y(n)] = Y(z)$. In Fig. 4.4, the residual signals $\Re(n)$ and $\widetilde{\Re}(n)$ obtained using (4.18) and (4.19), respectively, are plotted. It is seen from this figure that it is difficult to use the residual signals directly for the detection of the GC instants due to the bipolar fluctuations of their amplitudes around the GC instants [75]. Furthermore, at a very low SNR, the RS $\widetilde{\Re}(n)$ in (4.19) could be significantly different from the excitation signal due to the inaccurate estimates of the VTS parameters $\widetilde{a}_k$ [1]. Since, $Y(z) = T[x(n)] + T[v(n)] = X(z) + V(z)$, even though a better estimate of the $\widetilde{a}_k$

parameters is available to us, inverse filtering $y(n)$ by $\widetilde{A}(z)$ definitely introduces an error term $T^{-1}[\widetilde{A}(z)V(z)]$ and (4.19) can be rewritten as

$$\begin{aligned}\widetilde{\Re}(n) &= \left[x(n)+\sum_{k=1}^{p}\widetilde{a}_k x(n-k)\right]+\left[v(n)+\sum_{k=1}^{p}\widetilde{a}_k v(n-k)\right] \\ &= \widetilde{\Re}_x(n)+\widetilde{\Re}_v(n). \end{aligned} \tag{4.20}$$

Similar to $\Re(n)$ in (4.18), $\widetilde{\Re}_x(n)$ in $\widetilde{\Re}(n)$ is expected to maintain a close resemblance to a periodic impulse train excitation, whereas $\widetilde{\Re}_v(n)$ represents an error term in $\widetilde{\Re}(n)$ arising because of the noise. In order to reduce the effect of $\widetilde{\Re}_v(n)$ on $\widetilde{\Re}(n)$, a noise-compensation scheme is introduced in the ACF domain. To this end, the ACF of the RS $\widetilde{\Re}(n)$ can be written as

$$r_{\widetilde{\Re}}(m) = r_x(m)+r_v(m) \tag{4.21}$$

where $r_x(m)$ and $r_v(m)$ are the ACFs of $\widetilde{\Re}_x(n)$ and $\widetilde{\Re}_v(n)$, respectively and the cross correlation between $\widetilde{\Re}_x(n)$ and $\widetilde{\Re}_v(n)$ have been neglected assuming $x(n)$ is uncorrelated with noise $v(n)$. The ACF $r_x(m)$ of $\widetilde{\Re}_x(n)$ would follow a periodic pattern similar to that of $\widetilde{\Re}_x(n)$ and exhibits significant values at the origin and multiples of the pitch period T_0. On the other hand, it is seen from (4.20) that the error term $\widetilde{\Re}_v(n)$ can be treated as a p-th order moving average (MA) sequence and its ACF $r_v(m)$ can be expressed as

$$r_v(m) = \begin{cases} \sigma_v^2 \displaystyle\sum_{k=0}^{p-m}\widetilde{a}_k\widetilde{a}_{k+m}, & |m|\le p \\ 0, & \text{otherwise,} \end{cases} \tag{4.22}$$

where σ_v^2 represents the variance of the noise $v(n)$ and, clearly, $r_v(m)$ vanishes after p lags. This finite duration property of the MA correlation sequence $r_v(m)$ can be utilized to reduce the effect of $\widetilde{\Re}_v(n)$ on the RS $\widetilde{\Re}(n)$. Assuming that the ACF $r_x(m)$

of $\widetilde{\Re}_x(n)$ has negligible values at $|m| = p$, from (4.21) and (4.22), an estimate of the noise variance can be obtained as

$$\widetilde{\sigma}_v^2 = \frac{r_{\widetilde{\Re}}(m)}{\sum_{k=0}^{p-m} \widetilde{a}_k \widetilde{a}_{k+m}}, \quad 0 < m \leq p \tag{4.23}$$

where $m = p$ is used. Thus, by using the estimate of $\widetilde{\sigma}_v^2$ and the identified VTS parameters $\widetilde{a}_k$ in (4.21), noise compensation can be performed for $|m| \leq p$ and the resulting noise compensated ACF $\widetilde{r}_x(m)$ of the RS $\widetilde{\Re}(n)$ is given by

$$\widetilde{r}_x(m) = \begin{cases} r_{\widetilde{\Re}}(m) - \widetilde{\sigma}_v^2 \sum_{k=0}^{p-m} \widetilde{a}_k \widetilde{a}_{k+m}, |m| \leq p \\ r_{\widetilde{\Re}}(m), \ |m| > p. \end{cases} \tag{4.24}$$

In Fig. 4.5, the ACF of RS $\Re(n)$ and the noise-compensated ACF $\widetilde{r}_x(m)$ of the RS $\widetilde{\Re}(n)$ are plotted for a voiced frame. It can be observed that compared to $\widetilde{\Re}(n)$ in Fig. 4.4(b), which does not contain clearly distinguishable peaks at the GC instants, $\widetilde{r}_x(m)$ in Fig. 4.5(b) is less susceptible to noise and more suitable to be used in generating an ELS for pitch estimation.

It is to be noted that, in the case of colored noise, the noise term $\widetilde{\Re}_v(n) = T^{-1}[\widetilde{A}(z)V(z)]$ in (4.20) can also be modeled as a MA sequence with an order p' depending on the noise characteristics. In this case, the noise variance is no more constant. Assuming that $p' \leq p$, the ACF $r_v(m)$ of $\widetilde{\Re}_v(n)$ exhibits significant values within p lags and we can construct the noise-compensated ACF $\widetilde{r}_x(m)$ of $\widetilde{\Re}(n)$ by considering the ACF $r_{\widetilde{\Re}}(m)$ of $\widetilde{\Re}(n)$ as given by (4.21) for $m \geq p$.

4.4.2 Squared Hilbert Envelope of the ACF of RS as an ELS

It is found that the Hilbert Transform (HT) of the RS $\Re(n)$ of clean speech exhibits an unambiguous peak at the GC instant [74]. It is seen from Fig. 4.5(a) that the

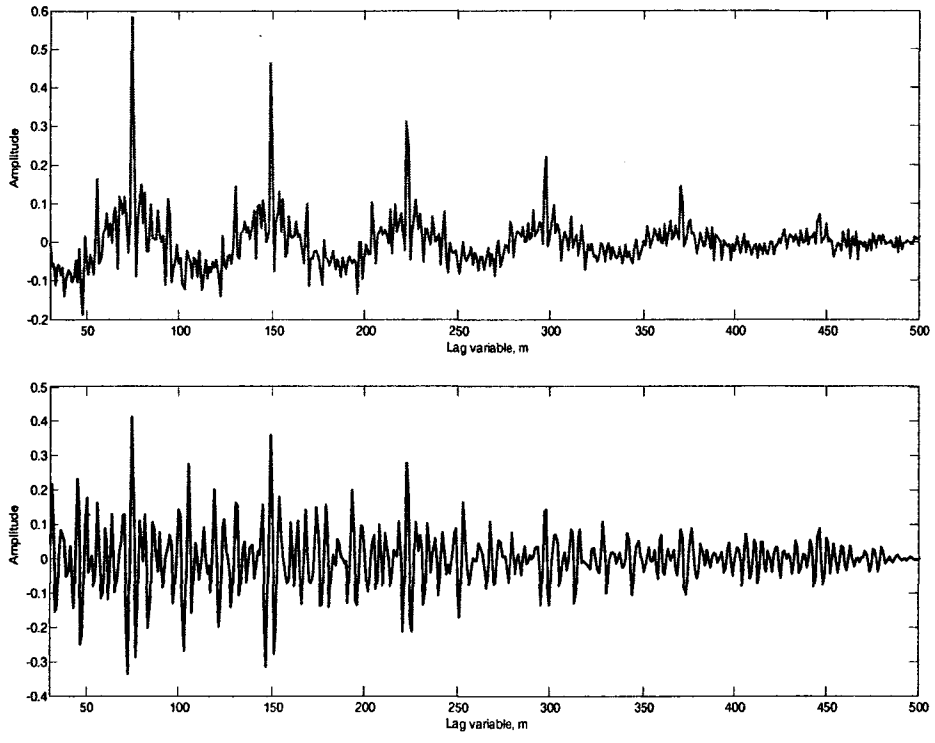


Figure 4.5: (a) ACF $r(m)$ of residual signal $\Re(n)$ (b) The noise-compensated ACF $\widetilde{r}_x(m)$ of the residual signal $\widetilde{\Re}(n)$.

ACF $r(m)$ of $\Re(n)$ contains more prominent peaks at the GC instants compared to that in $\Re(n)$ as plotted in Fig. 4.4(a). Hence, $r_h(n)$, the HT of the ACF $r(m)$ of $\Re(n)$, is expected to retain similar peaks at the GC instants, where $r_h(n)$ is defined as follows

$$r_h(n) = IDFT[R_h(k)], \tag{4.25}$$

where

$$R_h(k) = \begin{cases} -jR(k), & k = 0, 1, \ldots (N/2) - 1 \\ jR(k), & k = (N/2), (N/2) + 1, \ldots (N-1). \end{cases} \tag{4.26}$$

Here *IDFT* denotes the inverse discrete Fourier transform and $R(k)$ is the discrete Fourier transform of the ACF $r(n)$, where the lag variable of the ACF is changed

from m to n for convenience. The Hilbert transform $r_h(n)$ of the real signal $r(n)$ is also real, and $r(n)$ and $r_h(n)$ constitute, respectively, the real and imaginary parts of an analytic signal. Motivated by the features of $r(n)$ and the corresponding $r_h(n)$ as mentioned above, we propose a squared Hilbert envelope (SHE) of the ACF $r(n)$ of the RS as an ELS, which is defined as

$$E_r(n) = r^2(n) + r_h^2(n). \tag{4.27}$$

Even though $r(n)$ and $r_h(n)$ have positive and negative samples, according to (4.27), the ELS $E_r(n)$ of $r(n)$ is a positive function. In contrast to the RS $\Re(n)$, the ELS $E_r(n)$ in (4.27) is capable of removing the ambiguity caused by the phase angles of the formants in detecting the peaks at the GC instants. In Fig. 4.6 a frame of voiced speech, its RS $\Re(n)$, the ACF $r(n)$ of RS, and the ELS, i.e., the SHE $E_r(n)$ of $r(n)$, have been plotted. It is clearly observed from Fig. 4.6 that unlike $\Re(n)$, the ELS $E_r(n)$ exhibits unipolar impulse-like peak characteristics at the GC instants. Hence, the major peaks in $E_r(n)$ occurring at the GC instants can be exploited to determine the pitch period.

But in the presence of noise, the noise-compensated ACF $\widetilde{r}_x(n)$ of the RS that is obtained from the noisy speech is employed to generate the ELS. According to the definition of ELS given in (4.27), the ELS generated from $\widetilde{r}_x(n)$, i.e, the SHE of $\widetilde{r}_x(n)$, can be computed as

$$E_{\widetilde{r}}(n) = \widetilde{r}_x^2(n) + \widetilde{r}_{x_h}^2(n), \tag{4.28}$$

where $\widetilde{r}_{x_h}(n)$ is the Hilbert transform of $\widetilde{r}_x(n)$ obtained using (4.25) and (4.26). If $r_u(n)$ represents the error term remaining in $\widetilde{r}_x(n)$ even after noise-compensation, $\widetilde{r}_x(n)$ can be expressed as

$$\widetilde{r}_x(n) = r(n) + r_u(n). \tag{4.29}$$

By using the expression of $\widetilde{r}_x(n)$ given by (4.29) and exploiting the linearity property of the Hilbert transform, i.e.,

$$\widetilde{r}_{x_h}(n) = HT[r(n) + r_u(n)] = HT[r(n)] + HT[r_u(n)] = r_h(n) + r_{u_h}(n),$$

the ELS $E_{\widetilde{r}}(n)$ in (4.28) can be expressed as

$$\begin{aligned} E_{\widetilde{r}}(n) &= \{r(n) + r_u(n)\}^2 + \{r_h(n) + r_{u_h}(n)\}^2 \\ &= E_r(n) + [E_{r_u}(n) + E_c(n)] \\ &= E_r(n) + E_w(n) \end{aligned} \tag{4.30}$$

where

$$\begin{aligned} E_{r_u}(n) &= r_u^2(n) + r_{u_h}^2(n) \\ E_c(n) &= 2r(n)r_u(n) + 2r_h(n)r_{u_h}(n) \end{aligned}$$

In (4.30), $E_r(n)$, the SHE of $r(n)$, is given by (4.27). It is shown that in the presence of noise, the SHE $E_{\widetilde{r}}(n)$ of $\widetilde{r}_x(n)$ has an error term $E_w(n)$, which is composed of the SHE $E_{r_u}(n)$ of $r_u(n)$ and $E_c(n)$. In Fig. 4.7, the excitation-like signals $E_r(n)$ and $E_{\widetilde{r}}(n)$ obtained using (4.27) and (4.28), respectively, are plotted for a voiced frame. It can be observed from Fig. 4.7 that despite the introduction of the error term $E_w(n)$ in the presence of noise, $E_{\widetilde{r}}(n)$ closely matches $E_r(n)$ and demonstrates the prominently impulse-like nature at the GC instants up to a moderately low SNR.

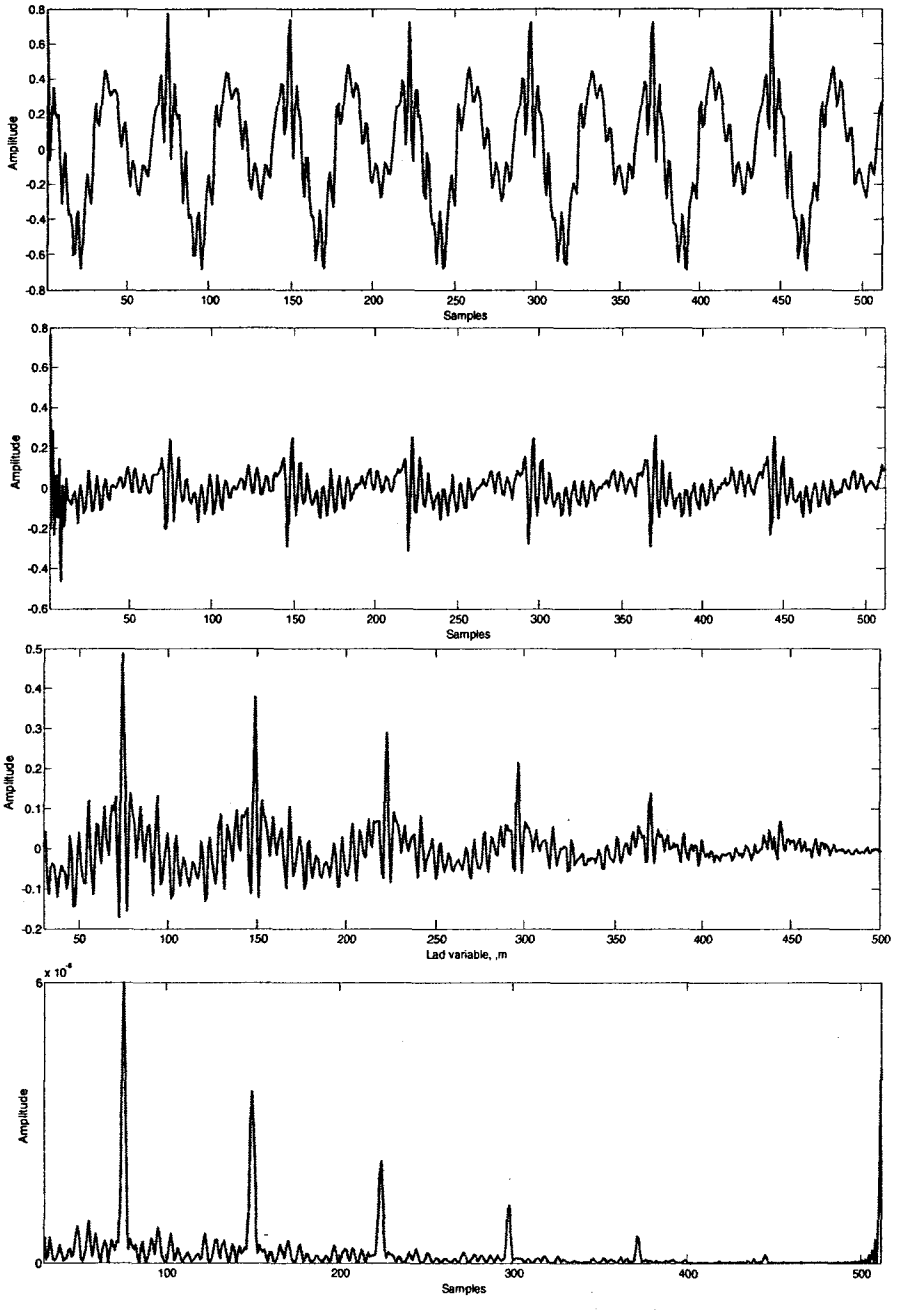


Figure 4.6: (a) A frame of voiced speech, (b) the residual signal $\Re(n)$, (c) the ACF $r(n)$ of residual signal, and (d) the ELS, the squared Hilbert envelope (SHE) $E_r(n)$ of $r(n)$.

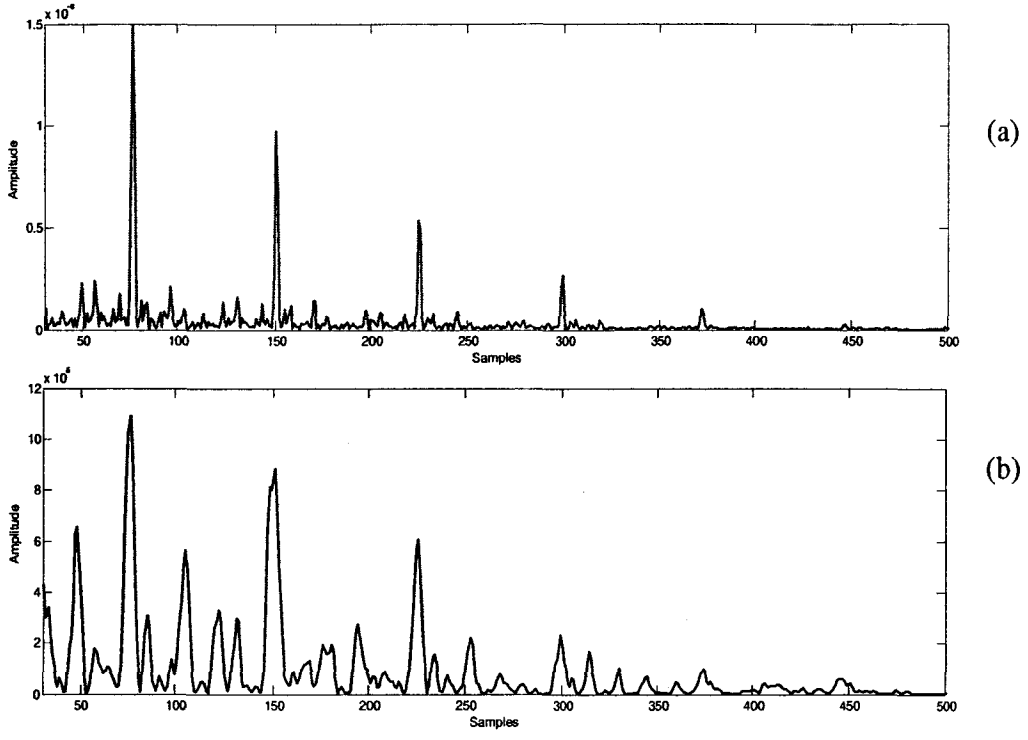


Figure 4.7: (a) The excitation-like signal $E_r(n)$ of $r(n)$, (b) the excitation-like signal $E_{\widetilde{r}}(n)$ of $\widetilde{r}_x(n)$.

4.5 Proposed Magnitude Difference Function of the ELS for Pitch Estimation

In a severe noisy condition, because of the pronounced effect of the error term $E_w(n)$ on the excitation-like signal $E_{\widetilde{r}}(n)$ in (4.30), the locations of the peaks of $E_{\widetilde{r}}(n)$ may provide only an approximation for the locations of the peaks at the GC events. Thus, the separation of the successive peaks of $E_{\widetilde{r}}(n)$ may not give information about the true pitch period. Hence, with a view to overcome the adverse effect of noise on the ELS $E_{\widetilde{r}}(n)$, we propose a new symmetric normalized magnitude difference function (SNMDF) of the ELS to determine an accurate pitch estimate under low SNR.

4.5.1 Symmetric Normalized Magnitude Difference Function

For the excitation-like signal $E_r(n)$, a time domain magnitude difference function is proposed as:

$$\psi(m) = \frac{\psi_{Num}(m)}{\psi_{Den}(m)} = \frac{\sum_{n=0}^{Q-1} |E_r(l) - E_r(n)|}{\sum_{n=0}^{K_m} |E_r(n)|}, \quad m \in [0, 1, \ldots, (Q-1)], \tag{4.31}$$

where the ELS $E_r(n)$ is the SHE of the ACF of RS $\Re(n)$ as defined in (4.27). In the numerator $\psi_{Num}(m)$ of $\psi(m)$,

$$l = mod(n + m, Q), \tag{4.32}$$

whereas in the denominator $\psi_{Den}(m)$ of $\psi(m)$, the upper limit K_m of the sum used for normalization operation is defined as

$$K_m = \left[\frac{Q}{2} - 1 + \left|\frac{Q}{2} - m\right|\right], \tag{4.33}$$

with Q representing the number of samples of $E_r(n)$ employed to compute $\psi(m)$ for every lag m. In the evaluation of (4.31), we choose $Q = N + m_{max}$, where m_{max} is the maximum possible pitch period of human speech. Since pitch usually changes slowly with time, such a choice of Q ensures a duration in which pitch parameter would not change significantly and at the same time at least two pitch periods would be covered to provide periodicity information for a wide range of speakers. By analyzing the terms on the right side of (4.31), we derive the following properties of $\psi(m)$:

1. At $m = 0$, for each term $|E_r(l) - E_r(n)|$ of the summation in the numerator $\psi_{Num}(m)$ of $\psi(m)$, we have $l = mod(n, Q) = n$ as $n \in [0 : (Q-1)]$ and $E_r(l) = E_r(n)$. Hence, $\psi(m)$ gives its minimum value $\psi(0) = 0$ at $m = 0$.

2. According to (4.32), by designing l in each term $|E_r(l) - E_r(n)|$ of $\psi_{Num}(m)$ as the modulo of $(n + m)$ with respect to Q, it can be shown that for $m \in [1 : \{(Q/2) - 1\}]$, $\psi_{Num}(m) = \psi_{Num}(Q - m)$. Thus, $\psi_{Num}(m)$ has an even symmetry about $m = m_s = Q/2$.

3. The upper limit K_m of the sum in the denominator $\psi_{Den}(m)$ governs the number of samples of the term $|E_r(n)|$ that has to be added for normalization operation at every lag m. According to (4.33), K_m is formulated such that for $m \in [1 : \{m_s - 1\}]$, $K_m = K_{(Q-m)}$. This yields an even symmetric normalizing function $\psi_{Den}(m)$ with a symmetry point m_s.

4. Because of the even symmetric nature of $\psi_{Num}(m)$ and $\psi_{Den}(m)$, it is imperative that $\psi(m)$ as defined by (4.31) has an even symmetry about m_s, i.e., for $m \in [1 : \{m_s - 1\}]$, $\psi(m) = \psi(Q - m)$. Thus, it is sufficient to compute $\psi(m)$ for $0 \leq m \leq m_s$ only. In view of this symmetry property as well as the normalizing operation, $\psi(m)$ is referred to as symmetric normalized difference function (SNMDF) of $E_r(n)$.

5. For voiced speech with periodic impulse-train excitation having a pitch period T_0, $E_r(n)$ exhibits periodic impulse-like characteristics at the GC events. Thus, $E_r(n)$ is similar to $E_r(n + \rho T_0)$, where ρ is a nonnegative integer, and value of $E_r(n)$ within the pitch period, i.e., $\rho T_0 < n < (\rho + 1)T_0$ is significantly smaller than that at $n = \rho T_0$. Using these facts and since $E_r(l) = E_r(mod(n+m, Q)) = E_r(n+m)$ for $(n+m) < Q$, it can be shown that at $m = \rho T_0$, $E_r(l) = E_r(n)$ for $Q' = (Q-\rho T_0)$ among Q terms of the numerator $\psi_{Num}(m)$, where $(n+\rho T_0) < Q$. On the other hand, $E_r(l) \neq E_r(n)$ for $\rho T_0 < m < (\rho + 1)T_0$ in each of the Q terms. As a result, $\psi_{Num}(m)$ at $m = \rho T_0$ has a smaller magnitude compared to

that within $\rho T_0 < m < (\rho + 1)T_0$. Moreover, the number of samples $[K_m + 1]$ of $|E_r(n)|$ added to constitute the normalizing function $\psi_{Den}(m)$ at $m = \rho T_0$ is higher than that within $\rho T_0 < m < (\rho+1)T_0$. Hence, the ratio of $\psi_{Num}(m)$ and $\psi_{Den}(m)$ is lower at $m = \rho T_0$ than that within $\rho T_0 < m < (\rho + 1)T_0$ yielding $\psi(\rho T_0) < \psi(\rho T_0 + m_b)$, where $0 < (\rho T_0 + m_b) \leq m_s$, $0 < m_b < T_0$. Clearly, the SNMDF $\psi(m)$ as defined by (4.31) exhibits local minima at $m = \rho T_0$, $\rho = 0, 1, \ldots \lfloor m_s/T_0 \rfloor$.

6. In the calculation of $\psi(\rho T_0)$, the contribution of $Q' = (Q - \rho T_0)$ terms is much lower than those of the remaining terms, where $E_r(l) \neq E_r(n)$. Hence, with increasing multiple ρ of T_0, $\psi_{Num}(\rho T_0) < \psi_{Num}((\rho + 1)T_0)$. Since K_m is monotonically decreasing with increasing m up to m_s, $K_{\rho T_0} > K_{((\rho+1)T_0)}$, making $\psi_{Den}(\rho T_0) > \psi_{Den}((\rho + 1)T_0)$. As a result, the ratio of $\psi_{Num}(\rho T_0)$ and $\psi_{Den}(\rho T_0)$ is ensured to be lower at $m = \rho T_0$ than that at $m = (\rho + 1)T_0$. This causes the local minima of $\psi(m)$ to increase when m increases with increasing multiple of T_0, i.e, $\psi(\rho T_0) < \psi((\rho + 1)T_0)$. This phenomenon prevents its minima from falling and thus avoiding the selection of a global minimum at a higher pitch-multiple.

In Fig. 4.8, the SNMDF $\psi(m)$ of $E_r(n)$ is plotted up to m_s for strongly and weakly voiced speech frames. The properties of SNMDF discussed above are evident in these plots through the fact that the SNMDF exhibits minima at lags equivalent to integer multiples of pitch period and the values of the front minima are deeper than those of the rear. From our detailed experimentations, it has been found that properties of the SNMDF remain valid for both strongly and weakly voiced frames.

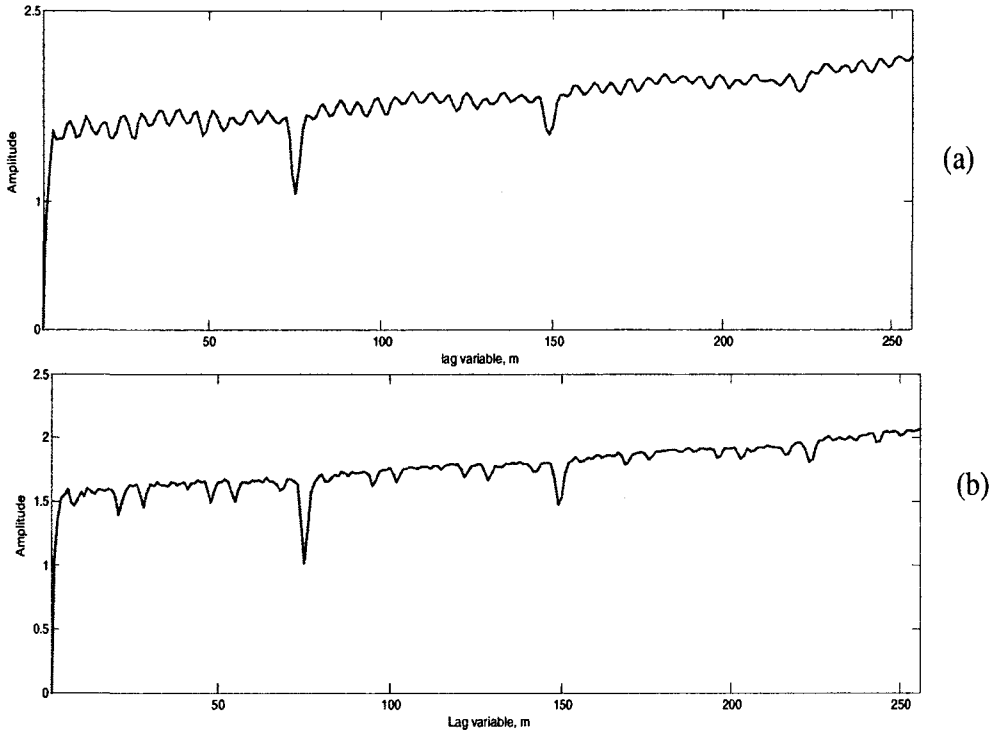


Figure 4.8: The SNMDF $\psi(m)$ of $E_r(n)$ for (a) a strongly (b) a weakly voiced speech frame.

These observations indicate that the SNMDF of $E_r(n)$ possesses useful properties that can be exploited to estimate pitch period T_0 (pitch frequency F_0) from the location of its global minimum at $m = T_0$. The properties of the SNMDF as illustrated through examples in Fig. 4.8 concern $E_r(n)$ of clean speech signals. In practical conditions, however, one has to extract the pitch information from $E_r(n)$ of speech signals corrupted by noise.

In the presence of noise, given Q samples of the excitation-like signal $E_{\widetilde{r}}(n)$, its SNMDF $\widetilde{\psi}(m)$ can be computed using (4.31) as

$$\widetilde{\psi}(m) = \frac{\sum_{n=0}^{Q-1} |E_{\widetilde{r}}(l) - E_{\widetilde{r}}(n)|}{\sum_{n=0}^{K_m} |E_{\widetilde{r}}(n)|}, \quad m \in [0, 1, \ldots, m_s] \tag{4.34}$$

where the ELS $E_{\widetilde{r}}(n)$ as defined in (4.28) is the SHE of the noise-compensated ACF $\widetilde{r}_x(n)$ of the RS that is obtained from the noisy speech. Inserting the expression of the ELS $E_{\widetilde{r}}(n)$ given by (4.30) into (4.34), we have

$$\widetilde{\psi}(m) = \frac{\sum_{n=0}^{Q-1} |\{E_r(l) - E_r(n)\} + \{E_w(l) - E_w(n)\}|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|}, \quad m \in [0, 1, \ldots, m_s]. \tag{4.35}$$

From the above expression, an inequality can be formulated as

$$\widetilde{\psi}(m) \leq \frac{\sum_{n=0}^{Q-1} |E_r(l) - E_r(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|} + \frac{\sum_{n=0}^{Q-1} |E_w(l) - E_w(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|}, \quad m \in [0, 1, \ldots, m_s] \tag{4.36}$$

By introducing an equality term equal to or greater than zero, (4.36) can be expressed as (see Appendix A for details)

$$\widetilde{\psi}(m) = \psi(m) + \widetilde{\Gamma}(m), \quad m \in [0, 1, \ldots, m_s]; \tag{4.37}$$

here $\psi(m)$, the SNMDF of $E_r(n)$, is given by (4.31) and $\widetilde{\Gamma}(m)$ is a term introduced in the presence of noise. It can be verified that $\widetilde{\Gamma}(m)$ vanishes in the absence of noise. We will now conduct experiments to examine and illustrate the behavior of $\widetilde{\psi}(m)$ in comparison to that of $\psi(m)$ in (4.31). We used voiced frames of different strengths corrupted by different types and levels of noises as a platform of this experiment.

In Fig. 4.9, we show examples of the typical behavior of $\widetilde{\psi}(m)$ observed from our experiment. Plots in Figs. 4.9 (a) and (b) correspond to a strongly voiced frame at SNR $= -5$ dB and those in Figs. 4.9(c) and (d) correspond to a weakly voiced frame at SNR $= -10$ dB. The type of noise embedding the speech is white in the case of Figs. 4.9(a) and (c), whereas it is a multi-talker Babble noise in the case of Figs. 4.9(b) and (d). It is evident from the figures that minima of $\psi(m)$ at multiples of a pitch period are well-preserved and remain prominent in $\widetilde{\psi}(m)$ whether the frame under consideration is strongly or weakly voiced, or whether the corruption is due to the presence of a white or a Babble noise regardless of its level. The property of exhibiting high-value minima with increasing pitch-multiple that is clarified for $\psi(m)$ is clearly retained in $\widetilde{\psi}(m)$, thus justifying the effectiveness of $\widetilde{\psi}(m)$ in extracting the pitch period from the severely noise-corrupted speech. By searching the location of the global minimum from the minima of $\widetilde{\psi}(m)$ in the possible pitch period range $[m_{min} : m_{max}]$, an estimate of T_0 can be obtained as

$$\widetilde{T_0} = \underset{m}{arg\,\min}[\widetilde{\psi}(m)] \tag{4.38}$$

and the corresponding estimate of pitch frequency in Hz can be determined as

$$\widetilde{F_0} = \frac{F_s}{\widetilde{T_0}}. \tag{4.39}$$

The complete method of pitch estimation for a time frame whose development started in the Section 4.2 will henceforth be referred to as the SNMDF-ELS method.

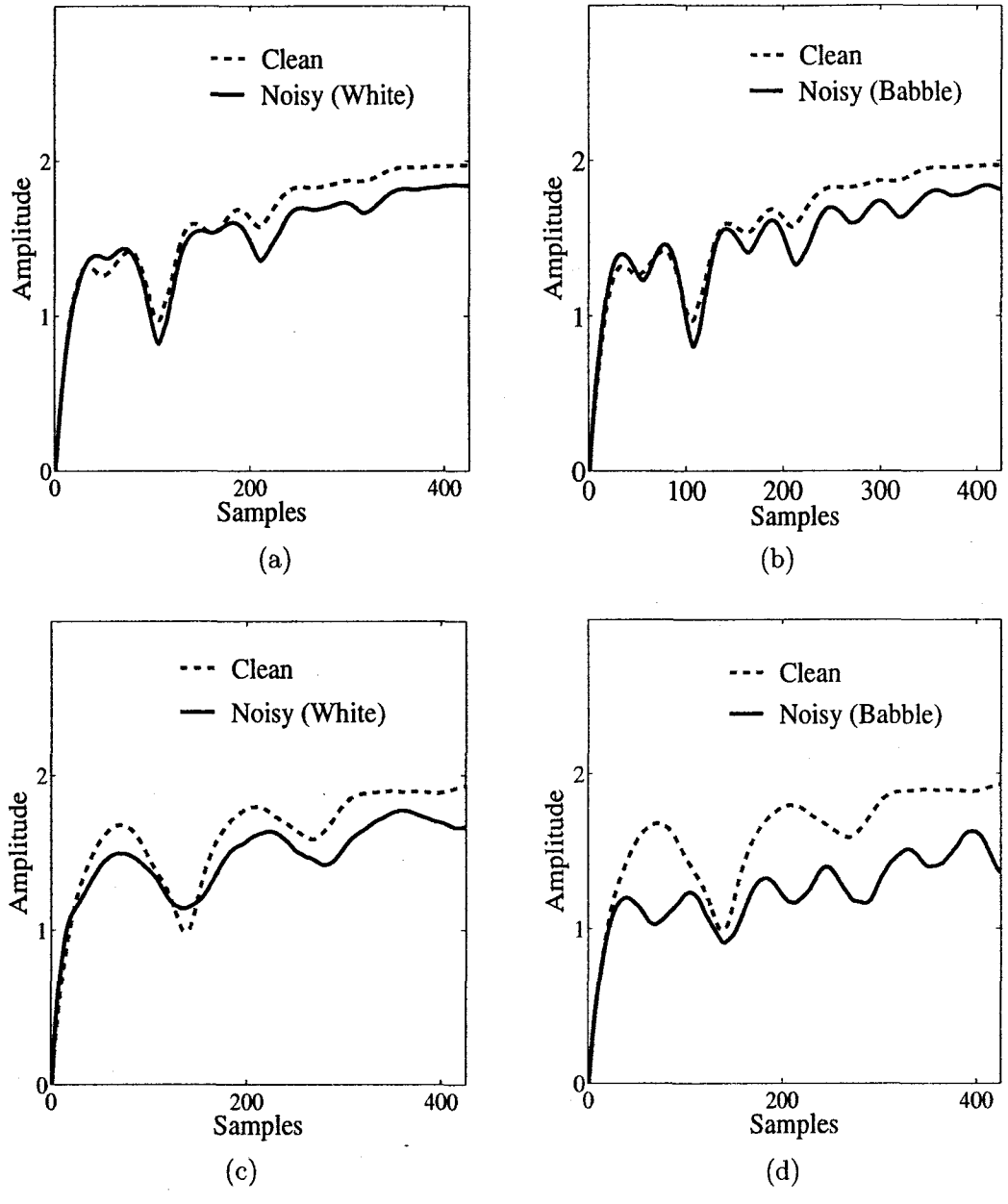


Figure 4.9: Plots for $\widetilde{\psi}(m)$ for typical voiced frames considering different levels and types of noise and strengths of voiced frames. A strongly voiced frame at SNR $= -5$ dB under (a) white noise and (b) Babble noise. A weakly voiced frame at SNR $= -10$ dB under (c) white noise and (d) Babble noise.

4.5.2 SNMDF Based Pitch Tracking using Dynamic Programming

In this Section, a pitch tracking scheme proposed in Chapter 2 can easily be employed to reduce the error in the F_0 contour. It is important to keep in mind the obvious changes required for candidate generation and cost computation due to the different approach of the development of the SNMDF-ELS method.

In the pitch tracking scheme of this Chapter, the pitch period $\widetilde{T}_0$ as given in (4.38), determined from the global minimum of the SNMDF $\widetilde{\psi}(m)$, is selected to be one of the members of the set of potential pitch candidates for a frame. For a given frame t, a certain number, say $(J_t - 1)$, of local minima of the SNMDF $\widetilde{\psi}(m)$ given by (4.34) lying within the pitch period range but excluding the one located at or nearest to the lag $m = \widetilde{T}_0$ are chosen, and the corresponding locations of these minima are selected as the other possible pitch candidates. Among the J_t number of potential candidates at frame t, $\widetilde{T}_0$ is assigned the highest priority to be the correct pitch value for the frame under consideration. The remaining $(J_t - 1)$ candidates are prioritized by sorting them according to increasing magnitude of the corresponding values of $\widetilde{\psi}(m)$. By letting $F^j_{0,t}$ represent the j-th pitch frequency candidate in the t-th frame, the local cost for the candidate $F^j_{0,t}$ is defined as

$$C_{local}(F^j_{0,t}) = \widetilde{\psi}(F^j_{0,t}) \tag{4.40}$$

where $\widetilde{\psi}(F^j_{0,t})$ is the value of the SNMDF evaluated at $F^j_{0,t}$. Obviously, the pitch frequency candidate having a lower score of $\widetilde{\psi}(F^j_{0,t})$ will result in a lower local cost $C_{local}(F^j_{0,t})$. Considering the deviation among pitch candidates from frame to frame, similar to (3.38), a transition cost measuring the cost of the pitch path going from

the i-th candidate of frame $(t-1)$ to the j-th candidate of frame t is computed as

$$C_{tran}(F^i_{0,t-1}, F^j_{0,t}) = \left| \frac{F^i_{0,t-1}}{F^j_{0,t}} \right|. \tag{4.41}$$

Considering that there are O_c consecutive frames, i.e., $t = 1, 2, \ldots, O_c$, similar to (3.39), the total cost function $C(A)$ corresponding to an arbitrarily chosen trajectory $A = \left\{ F^{j_1}_{0,1}, F^{j_2}_{0,2}, \ldots, F^{jO_c}_{0,O_c} \right\}$ is defined as

$$C(A) = C_{local}(F^{j_1}_{0,1}) + \sum_{t=2}^{O_c} \left[\varpi . C_{tran}(F^{j_{t-1}}_{0,t-1}, F^{j_t}_{0,t}) + C_{local}(F^{j_t}_{0,t}) \right] \tag{4.42}$$

and the task of pitch tracking is now to minimize this total cost function. Due to the suitability of dynamic programming (DP) in handling such a problem, we employ DP for efficiently minimizing the total cost function for pitch tracking [48], [59]. For this purpose, we use only the first three pitch candidates from the prioritized set for each frame [62].

4.6 Simulation Results

In this Section, computer simulations are conducted in order to demonstrate the effectiveness of the proposed method in estimating the pitch in the presence of noise. We investigate the pitch estimation performance for natural speech signals corrupted by additive noise. The estimation performance of the proposed SNMDF-ELS method in terms of the accuracy and consistency of the estimated pitch is obtained and compared with that of some of the state-of-the-art pitch estimation methods.

4.6.1 Simulation Conditions

(a) Database and other details: The proposed SNMDF-ELS method for pitch estimation is tested using speech signals from the *Keele* database [63], [64] as done

in the previous chapters. Each speech utterance in the database is divided into frames, each of size $N = 25.6$ ms at a frame rate of 100 Hz (i.e., a frame shift of 10 ms). In order to imitate a condition for a noisy environment, the noisy speech $s_y(n)$ is generated according to (2.26) by adding noise to the original clean speech $s_x(n)$. White or multi-talker babble noise is used in the simulation and the *Noisex92* database [65] is adopted as their source. In the proposed method, pitch estimation is performed on a frame by frame basis. In order to carry out the short-time analysis on the observed noisy speech $s_y(n)$, first windowing is performed so as to reduce the edge effects at the beginning and the end of the frame. An N-sample normalized Hamming window is used to obtain an windowed noisy frame, $y'(n)$. In our simulations, we used the same values for the parameters, such as frame rate and basic frame (or window) size as specified in the Keele database. Since voiced speech spectra normally have a roll-off about -6 dB/Octave, they are tilted into a slightly low-pass form [1]. In order to reduce the natural spectral tilt of the windowed speech, a high-pass pre-emphasis filter with the following input output relation is employed

$$y''(n) = y'(n) - u_p y'(n), \tag{4.43}$$

where $y''(n)$ is the filtered output with respect to the input $y'(n)$ with a pre-emphasis factor $0 < u_p < 1$. The effect of applying a pre-emphasis filter can be viewed as introducing an extra zero into the transfer function of the vocal tract filter. Note that the introduction of such a zero neither alters the pole locations in the transfer function plane, nor alters their resonance center frequencies and bandwidths in the associated frequency response. In addition to signal pre-emphasis, a low-pass filtering is performed to remove the effect of very high frequencies (> 5 kHz) while preserving 4-5 formants (one formant per kHz) that helps the VTS parameter identification

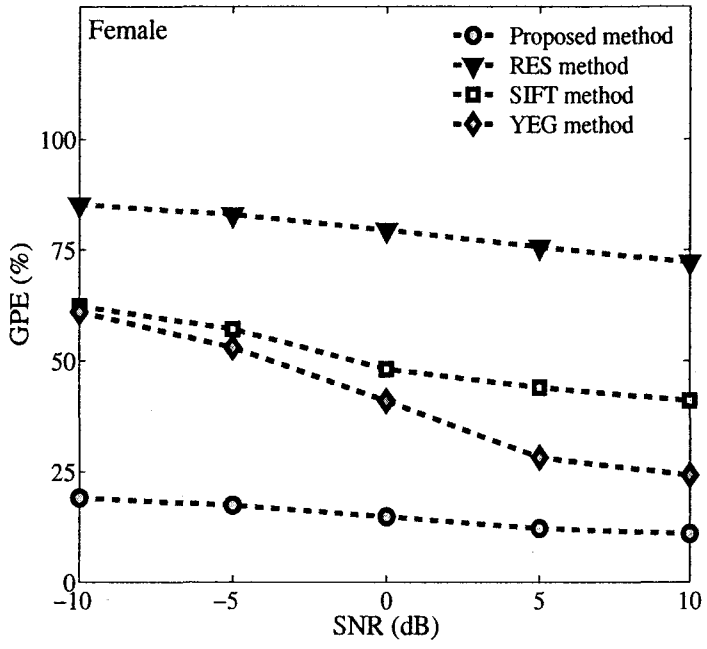


Figure 4.10: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in white noise.

required for the ELS generation. Low-pass filtering can be performed by using N-point FFT and IFFT operations.

As the homomorphic deconvolution principle, the standard cepstrum analysis [89] is employed on the noisy ACF $\phi_y(m)$ because of its straightforwardness in implementation over the others (e.g., [94], [95], [96]). In the real cepstrum $c_{\phi_y}(m)$ of $\phi_y(m)$ as given by (4.14), the effect of the term $c_{\phi_w}(m)$ arising because of the noise can be reduced by replacing the actual value $\phi_y(0)$ by a smaller value while computing $c_{\phi_y}(m)$ from $\phi_y(m)$. Noting that $\phi_y(0) > |\phi_y(m)|$ for $m \neq 0$, we replace $\phi_y(0)$ by $u_n\phi_y(0)$ with $\frac{|\phi_y(1)|}{\phi_y(0)} \leq u_n < 1$. This process efficiently suppresses the level of $c_{\phi_w}(m)$ in $c_{\phi_y}(m)$. In the liftering operation for removing the effect of the excitation component

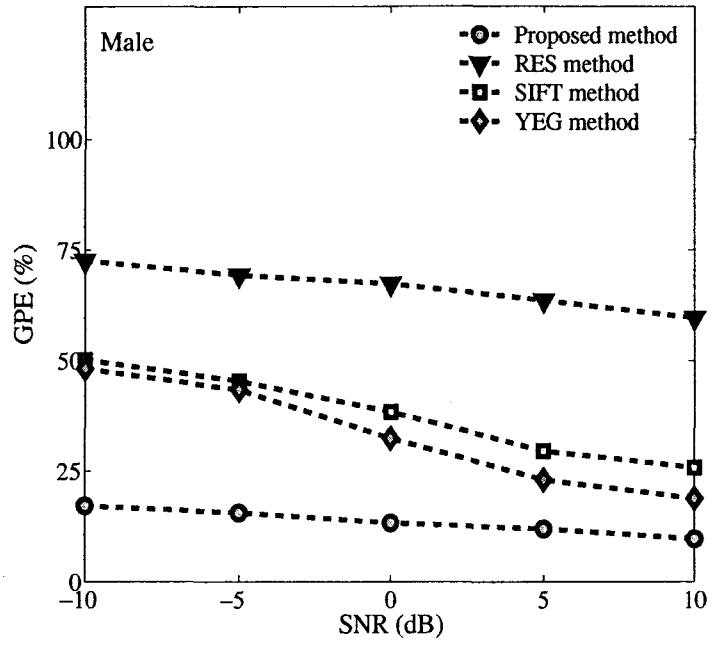


Figure 4.11: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in white noise.

$c_{\phi_e}(m)$ from $c_{\phi_y}(m)$, rather than using a fixed length cepstral window independent of T_0 of the underlying speech, a larger cepstral window of 4.5 milliseconds and a shorter cepstral window of 2.5 milliseconds are adopted as more rational choices for male and female speakers, respectively. A 12th order ($p = 12$) autoregressive (AR) model is assumed to characterize the VTS filter. In the computation of the p number of VTS $\tilde{a}_k$ parameters by using (4.16), a few lower lags of the ACF are avoided and a combination of more than p equations is employed in order to handle the noisy environment. The number of equations in (4.16) is governed by S, and we have chosen $S = 5p$.

(b) Metrics and Comparison Methods Used for Performance Evaluation: The performance metrics considered here, as defined in Chapter 2, are 1)

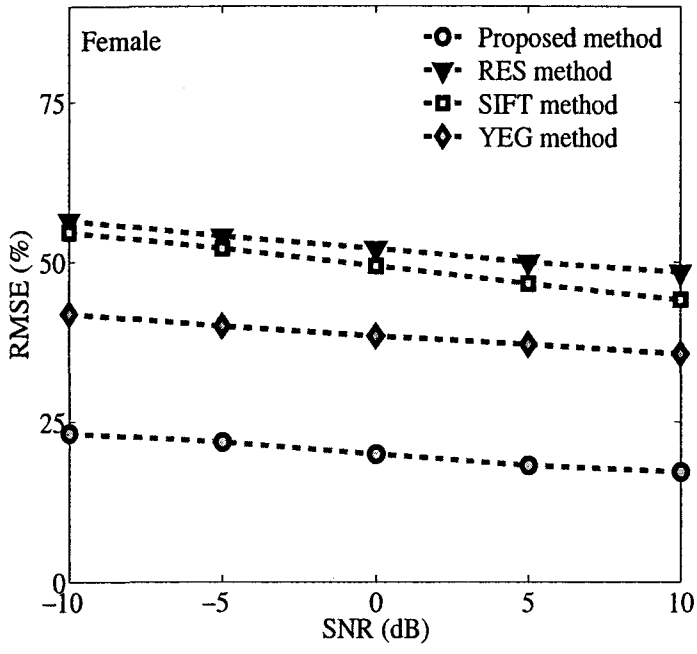


Figure 4.12: RMSE (%) as a function of SNR for female speaker group in white noise.

the percentage gross pitch error (PGPE); 2) the mean of fine pitch error (m_{FPE}); 3) the standard deviation of fine pitch error (σ_{FPE}); and 4) the root-mean-square-error (RMSE). At different SNR levels, for the performance comparison of the pitch estimates obtained by using the proposed SNMDF-ELS method at the voiced frames based on the voiced/unvoiced labels provided by the *Keele* database, we consider the residual (RES) method [41], the simple inverse filter tracking (SIFT) method [23] and the YEG method [47]. All the methods used for comparison follow the basic strategy of pre-whitening the speech and the pre-whitening step involves the use of the LP based inverse filter. We implemented the methods independently using the default parameters specified by the authors.

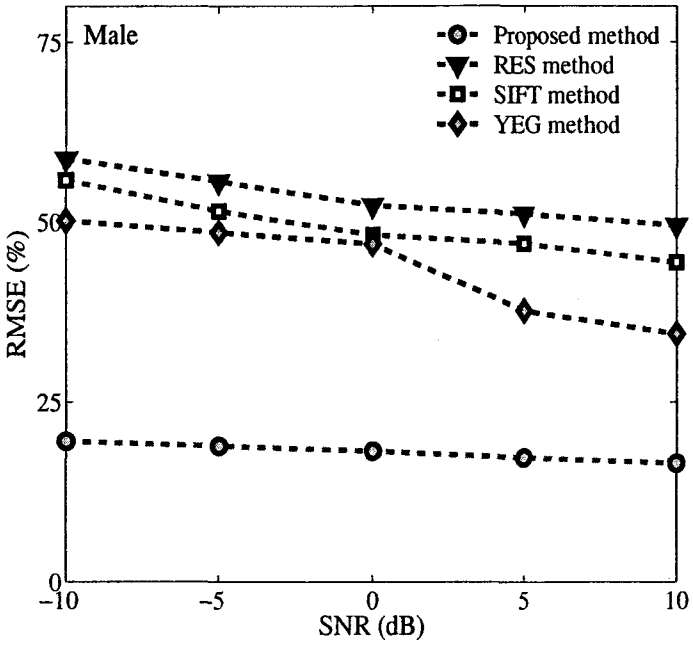


Figure 4.13: RMSE (%) as a function of SNR for male speaker group in white noise.

4.6.2 Results and Comparisons

(a) Results on white-noise corrupted speech: The pitch estimation results obtained from the RES, SIFT, YEG and the proposed SNMDF-ELS methods are investigated at first for the white-noise corrupted speech signals. Figs. 4.10 and 4.11 show the PGPE values as a function of SNR obtained from these methods for the female (5) and male (5) speaker groups, respectively, where the SNR varies from a very low value of -10 dB to a high value of 10 dB. Note that the RES and SIFT methods are for the estimation of pitch in the clean speech and thus not expected to perform well under a low SNR. It is seen from these figures that the estimation results of the RES method are unsatisfactory for the whole range of SNR. At SNR $= 10$ dB, the SIFT method outperforms the RES method drastically but provides not

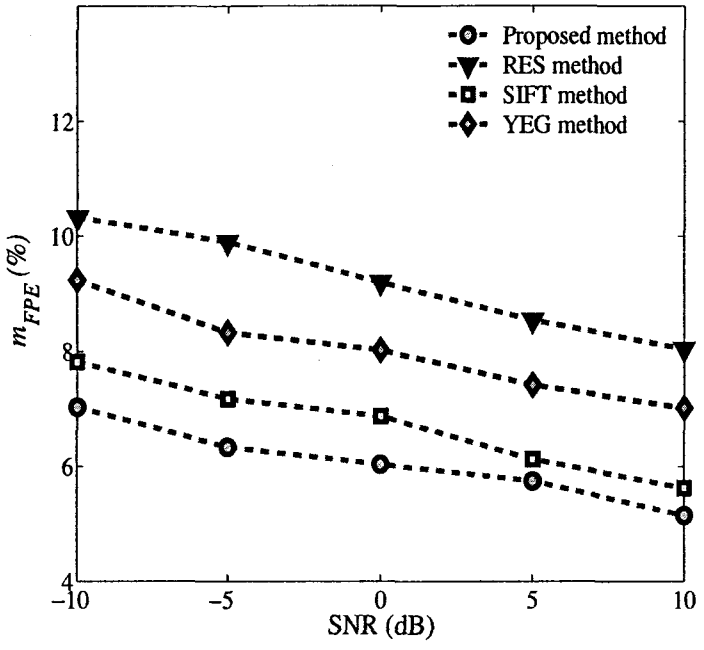


Figure 4.14: m_{FPE} (%) as a function of SNR for all female and male speakers in white noise.

much higher PGPE values compared to the YEG method. At such a level of SNR, although the SIFT and YEG methods show larger errors, the estimation accuracy achieved by the proposed method is much higher. It is also seen from these figures that the YEG method performs better than the SIFT method above SNR = 0 dB but both the methods show a poor performance when the SNR is low. Even at an SNR as low as −10 dB, when the other methods fail to estimate the pitch, the proposed method shows significantly smaller values of PGPE.

The variation of RMSE with respect to the level of SNR are plotted in Figs. 4.12 and 4.13 for all four methods using the same female and male speaker groups as above. It is observed from these figures that the RES method completely fails to estimate pitch not only at a very low SNR but also at a high SNR of 10 dB. The SIFT and

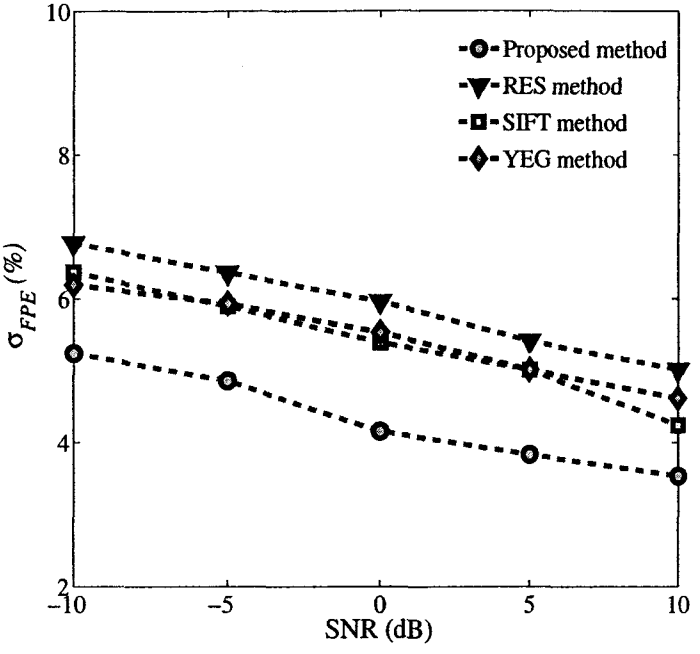


Figure 4.15: σ_{FPE} (%) as a function of SNR for all female and male speakers in white noise .

YEG methods provide better RMSE values compared to the RES method. However, their estimation errors are much higher over the SNR range considered in comparison to that of the proposed method. Clearly, the proposed method is able to estimate the pitch quite accurately even at a low level of SNR (as low as -10 dB).

Figs. 4.14 and 4.15 depict the mean m_{FPE} and standard deviation σ_{FPE} resulting from the four methods for the set of 10 mixed (5 female plus 5 male) speakers of the database. These figures demonstrate that the overall mean and standard deviation values of the fine-pitch errors obtained by using the proposed SNMDF-ELS method are lower throughout the whole range of SNR. Much smaller values of PGPE and RMSE achieved by using the proposed method under a wide range of SNR levels, along with lower overall mean and standard deviation of the fine-pitch errors, indicate

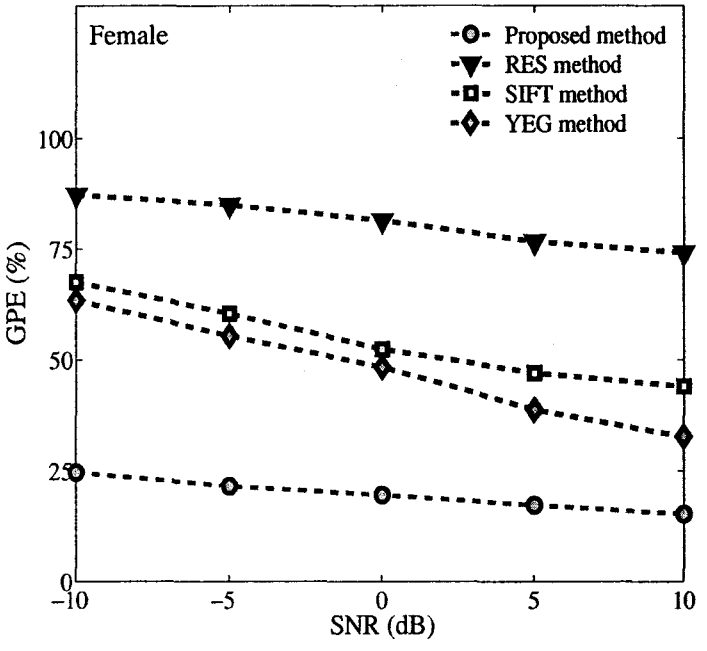


Figure 4.16: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in babble noise.

its high degree of estimation accuracy, robustness and consistency.

(b) Results on Multi-Talker Babble-Noise Corrupted Speech: We now examine the ability of the proposed SNMDF-ELS and the other three methods to deal with an environmental babble noise. The multiplicity of speakers in the babble noise produces a flatter short-term spectrum, which has greater spectral and temporal modulation than white noise. In Fig. 4.16 through Fig. 4.21, the pitch estimation results in terms of the PGPE, RMSE, mean m_{FPE}, and standard deviation σ_{FPE} for each of the methods are presented. The presence of multiple background competing speakers makes pitch estimation difficult and generally the performance of all the four methods degrades in the presence of babble noise compared to that in the white noise.

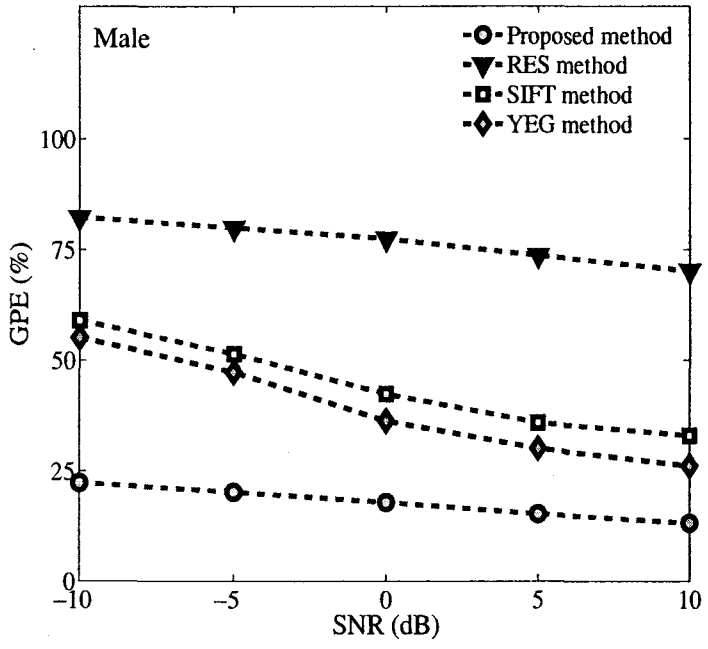


Figure 4.17: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in babble noise.

However, the proposed SNMDF-ELS method is still able to overcome the difficulty mentioned above, showing superiority with respect to all the four performance metrics at all the levels of SNRs for the same male and female speaker groups as the ones considered for the white-noise corrupted speech.

The PGPE values as a function of SNR obtained from the four methods for the female and male speaker groups, respectively, are portrayed in Figs. 4.16 and 4.17. It is seen that the proposed method continues to provide acceptable results at a high SNR, such as 10 dB. Also, the proposed method still maintains satisfactory performance at a very low level of SNR of -5 dB or even lower than that.

The plots for the RMSE resulting from using the four pitch estimation methods for the female and male speaker groups are, respectively, provided in Figs. 4.18 and

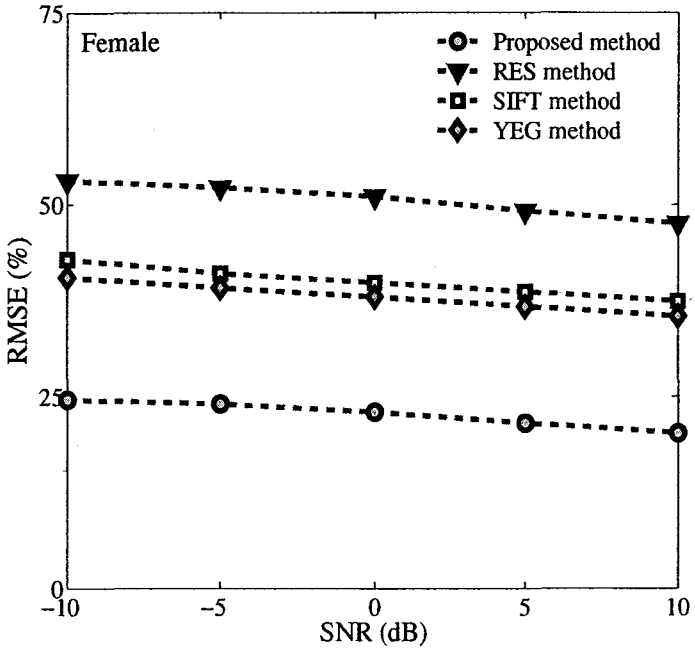


Figure 4.18: RMSE (%) as a function of SNR for female speaker group in babble noise.

4.19. It may be pointed out that the estimation error occurred in the case of the proposed SNMDF-ELS method remains significantly lower even for the low levels of SNR, such as -10 dB. As seen, the RMSE values obtained by the other three methods remain very high at SNR levels below 10 dB.

Figs. 4.20 and 4.21, respectively, plot the mean m_{FPE}, and standard deviation σ_{FPE} obtained from the four methods as a function of SNR for the same set of 10 mixed (5 female plus 5 male) speakers as the one used in Figs. 4.14 and 4.15. These figures testify that the estimation error of the proposed SNMDF-ELS method is increased as expected in comparison to the white noise case, but it still provides quite better results compared to that provided by the other three methods for the entire range of SNR considered.

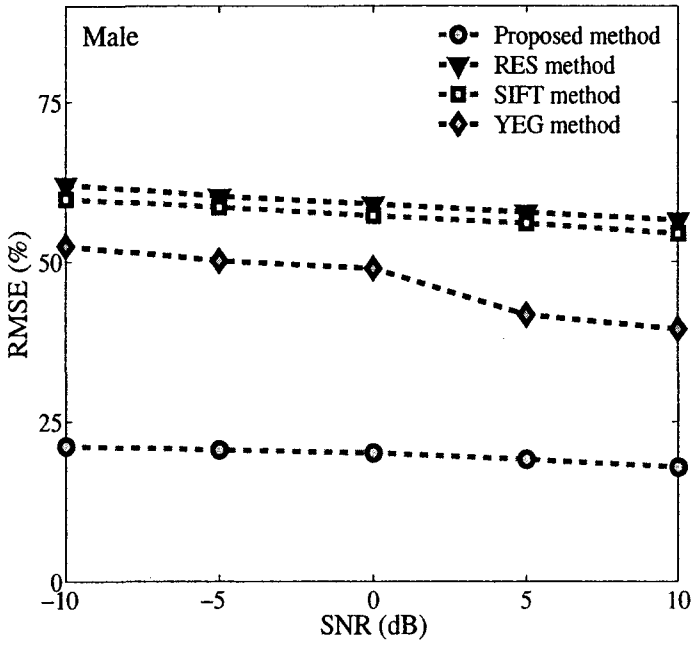


Figure 4.19: RMSE (%) as a function of SNR for male speaker group in babble noise.

In order to evaluate the net effect of pitch tracking by using dynamic programming (DP) on the pitch estimation performance of the proposed SNMDF-ELS method, finally we present the pitch contours obtained from the four pitch estimation methods. Fig. 4.22 plots a reference pitch contour for a 1.4-second excerpt of the clean speech of a male speaker from the reference database. The reference pitch contour is accompanied by the spectrogram of clean speech to clearly show the pitch track. Also, the pitch values estimated by the four methods at SNR of -10 dB in the presence of white noise are plotted on the spectrograms of the noise corrupted speech. If we compare the estimation accuracy, it is visible from the figure that the proposed method is able to track pitch and yields a comparatively smoother pitch contour in contrast to the other methods. Similarly, Fig. 4.23 shows a similar plot of the pitch contours

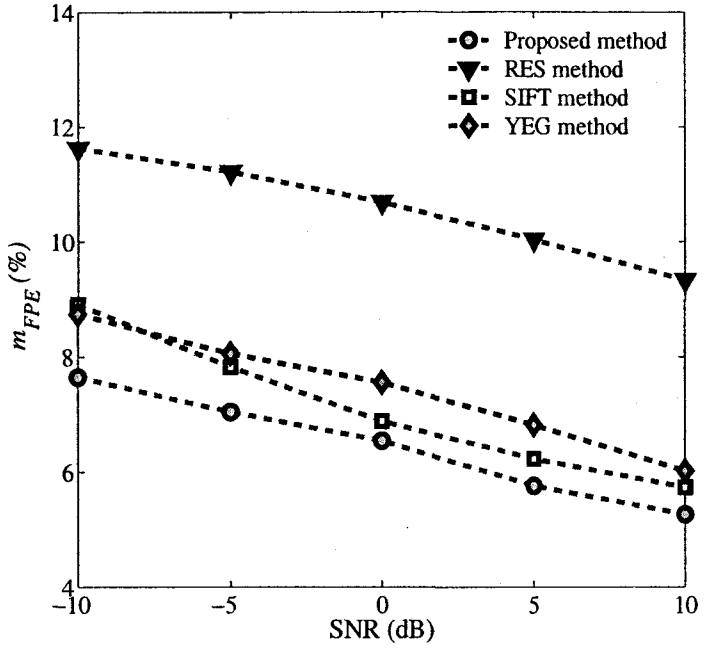


Figure 4.20: m_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

resulting from the four methods for female speech corrupted by babble noise at SNR −10 dB. From Fig. 4.23, it is clear that the proposed method is able to estimate pitch accurately giving a smoother contour even in the presence of babble noise. The pitch contours in Figs. 4.22 and 4.23 obtained from all the methods have convincingly illustrated that the SNMDF-ELS method is capable of reducing the double and half-pitch errors to a significant extent by the use of the proposed pitch tracking scheme. Thus, it is inferred that that the proposed method is suitable for real-life applications in a heavy noisy environment.

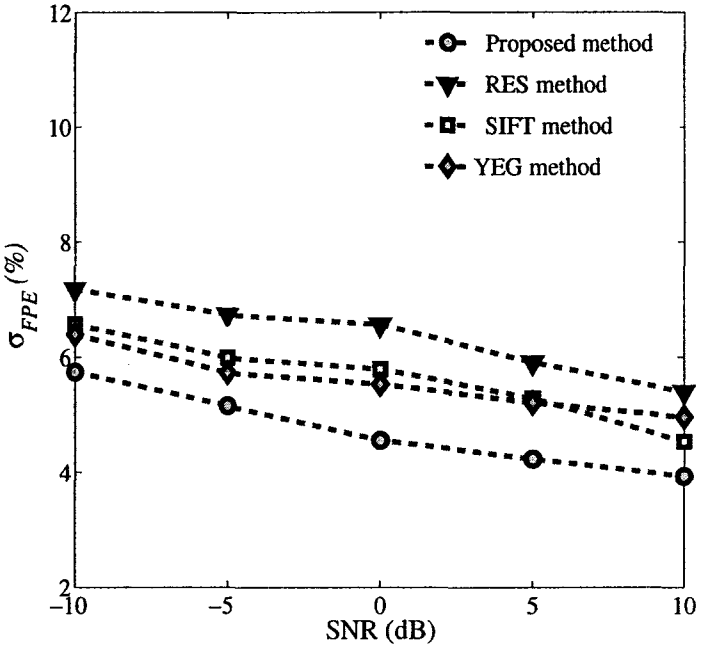


Figure 4.21: σ_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

4.7 Conclusion

In this chapter, apart from the model fitting based pitch estimation methods developed in Chapters 2 and 3, a new method for pitch estimation under noisy conditions has been presented. Instead of employing the noisy speech directly for pitch estimation, this method is developed by employing an ELS obtained from the noisy speech.

In order to overcome the limitation of the conventional ACF based LP residual, a time-frequency domain homomorphic deconvolution scheme has been devised to identify the VTS parameters from noisy speech. As a result of zero-lag compensation and ignoring few low-order lags of ACF, which suffer from more noise corruption than the high-order lags, more accurate VTS parameters are obtained and employed for

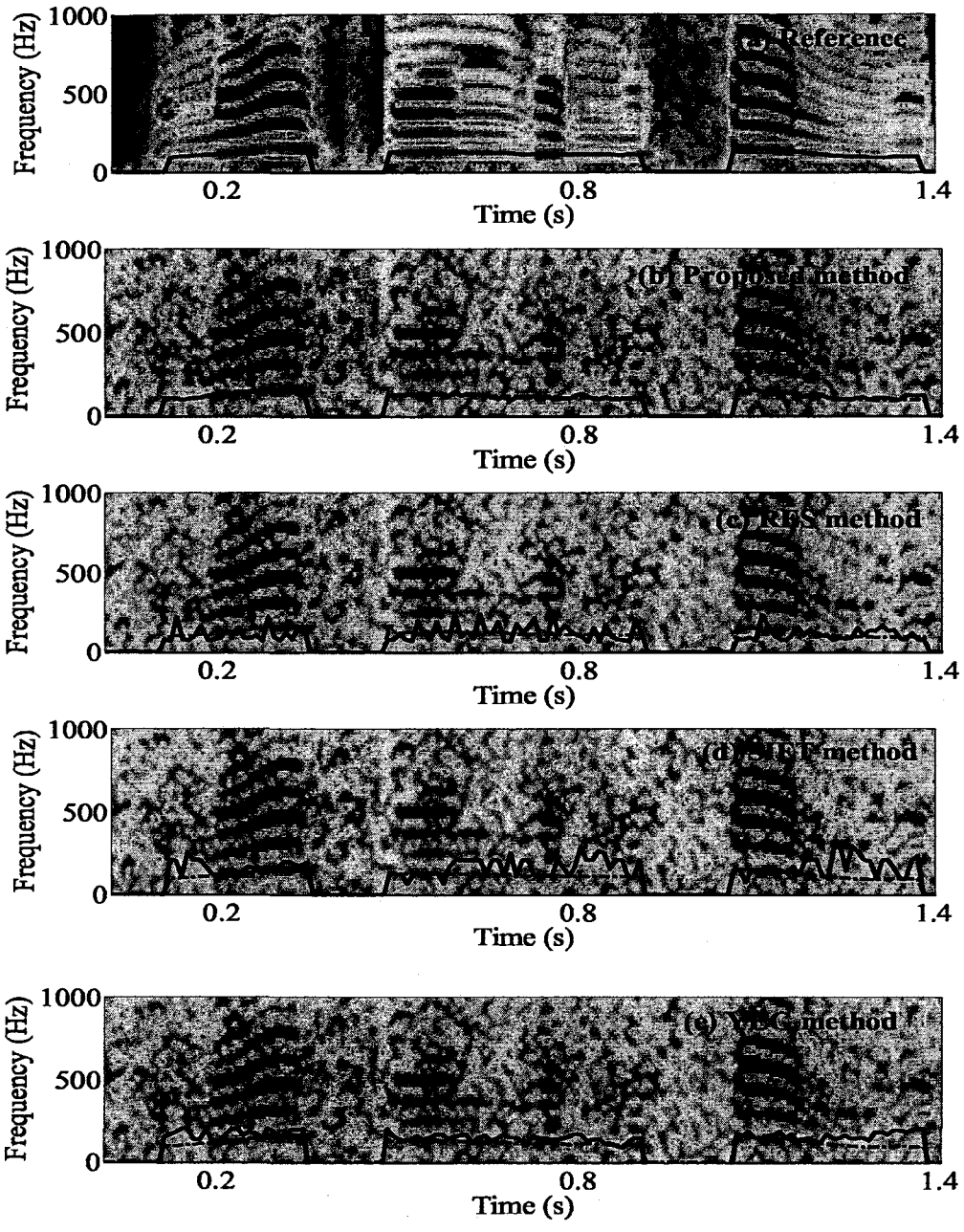


Figure 4.22: Pitch contours of different methods at SNR $= -10$ dB in white noise.

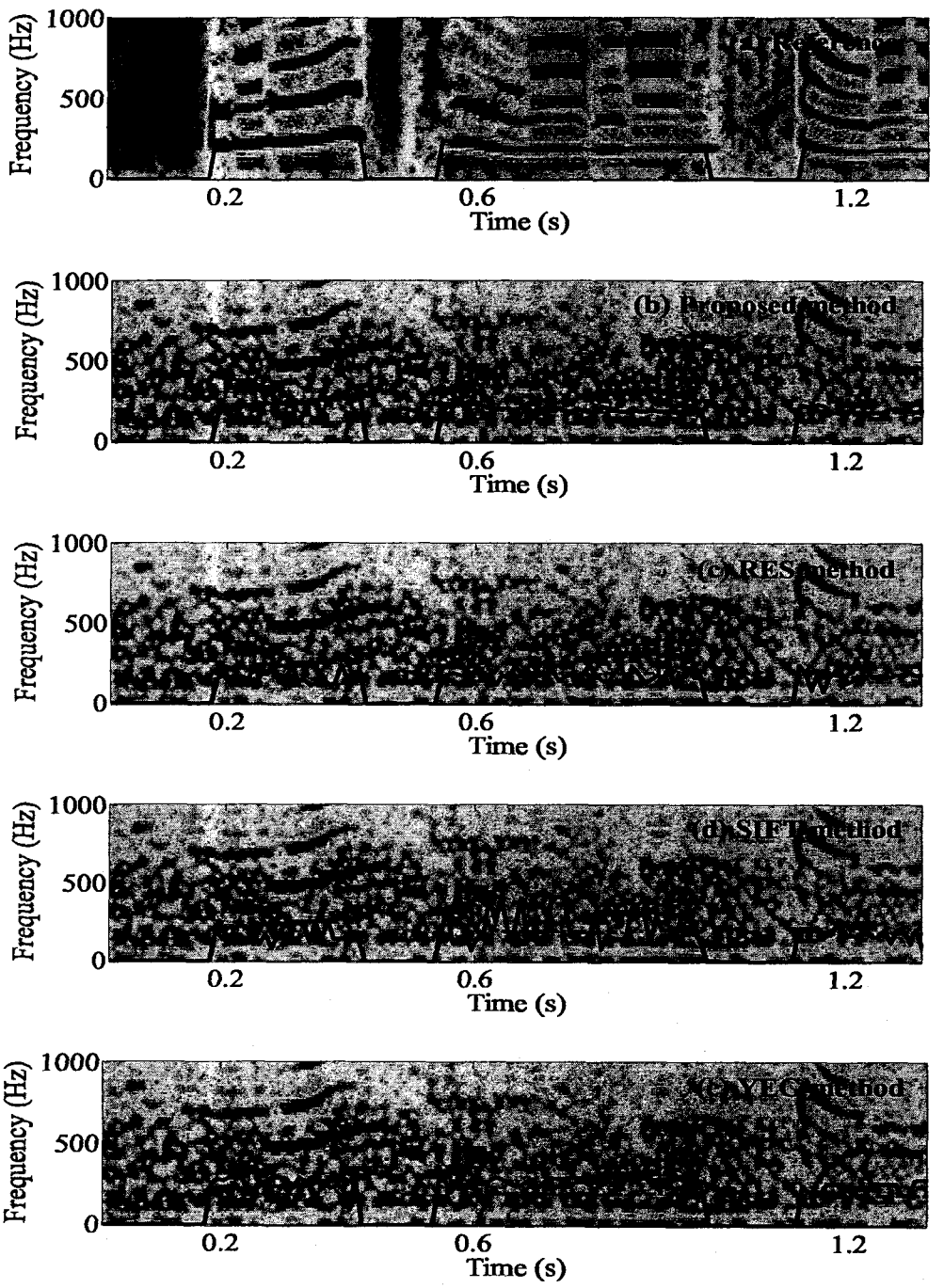


Figure 4.23: Pitch contours of different methods at SNR = −10 dB in Babble noise.

inverse-filtering the noisy speech, thus producing a reasonable RS. Since the presence of the unavoidable noise in the RS makes the generation of ELS difficult, an explicit noise-compensation scheme based on the identified VTS parameters and the estimated noise variance has been proposed to reduce the effect of noise on the ACF of the RS. Then, for the generation of the ELS, an SHE has been computed by employing the noise-compensated ACF of the RS. With a view to further overcome the adverse effect of noise on the ELS, a scheme for pitch estimation and tracking using a proposed symmetric normalized magnitude difference function of the ELS has been presented.

A comprehensive simulation study has been performed on speech signals of different female and male speakers, demonstrating the proposed time-frequency domain method, referred to as SNMDF-ELS, is sufficiently accurate and consistent in estimating the pitch at very low levels of SNR. The method has also been applied to pitch estimation of speech signals corrupted by a multi-talker babble noise. The simulation results have reinforced that the proposed method is superior to some of the state-of-the-art methods in dealing with pitch estimation of natural speech signals corrupted by white or real life babble noise and thus readily suitable for practical applications.

Chapter 5

Pitch Estimation Using the Pseudo-Cepstrum of an Excitation-Like Signal Obtained From Enhanced Speech

5.1 Introduction

It is well known that the peaks of the residual signal (RS) of speech is supposed to indicate the glottal closure (GC) instants [76], [77] and determination of those peaks gives the pitch periods [74]. Hence, pitch estimation using the GC instants in the RS has been investigated by some researchers [23] and [47]. These methods use ACF based linear prediction (LP) analysis [78], [79] for RS production and their performance solely depends on the accuracy of the vocal-tract system (VTS) parameter identification. Due to the inherent limitation of the ACF based LP analysis, the extracted RS exhibits bipolar fluctuations around the GC instants even in clean speech. Since the task of VTS parameter identification becomes very difficult in the presence of noise, due to inaccurate estimates of the VTS parameters, the RS of noisy speech shows indistinguishable peaks at the GC instants [82]. Even an attempt of computing the ACF from the RS or the processed RS cannot prevent the significant degradation

of the pitch estimation performance in a noisy environment [47]. Although the VTS parameter identification method based on Homomorphic deconvolution overcomes the overlapping problem of the conventional LP analysis and is capable of handling noisy speech as presented in the previous chapter, inverse filtering the noisy speech introduces an error term. To minimize its effect, a noise compensation scheme was proposed there in the correlation domain and the bipolar fluctuations at the GC instants were overcome by employing a noise-compensated ACF of the RS to generate a squared Hilbert Envelope (SHE) as an excitation-like signal (ELS). Furthermore, at a very low SNR, in order to overcome the undesirable effect of noise on the ELS, a new symmetric normalized magnitude difference function (SNMDF) of the ELS was proposed and used for pitch estimation.

In order to generate a noise-robust ELS for pitch estimation at a very low SNR, one possible solution is to introduce a noise-reduction block prior to pitch estimation as exists in automatic speech recognition system (ASR) for hands free mobile communication and multiparty meetings [97], [98]. In such applications, the recognition rate of ASR is always influenced by noise because of the training and recognition model mismatch in noisy environments. But ASR must remain usable under a large variety of noisy conditions. Besides the quest for robust features for ASR, one main line of research is aimed at handling noisy environments by employing prior speech enhancement. In particular, in many applications, such as car speech technologies and hearing aids [99], [100] there has been a framework of noise reduction followed by pitch estimation using a conventional method like the cepstrum [21]. The reason behind introducing such a noise reduction scheme prior to pitch estimation is twofold. First, a noise reduction scheme offers an advantage of handling a noise reduced speech frame, which would be much easier than handling directly a noise corrupted

frame. Secondly, the noise reduction scheme itself is capable of handling different types of noises, making the overall pitch estimation methodology more robust to different noisy scenarios.

The main objective of a noise reduction technique is to improve one or more perceptual aspects of speech, such as the quality or intelligibility while preserving necessary speech information [101]. For pitch estimation, suppose we only have a single channel available that provides the noisy observations, since usually a second channel is not available in most of the applications that use pitch information. As no reference noise is available, performing the task of noise reduction is very difficult and for such a scenario, the spectral subtraction scheme has become one of the most well-known techniques for noise reduction. The spectral subtraction based noise reduction methods have received a great deal of attention over the past several years because of their low complexity and relatively simple implementation [102]–[106]. This method is based on the assumption that the additive noise is uncorrelated with speech and does not change its amplitude drastically with time, which is rare in practical real-world noisy situations. In view of real life applications, possible cross-correlation between speech and noise cannot be ignored. Also, the noise power spectrum that is usually estimated for subtraction from the initial silence frames must be updated time to time adequately in order to track the time variation of the noise. In reference to the previous chapter, a noise-reduction scheme thus suitably designed will help identify the VTS parameters in the presence of noise with a sufficient accuracy for the RS generation [107]–[113]. A noise compensation on the RS will be no longer needed and the RS can be directly employed to generate an ELS for pitch estimation.

Motivated by the advantage of the existing framework, in this chapter we design an effective noise reduction scheme prior to pitch estimation. Encouraged by the

potential of the ELS over the speech itself in providing clues for pitch information as disclosed in the previous chapter, our target is to generate an ELS from the noise-reduced speech. It is noted that cepstral analysis has been an important tool in many fields, such as biomedical signal processing and speech processing [96], [114]–[117]. However, the conventional cepstrum, a nonlinearly transformed version of the spectrum of a signal computed from its phase blind autocorrelation, has been widely used as a feature in speech recognition [118], [119]–[121]. In this Chapter, instead of using a conventional cepstrum computed from the noisy speech or its LP residual [122], a phase information incorporated cepstrum obtained from the ELS of the enhanced speech is employed for accurate pitch estimation under severe noise.

In this chapter, a cepstral domain approach using an ELS obtained from the noise-reduced speech is presented for the estimation of pitch from speech signals in the presence of heavy noise [123]. The first task here is to develop a noise subtraction scheme in the frequency domain, which takes into account the possible cross-correlation between speech and noise and has an advantage of noise being updated from time to time and adjusted at each frame. The enhanced frame thus obtained is utilized to identify the VTS parameters via the homomorphic deconvolution technique. A RS is then produced by inverse-filtering the enhanced speech. The next task is to generate an SHE of the RS, which represents an ELS of the enhanced speech. In order to overcome the adverse effect of noise on the ELS under a severe noisy condition and the limitation of the conventional cepstrum in handling different types of noises, a time-frequency domain pseudo cepstrum of the ELS (PCELS) of the enhanced speech, incorporating information of both magnitude and phase spectra of the ELS, is proposed for pitch estimation on a frame-by-frame basis. Extensive simulations are carried out on the naturally spoken speech signals of the *Keele* data-

base in the presence of additive white or multi-talker babble noise available from the *Noisex92* database. Simulation results demonstrate quite a satisfactory pitch estimation performance by the proposed time-frequency domain method, referred to as PCELS-ES, for both female and male speakers even at an SNR of -10 dB. It is shown that the proposed method is also capable of performing well for multi-talker babble noise-corrupted speech signals, and thus is suitable for practical applications.

The rest of the chapter is organized as follows. In Section 5.2, a brief overview of the proposed pitch estimation method is presented through a block diagram. In Section 5.3, we propose a frequency domain noise reduction scheme. The way of generating the ELS from the enhanced speech is described in Section 5.4. In Section 5.5, a PCELS of the enhanced speech is introduced for pitch estimation in a heavy noisy condition. Section 5.6 demonstrates the pitch estimation performance of the proposed method through extensive computer simulations utilizing speech signals corrupted by both white and multi-talker babble noise. Finally, in Section 5.7, the attractive features of this investigation are summarized.

5.2 Brief Description of the Proposed Method

An overview of the proposed pitch estimation method is presented through a block diagram in Fig. 5.1. A pre-processed frame of the observed noisy speech $s_y(n)$ can be expressed in the time-domain as

$$y(n) = x(n) + v(n), \tag{5.1}$$

where $x(n)$ and $v(n)$ represent the pre-processed versions of clean speech $s_x(n)$ and additive noise $s_v(n)$, respectively. Similar to Chapter 4, here the pre-processing of the observed noisy speech $s_y(n)$ involves a windowing operation, pre-emphasis and

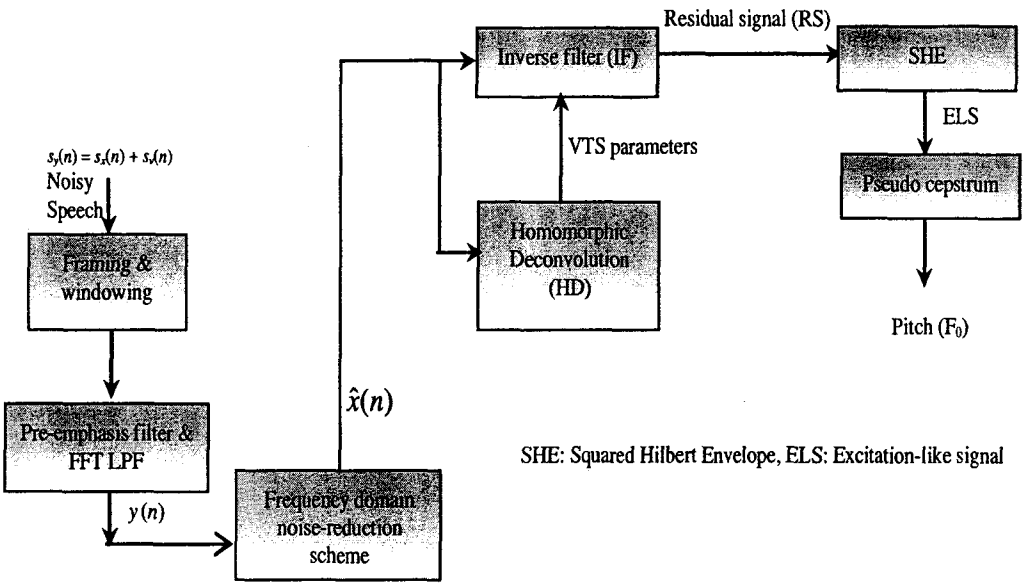


Figure 5.1: A block diagram representing the overview of the proposed pitch estimation method.

low-pass filtering (to exclude frequencies above 5 kHz) of a noisy speech frame. Unlike the previous chapter, we first attempt to reduce noise from the pre-processed noisy speech frame $y(n)$ to generate an ELS of the enhanced speech and then compute a pseudo-cepstrum of the ELS (PCELS) of the enhanced speech for pitch estimation.

Recall that, in the previous chapter, the RS obtained from $y(n)$ was first noise-compensated in the autocorrelation domain and then the noise-compensated ACF of the RS was employed to generate an SHE as the ELS of the voiced speech. Now, in this chapter, since the RS is obtained from the noise-reduced speech, it is found that the SHE of the RS of the enhanced speech without noise compensation in the autocorrelation domain, is accurate enough to be an ELS.

In the computation of conventional real cepstrum, the magnitude spectrum or

the power spectrum of the speech signal is usually used in order to avoid the phase unwrapping problem [90], [95]. Here, motivated by the role of the phase spectrum in speech perception, which is recently reported, we incorporate the information of both magnitude and phase spectra of the ELS while computing the PCELS of the enhanced speech.

5.3 A Frequency-Domain Noise Reduction Scheme

In the spectral subtraction based noise reduction methods, an estimate of the noise power spectrum is subtracted from the noisy speech power spectrum in order to obtain an estimate of the power spectrum of the clean speech. We propose a spectral subtraction based noise reduction scheme where, unlike conventional methods, the noise spectrum estimate is updated in every silence (or noise-only) period and moreover, the cross-correlation between speech and noise is taken into consideration. Our target is to obtain an estimate of $x(n)$ using the proposed power spectral subtraction based noise reduction scheme applied to $y(n)$. From (5.1), the discrete Fourier transform (DFT) of $y(n)$ can be written as

$$Y(k) = X(k) + V(k), \tag{5.2}$$

where $X(k)$ and $V(k)$ are the DFTs of $x(n)$ and $v(n)$, respectively. For an input vector $\{y(1), y(2), \ldots, y(N)\}$, its DFT can be computed as

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j2\pi nk/N}, \quad 0 \leq k \leq N-1. \tag{5.3}$$

The instantaneous power spectrum of $y(n)$ can be estimated as

$$|Y(k)|^2 = |X(k)|^2 + |V(k)|^2 + C(k) + C^*(k), \tag{5.4}$$

where $C(k) = X(k)V^*(k)$. Since in a noisy environment, we do not have access to $x(n)$, we would like to obtain an estimate of $|X(k)|^2$ from $|Y(k)|^2$. The cross-terms $C(k)$ and $C^*(k)$ in (5.4) reduce to zero if $v(n)$ and $x(n)$ are uncorrelated and one of the two has zero mean. However, in the presence of correlated real world noise, these cross terms can no longer be neglected. Considering the effect of cross-terms, an FFT based power spectral subtraction scheme with cross-correlation compensation is derived from (5.4) as

$$|\hat{X}(k)|^2 = \begin{cases} H(k), & \text{if } H(k) > 0 \\ \beta_s |\hat{V}(k)|^2, & \text{otherwise} \end{cases} \tag{5.5}$$

where

$$H(k) = |Y(k)|^2 - \alpha|\hat{V}(k)|^2 - \varsigma|Y(k)||\hat{V}(k)|. \tag{5.6}$$

In (5.5), $|\hat{X}(k)|^2$ represents an estimate of the short-time FFT power spectra of $x(n)$ and β_s refers to the spectral floor parameter introduced to prevent the negative value of $|\hat{X}(k)|^2$. In (5.6), α symbolizes the over-subtraction factor used to prevent the over-estimation of the noise power spectrum, and the last term is introduced to consider the effect of instantaneous cross-correlation between $Y(k)$ and $\hat{V}(k)$, where ς is the cross-correlation compensation factor.

In the conventional spectral subtraction scheme, an estimate $|\hat{V}(k)|^2$ of the noise power spectrum $|V(k)|^2$ is obtained from the beginning silence frames. In the proposed spectral subtraction based noise reduction scheme, we update the thus estimated noise power spectrum during each silence period as follows

$$|\hat{V}^t(k)|^2 = \begin{cases} |Y^t(k)|^2, t = 1 \\ \nu_n |\hat{V}^{t-1}(k)|^2 + (1 - \nu_n)|Y^t(k)|^2, \end{cases} \tag{5.7}$$

where t is the frame index, ν_n is the forgetting factor, $|\hat{V}^t(k)|^2$ and $|Y^t(k)|^2$, respectively, represent the estimated noise power spectrum and the power spectrum of the

noisy speech in the t-th frame. For a frame t after a silence period, we rely on the use of a preliminary noise power spectrum estimated as $|\hat{V}^t(k)|^2 = |\hat{V}^{t_I}(k)|^2$, where t_I refers to the index of the immediately last silence frame before the beginning of a speech frame. Considering that this estimate of the noise power spectrum is updated only during a silence period while it may change drastically with time, it is insufficient to use a constant value of the over-subtraction factor α to compensate for the errors induced in the noise power spectrum to be subtracted in each frame. In order to track the time variation of the noise, α should be adjusted in each frame t after a silence period. According to the spectral characteristics of the human speech, the low-frequency band typically from 0 to 50 Hz contains no speech information. Thus, for noisy speech, the initial low frequency band, say $\triangle = [0, 50]$ Hz contains only noise. In view of this fact, in order to change the value of α_t for the t-th frame after a silence period, we propose to use the ratio between the powers of $|Y^t(k)|$ and $|\hat{V}^{t_I}(k)|$ in the low frequency band $\triangle$ as

$$\alpha_t = \frac{\sum_{k\in\triangle} |Y^t(k)|^2}{\sum_{k\in\triangle} |\hat{V}^{t_I}(k)|^2}, \quad \triangle = [0,\ 50]\ Hz \tag{5.8}$$

where $|\hat{V}^{t_I}(k)|^2$ represents the estimated noise power spectrum in the immediately last silence frame t_I before the beginning of a speech frame. In the low-frequency band $\triangle$ of the t-the frame, the time variation of the noisy speech power spectrum $|Y^t(k)|$ is equivalent to the noise power spectrum of that frame. Thus, the use of α_t defined in (5.8) clearly serves as a relative weighting factor with respect to the estimated preliminary noise power spectrum,$|\hat{V}^t(k)|^2 = |\hat{V}^{t_I}(k)|^2$, leading to a reasonable tracking for the time-variation of the noise $v(n)$. The cross-correlation compensation factor

$0 \leq \varsigma \leq 1$ in (5.6) can be expressed as

$$\varsigma = \left| \frac{\sigma_{yv}}{\sigma_y \sigma_v} \right| \tag{5.9}$$

with

$$\sigma_{yv} = \frac{1}{N/2} \sum_k |Y(k) - \partial_y||\hat{V}(k) - \partial_v| \tag{5.10}$$

$$\sigma_y^2 = \frac{1}{N/2} \sum_k \{|Y(k)| - \partial_y\}^2 \tag{5.11}$$

$$\sigma_v^2 = \frac{1}{N/2} \sum_k \{|\hat{V}_w(k)| - \partial_v\}^2 \tag{5.12}$$

$$\partial_y = \frac{1}{N/2} \sum_k |Y(k)|, \quad \partial_v = \frac{1}{N/2} \sum_k |\hat{V}(k)| \tag{5.13}$$

Here σ represents the standard deviation and for a zero mean uncorrelated white noise, ς reduces to zero (as $\hat{V}(k) = \partial_v$). Once the subtraction is performed in the power spectral domain based on (5.5)-(5.13), an enhanced speech frame is obtained by using the estimated magnitude spectrum along with the phase of the noisy speech as

$$\hat{x}(n) = IFFT\{|\hat{X}(k)|e^{j \arg(Y(k))}\}, \tag{5.14}$$

where the inverse Fourier transform is performed via inverse FFT (IFFT). The clean speech estimate $\hat{x}(n)$ thus obtained can be assumed to preserve the desired pitch information. Fig. 5.2 plots clean speech, its noisy version at SNR $= -10$ dB as well as the corresponding noise-reduced version. It is seen that the proposed frequency domain noise-reduction scheme is capable of yielding an enhanced speech that is expected to help pitch estimation even in a severe noisy condition.

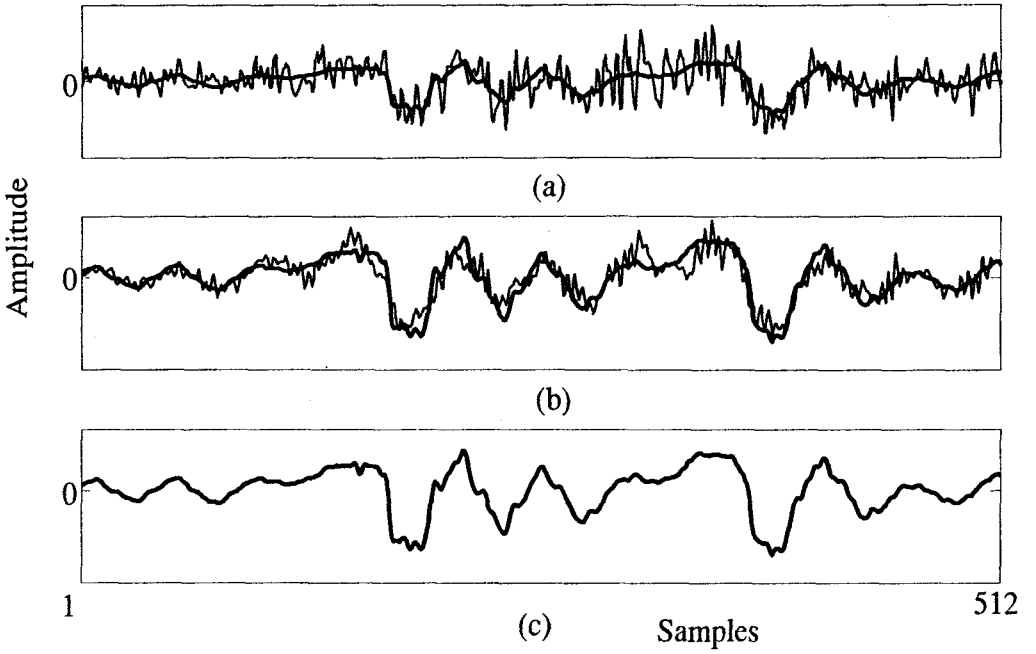


Figure 5.2: (a) Noisy signal at SNR = −10 dBdB, (b) Noise-reduced speech at SNR = −10 dB, and (c) clean speech.

5.4 Generation of an Excitation-like Signal from the Enhanced Speech

During a voiced sound, instants of GC with significant excitation occur at an interval of the pitch period. Thus, pitch information is well preserved in the excitation signal of the vocal-tract system. Instead of directly using the noise-reduced speech signal $\hat{x}(n)$ obtained in the previous section, we want to first extract an excitation-like signal of $\hat{x}(n)$ and then use it for the purpose of pitch estimation. In view of this, $\hat{x}(n)$ is passed through an inverse vocal tract system filter yielding an estimate of the RS as given by

$$\hat{\Re}(n) = \hat{x}(n) + \sum_{k=1}^{p} \hat{a}_k \hat{x}(n-k). \tag{5.15}$$

Here the vocal tract system parameters $\hat{a}_k$ are identified first from the ACF of $\hat{x}(n)$, which can be written as

$$\phi_{\hat{x}}(m) = \frac{1}{N}\sum_{n=0}^{N-1-|m|}\hat{x}(n)\hat{x}(n+|m|), \quad m=0,\pm1,\ldots,\pm M, M<N \tag{5.16}$$

$$= \phi_x(m)+\phi_w(m), \tag{5.17}$$

where $\phi_x(m)$ represents the ACF of clean speech $x(n)$ in a frame and $\phi_w(m)$ refers to a term arising from the remaining noise in $\hat{x}(n)$. Unlike $\phi_y(m)$ in (4.12) in the previous chapter, $\phi_{\hat{x}}(m)$ does not contain the ACF $\phi_v(m)$ of $v_(n)$ and the sum of cross-correlation terms $\phi_c(m)$ but carries a noise term $\phi_w(m)$ which, under a heavy noisy condition, cannot be neglected even after noise reduction. It is already known that the ACF $\phi_x(m)$ in (5.17) can be considered as a convolution of $\phi_g(m)$, the ACF of the vocal-tract impulse response $g(n)$, and $\phi_e(m)$, the ACF of the frame $e(n)$ of the impulse-train excitation. Thus, a cepstral domain homomorphic deconvolution on $\phi_{\hat{x}}(m)$ is performed to remove the effect of excitation and some portion of the remaining noise. To this end, in order to compute the cepstrum of $\phi_{\hat{x}}(m)$, we apply a logarithm operation on the discrete Fourier transform of $\phi_{\hat{x}}(m)$ as given by (5.17) and obtain

$$\log(\Phi_{\hat{x}}(k)) = \log\left[\Phi_{\hat{x}}(k)\left((1+\frac{\Phi_w(k)}{\Phi_{\hat{x}}(k)}\right)\right] = \log[\Phi_{\hat{x}}(k)\Phi_u(k)] = \log[\Phi_{\hat{x}}(k)]+\log[\Phi_u(k)]; \tag{5.18}$$

here $\Phi_{\hat{x}}(k)$, $\Phi_x(k)$ and $\Phi_w(k)$ are the DFTs of $\phi_{\hat{x}}(m)$, $\phi_x(m)$ and $\phi_w(m)$, respectively. By taking the inverse DFT of $\log(\Phi_{\hat{x}}(k))$, the real cepstrum $c_{\phi_{\hat{x}}}(n)$ of $\phi_{\hat{x}}(m)$ can be computed as

$$c_{\phi_{\hat{x}}}(n) = c_{\phi_g}(n)+c_{\phi_e}(n)+c_{\phi_u}(n), \quad 0\le n\le N-1. \tag{5.19}$$

In (5.19), $c_{\phi_g}(n)$ and $c_{\phi_e}(n)$ are, respectively, the real cepstra of $\phi_g(m)$ and $\phi_e(m)$, and $c_{\phi_u}(n)$ arises because of the noise. The term $c_{\phi_u}(n)$ determines as to how the noise affects $c_{\phi_{\hat{x}}}(n)$ and it vanishes altogether in the absence of noise. A low-time liftering operation in the quefrency domain is performed on $c_{\phi_{\hat{x}}}(n)$ to extract a better estimate of $c_{\phi_g}(n)$. The liftering operation on $c_{\phi_{\hat{x}}}(n)$ not only removes the effect of $c_{\phi_e}(n)$, it also further reduces the effect of $c_{\phi_u}(n)$ from the high quefrency portion of $c_{\phi_{\hat{x}}}(n)$. By using the inverse cepstrum operation on the liftered cepstrum $c_{\phi_{\hat{x}}}(n) \approx c_{\phi_g}(n) + \hat{c}_{\phi_u}(n)$, an estimate $\hat{\phi}_g(m)$ of $\phi_g(m)$ is obtained. Such an estimate $\hat{\phi}_g(m)$ is used in the following autoregressive relation

$$\hat{\phi}_g(m) = -\sum_{k=1}^{p} \hat{a}_k \hat{\phi}_g(m-k), \quad 0 < m \leq p. \tag{5.20}$$

Since, in the presence of noise, lower lags of $\hat{\phi}_g(m)$ generally become more corrupted than that of the higher lags, a few lower lags of $\hat{\phi}_g(m)$ are avoided in the computation of VTS parameters $\hat{a}_k$. Using $\hat{\phi}_g(m)$ for $m = p+1 \ldots p+S$, (5.20) yields a similar set of linear equations as in (4.16) of the previous Chapter. The VTS parameters $\hat{a}_k$ are then identified from the least-squares solution of the set of equations as mentioned above.

With the proposed frequency domain noise reduction scheme, it is expected that the effect of $v(n)$ has been significantly reduced. However, even with an accurate estimate of the vocal tract $\hat{a}_k$ parameters, because of the remaining noise in $\hat{x}(n)$, the estimated residual $\hat{\Re}(n)$ in (5.15) contains an error term $\Re_w(n)$, namely

$$\hat{\Re}(n) = \Re(n) + \Re_w(n), \tag{5.21}$$

where $\Re(n)$ is the RS obtained in the noise-free environment from $x(n)$. Clearly, if the VTS parameters can be accurately identified from $x(n)$, $\Re(n)$ is expected to closely resemble the excitation signal that produces $x(n)$.

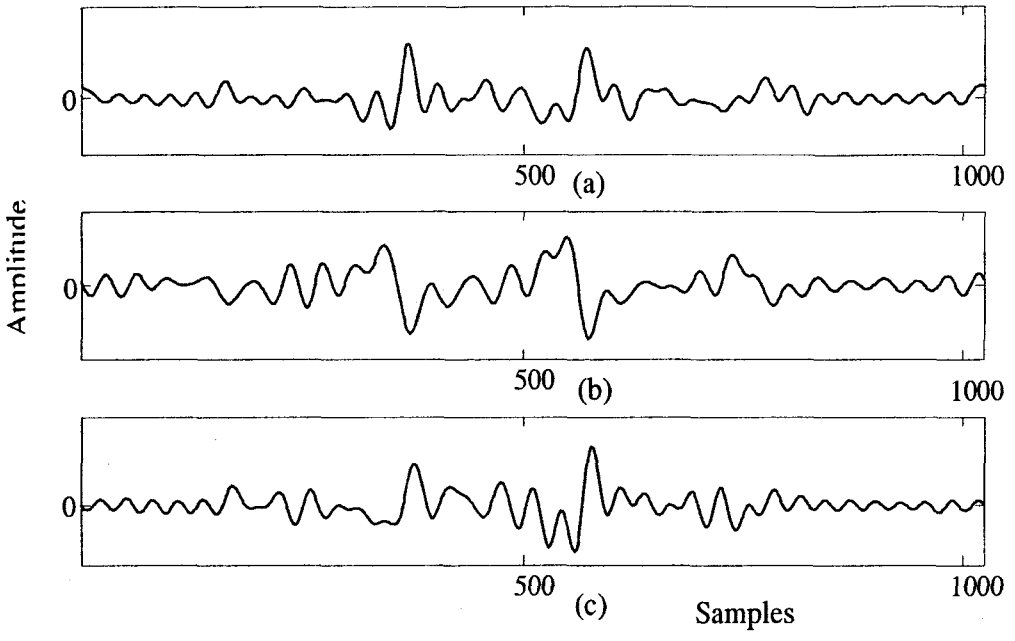


Figure 5.3: Residual signal: (a) noise-free environment, (b) without prior noise reduction in a noisy environment (SNR $= -10$ dB) and (c) with prior noise reduction at SNR $= -10$ dB.

In Fig. 5.3, three residual signals obtained from clean speech, noisy speech and the enhanced one are plotted for a voiced frame. The effect of noise reduction is clearly observable by comparing the residual signals obtained from the noisy speech and that of the enhanced one. From this figure, a bipolar fluctuation of the RS around the GC instants can also be observed, which makes the task of pitch estimation difficult if the RS is directly used.

It is found that the Hilbert Transform (HT) of the RS $\Re(n)$ of clean speech exhibits an unambiguous peak at the GC instant [74]. The RS $\hat{\Re}(n)$ in (5.21) obtained from the enhanced speech contains reasonably prominent peaks at the GC instants, as plotted in Fig. 5.3(c). Hence, the HT $\hat{\Re}_h(n)$ of the RS $\hat{\Re}(n)$ is expected to retain similar peaks at the GC instants. If $\hat{\Re}(k)$ is the discrete Fourier transform of $\hat{\Re}(n)$, then

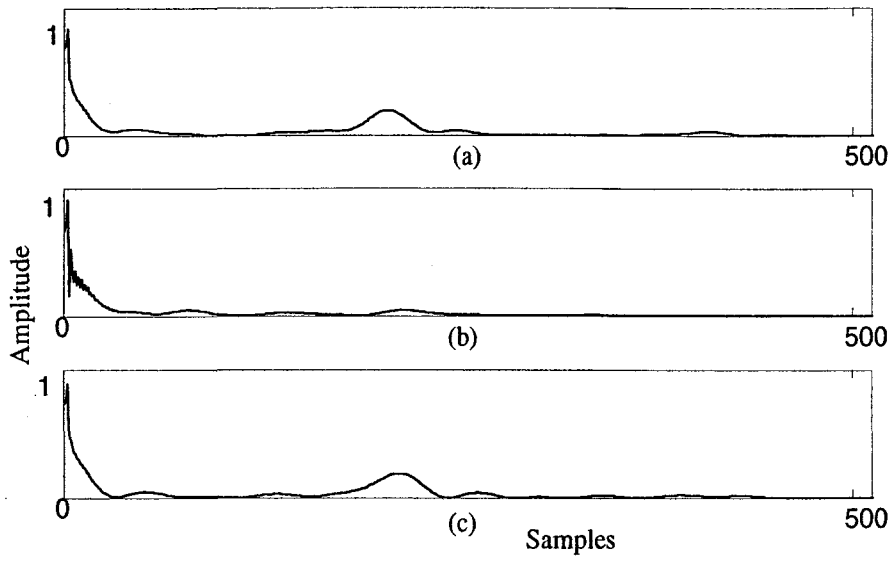


Figure 5.4: Square Hilbert Envelope (ELS): (a) noise-free environment, (b) without prior noise reduction in a noisy environment (SNR = -10 dB) and (c) with prior noise reduction at SNR = -10 dB.

the HT $\hat{\Re}_h(n)$ can be defined following (4.25) and (4.26) from the previous Chapter. Recall that the Hilbert transform $\hat{\Re}_h(n)$ of a real signal $\hat{\Re}(n)$ is also real, and $\hat{\Re}(n)$ and $\hat{\Re}_h(n)$ constitute, respectively, the real and imaginary parts of an analytic signal. Motivated by the features of the $\hat{\Re}(n)$ and the corresponding $\hat{\Re}_h(n)$ as mentioned above, we propose an SHE of the RS $\hat{\Re}(n)$ as an ELS of the enhanced speech, which is defined as

$$E_{\hat{\Re}}(n) = \hat{\Re}^2(n) + \hat{\Re}_h^2(n). \tag{5.22}$$

Even though $\hat{\Re}(n)$ and $\hat{\Re}_h(n)$ have positive and negative samples, according to (5.22), the SHE $E_{\hat{\Re}}(n)$ of $\hat{\Re}(n)$ is a positive function. Fig. 5.4 shows the SHEs computed in the clean environment as well as in a noisy environment with and without prior noise reduction. It is seen from Fig. 5.4 that unlike $\Re(n)$, the SHE, i.e., the ELS generated from the clean speech, exhibits the desired unipolar characteristics at the

GC instants. It can also be observed that the ELS $E_{\hat{\Re}}(n)$ obtained from the enhanced speech closely matches the ELS in clean speech and demonstrates a similar unipolar nature at the GC instants. Hence, the major peaks in $E_{\hat{\Re}}(n)$ occurring at the GC instants can be exploited to determine the pitch period at a moderately low SNR.

5.5 Pseudo-Cepstrum of the ELS of the Enhanced Speech

In a severe noisy condition, because of the pronounced effect of the error term $\Re_w(n)$ on the RS $\hat{\Re}(n)$ in (5.21), the peaks of the excitation-like signal $E_{\hat{\Re}}(n)$ generated from $\hat{\Re}(n)$ using (5.22) may provide only an approximation for the locations of the peaks at the GC events. Thus, the separation of the successive peaks of $E_{\hat{\Re}}(n)$ may not give information about the true pitch period. Hence, with a view to overcome the undesirable effect of noise on the ELS $E_{\hat{\Re}}(n)$, we propose a time-frequency domain pseudo-cepstrum of $E_{\hat{\Re}}(n)$ to determine an accurate pitch estimate under a very low SNR.

In speech analysis, it is commonly believed that the human auditory system is phase-deaf, i.e., it ignores phase spectrum and uses only magnitude spectrum for speech perception. Recently, it has been shown that the phase spectrum is also useful in human speech perception [124], [125], [126]. This inspires us to incorporate some kind of phase information derived from the phase spectrum of a signal. To this end, our signal of interest is the ELS $E_{\hat{\Re}}(n)$ and the Fourier transform of $E_{\hat{\Re}}(n)$ can be written as

$$F[E_{\hat{\Re}}(n)] = E(\omega) = |E(\omega)|e^{j\Xi(\omega)}. \tag{5.23}$$

Here, $\Xi(\omega)$ corresponds to the phase spectrum of $E_{\hat{\Re}}(n)$. A commonly used function dealing with the phase spectrum is the group delay function [127], [128], [129]. The

group delay function of $E_{\hat{\Re}}(n)$ is given by

$$\tau_d(\omega) = -\frac{d\Xi(\omega)}{d\omega}. \tag{5.24}$$

It can be simplified as follows [66]

$$\begin{aligned} \tau_d(\omega) &= -\mathrm{Im}\frac{d(\log E(\omega))}{d\omega} \\ &= \frac{E_r(\omega)G_r(\omega) + E_i(\omega)G_i(\omega)}{|E(\omega)|^2}, \end{aligned} \tag{5.25}$$

where $G(\omega)$ is the Fourier Transform of $nE_{\hat{\Re}}(n)$, and the subscripts denote the real and imaginary parts. We define the product spectrum $P(\omega) = |E(\omega)|^2\tau_d(\omega)$, which from (5.25) can be written as

$$P(\omega) = E_r(\omega)G_r(\omega) + E_i(\omega)G_i(\omega). \tag{5.26}$$

Although the product spectrum $P(\omega)$ is determined by both the magnitude spectrum and the phase spectrum, it does not involve phase computation and is a real function. Hence, we are motivated to propose a cepstrum of the product spectrum of the excitation-like signal $E_{\hat{\Re}}(n)$ as defined by

$$c_E(n) = F^{-1}[\log(P(\omega))]. \tag{5.27}$$

We refer to $c_E(n)$ as the pseudo-cepstrum of ELS (PCELS) of the enhanced speech. In Fig. 5.5 the PCELS of the enhanced speech is plotted along with the PCELS corresponding to the noisy and noise free case. It can be observed from the figure that the proposed pseudo-cepstrum of the ELS of the enhanced speech emphasizes the pitch-peak compared to that of the PCELS of noisy speech $y(n)$, thus justifying the effectiveness of $c_E(n)$ in extracting the pitch period.

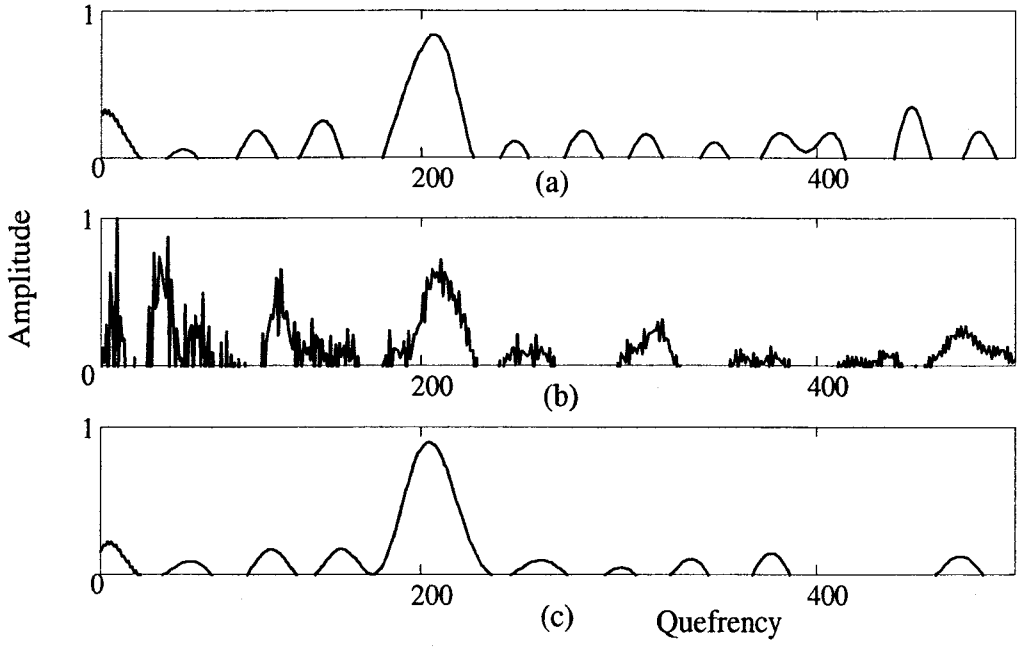


Figure 5.5: Pseudo-cepstrum of the ELS: (a) noise-free environment, (b) without prior noise reduction in noisy environment (SNR = -10 dB) and (c) with prior noise reduction at SNR = -10 dB.

Finally, by searching for the location of the global maximum from $c_E(n)$ in the possible range of pitch period $[n_min : n_max]$, an estimate of T_0 can be obtained as

$$\widetilde{T_0} = \underset{n}{arg\ max}[c_E(n)], \tag{5.28}$$

which gives the corresponding estimate of pitch frequency in Hz as

$$\widetilde{F_0} = \frac{F_s}{\widetilde{T_0}}. \tag{5.29}$$

The complete method for pitch estimation whose development started in the Section 5.2 will henceforth be referred to as the PCELS-ES method.

5.6 Simulation Results

Extensive simulations are performed for the estimation of pitch under noisy conditions; results along with some comparative analysis are investigated in this section for the proposed PCELS-ES method.

5.6.1 Simulation Conditions

(a) Database and other details: For the purpose of simulations, we employ naturally spoken speech signals obtained from the *Keele* database [63], [64]. A speech sequence in the *Keele* database is divided into overlapped analysis frames, each of size $N = 25.6$ ms at a frame rate of 100 Hz (i.e., a frame shift of 10 ms). In order to test the method in a noisy condition, white and multi-talker babble noises from the *Noisex92* database [65] are considered to corrupt the original clean speech $s_x(n)$ according to (2.26). The noisy speech $s_y(n)$ is segmented into N-sample overlapped frames by the application of a normalized hamming window function and we use the same values for the parameters, such as frame rate and basic frame (or window) size as specified in the *Keele* database. As in Chapter 4, in order to reduce the natural spectral tilt of the windowed speech, a high-pass pre-emphasis filter with input-output relation given by (4.43) is employed. In addition to signal pre-emphasis, a low-pass filtering by using N-point FFT and IFFT operations is also performed to exclude the high frequencies above 5 kHz, which are not of our interest in the generation of the RS as well as ELS.

In the frequency domain noise reduction scheme proposed in Section 5.3, the initial estimate of noise statistics is made from averaging over a few beginning frames of silence [104], [130] and the estimated noise power spectrum is updated in every

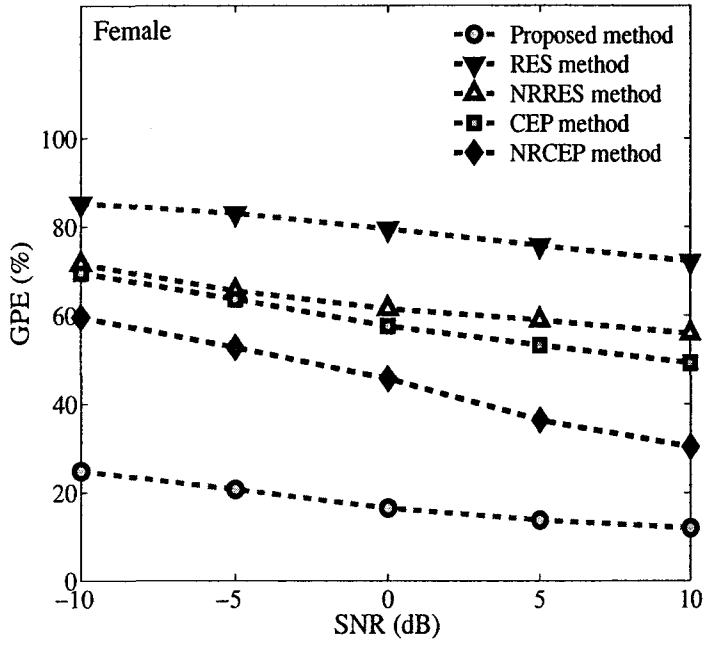


Figure 5.6: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in white noise.

silence period using (5.7) with a forgetting factor $\upsilon_n = 0.9$. We incorporated a statistical model-based voice activity detector (VAD) [131] to detect silence frames for the estimation and update of the noise. The parameter β_s was set to 0.002 in (5.5). As the homomorphic deconvolution principle, the standard cepstrum analysis [89] is employed on the ACF $\phi_{\hat{x}}(m)$ of the enhanced speech $\hat{x}(n)$. In the liftering operation performed on the real cepstrum $c_{\phi_{\hat{x}}}(n)$ of $\phi_{\hat{x}}(m)$, 4.5 milliseconds and 2.5 milliseconds cepstral windows are chosen for male and female speakers, respectively. A 12-th order ($p = 12$) autoregressive (AR) model is assumed to characterize the VTS filter. In the computation of the p number of VTS $\hat{a}_k$ parameters, the number of equations used is governed by S, and we have chosen $S = 5p$.

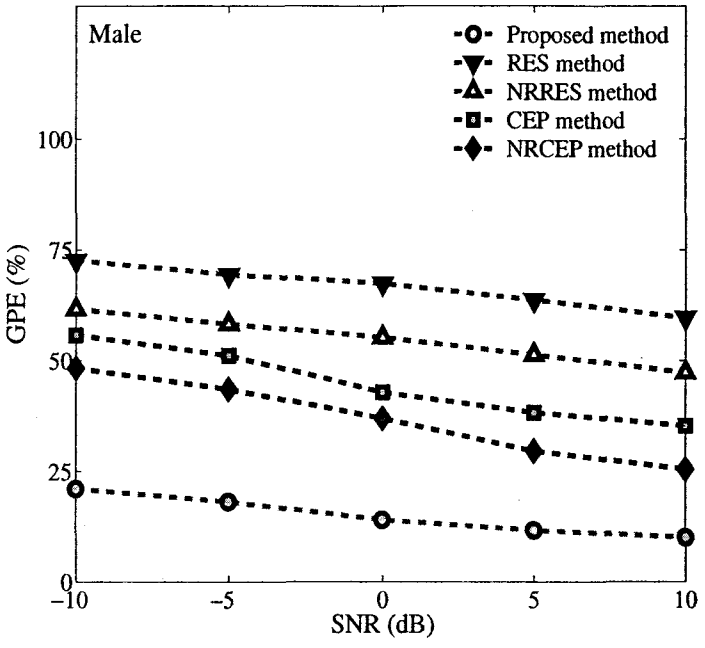


Figure 5.7: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in white noise.

(b) Metrics and Comparison Methods Used for Performance Evaluation: The performance metrics 1) the percentage gross pitch error (PGPE); 2) the mean of fine pitch error (m_{FPE}); 3) the standard deviation of fine pitch error (σ_{FPE}); and 4) the root-mean-square-error (RMSE), as defined in Chapter 2, are considered in our simulation study. For the purpose of comparison, we use the RES method [41] and the conventional cepstrum (CEP) method [21]. For fair comparison of the proposed PCELS-ES method to the methods mentioned above, we also implemented each comparison method preceded by the frequency domain noise reduction scheme proposed in Section 5.3 and refer to them as the NRRES method and the NRCEP method, respectively. While implementing the methods independently, we used the default parameters specified by the authors. The performance of the proposed method as

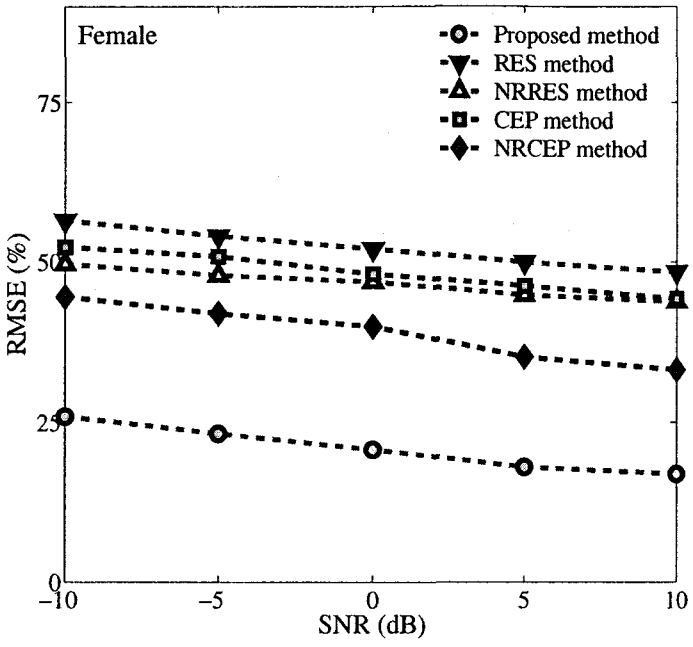


Figure 5.8: RMSE (%) as a function of SNR for female speaker group in white noise.

well as the other methods is evaluated at different levels of SNR in terms of the pitch estimates for the voiced frames based on the voiced/unvoiced labels provided by the *Keele* database.

5.6.2 Simulation Results and Comparisons

(a) Results on white-noise corrupted speech: We first investigate the pitch estimation performance of the RES, NRRES, CEP, NRCEP and the proposed PCELS-ES methods for the speech signals of the female (5) and male (5) speakers of the database in the presence of white noise. Figs. 5.6 and 5.7 show the PGPE values as a function of SNR obtained from all the five methods for the female and male speaker groups, respectively, where the SNR varies from a very low value of -10 dB

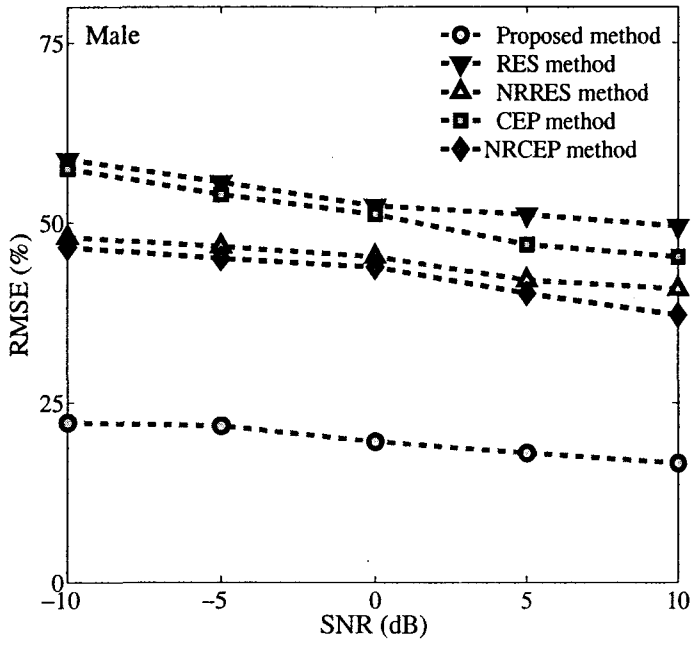


Figure 5.9: RMSE (%) as a function of SNR for male speaker group in white noise.

to a high value of 10 dB. It is seen from these figures that, although the RES method with a prior noise reduction (NRRES) shows an improvement in the performance over the RES method, the estimation results of the RES and the NRRES methods are not acceptable for the whole range of SNR. Although the CEP and the NRCEP methods give lower values of the PGPE than that given by the RES and the NRRES methods at an SNR=10 dB, their PGPE values are much higher even at this high SNR compared to that achieved by the proposed method. The proposed PCELS-ES method exhibits a superior performance with respect to the PGPE metric at all the levels of SNR considered. It is important to note that the proposed method provides an acceptable performance, when all the methods including the NRRES and NRCEP methods with prior noise reduction are unable to estimate the pitch at a very low

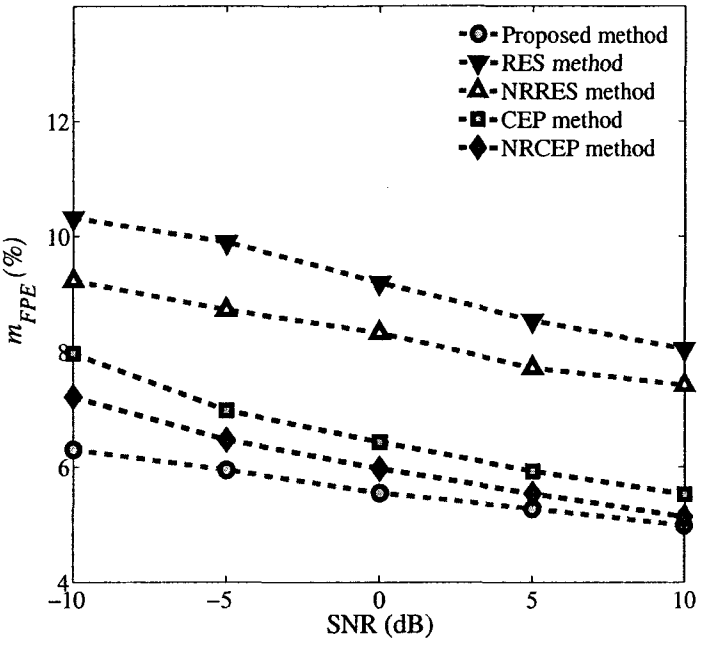


Figure 5.10: m_{FPE} (%) as a function of SNR for all female and male speakers in white noise.

SNR of -10 dB.

Figs. 5.8 and 5.9 present the values of RMSE as a function of SNR obtained by using the five methods for the same female and male speaker groups as above. It is demonstrated through these figures that the RES method with prior noise reduction (NRRES) outperforms the RES and CEP methods but it completely fails to estimate pitch not only at a very low SNR but also at a high SNR of 10 dB. Although the NRCEP method shows an improvement in performance over the RES, CEP and NRRES methods at high SNR, the RMSE values obtained from the proposed method are much lower at such an SNR. Also, the proposed method provides quite a high estimation accuracy for low levels (as low as -10 dB) of SNR, whereas the estimation results of the other methods are not satisfactory at lower levels of SNR.

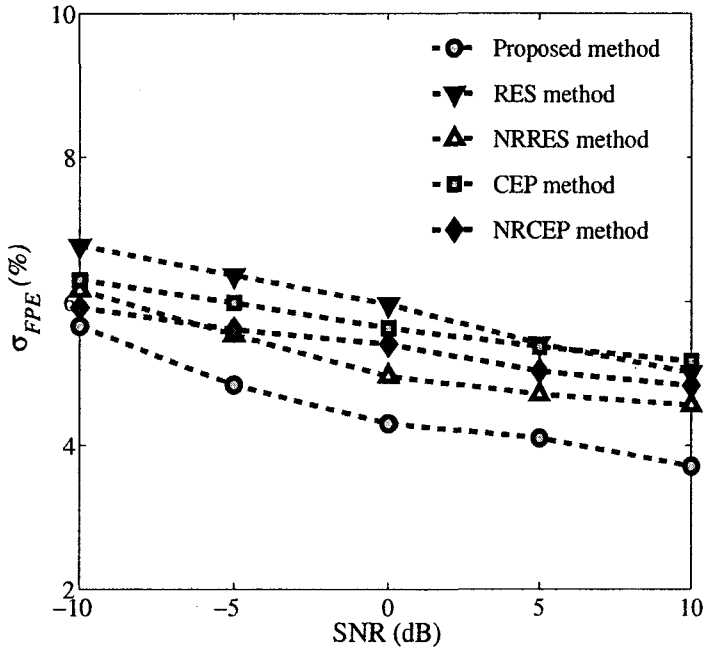


Figure 5.11: σ_{FPE} (%) as a function of SNR for all female and male speakers in white noise .

Fig. 5.10 and Fig. 5.11 depict, respectively, the mean m_{FPE} and standard deviation σ_{FPE} obtained by using the five methods for the set of 10 mixed (5 female plus 5 male) speakers of the database. It is seen that the overall mean of the fine-pitch errors resulting from the proposed PCELS-ES method is lower in comparison to that obtained by the other methods in the entire range of the SNR levels considered. The overall standard deviation values of the fine-pitch errors obtained from the proposed method are comparable to the NRRES, CEP and NRCEP methods at all levels of SNR.

Smaller values of the PGPE and RMSE achieved by using the proposed method under a wide range of SNR levels, along with lower and competitive values of the overall mean and standard deviation of the fine-pitch errors, indicate its superior ability

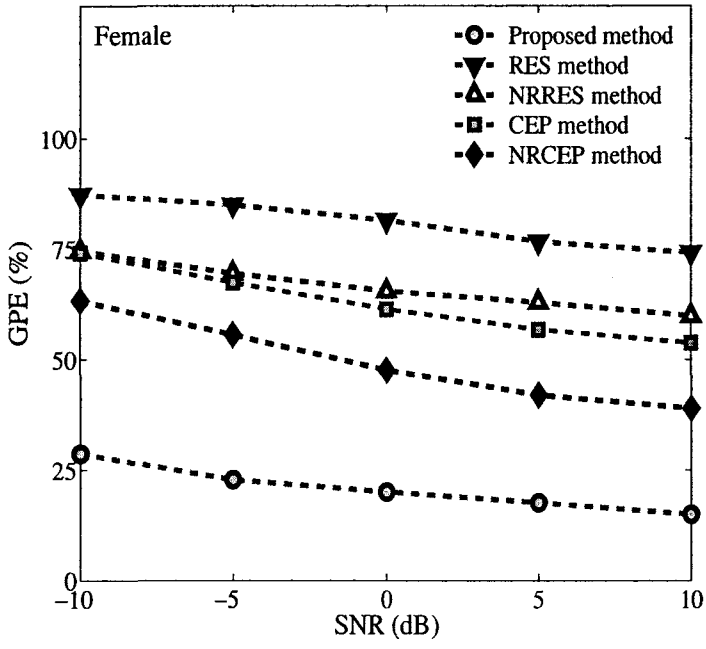


Figure 5.12: Percentage GPE [GPE (%)] as a function of SNR for female speaker group in babble noise.

for accurate pitch estimation and its high degree of robustness as well as consistency.

(b) Results on Multi-Talker Babble-Noise Corrupted Speech:

We now study the robustness of the proposed PCELS-ES and the other two methods in the presence of babble noise, which is a non-Gaussian non-stationary colored noise. The pitch estimation results obtained from the babble-noise-corrupted speech in terms of the PGPE, RMSE, mean m_{FPE} and standard deviation σ_{FPE} for each of the methods are portrayed in Fig. 5.12 through Fig. 5.17. It is expected that that the performance of all the five methods would degrade in the presence of babble noise compared to that in the white noise due to presence of the harmonic components coming from the multiple background competing speakers in the babble noise itself.

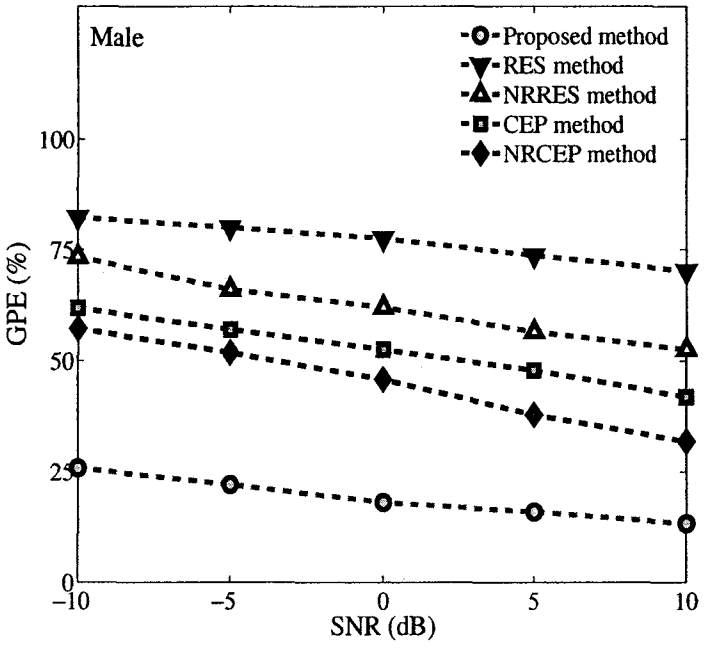


Figure 5.13: Percentage GPE [GPE (%)] as a function of SNR for male speaker group in babble noise.

However, it is seen from these figures that for the same male and female speaker groups as the ones considered for the white-noise corrupted speech, the proposed PCELS-ES method retains its superiority with respect to all the four performance metrics at all the levels of SNRs.

Figs. 5.12 and 5.13 provide plots for the PGPE values as a function of SNR obtained from the three methods for the female and male speaker groups. It is seen that the proposed method maintains satisfactory performance at high SNRs, such as 10 dB. Moreover, the PGPE values achieved by the proposed method still remain acceptable for low levels of SNR (as low as -10 dB).

The RMSE resulting from using the five pitch estimation methods for the female and male speaker groups is shown in Figs. 5.14 and 5.15, respectively. It is observed

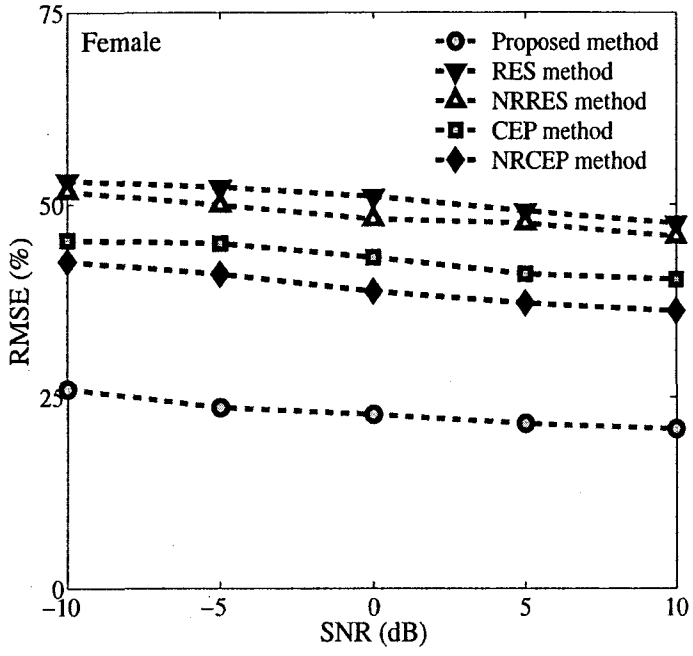


Figure 5.14: RMSE (%) as a function of SNR for female speaker group in babble noise.

from these figures that the proposed PCELS-ES method continues to provide significantly lower errors even at a very low SNR of −10 dB, whereas the performance of the other four methods remains poor at SNR levels below 10 dB.

For the same set of 10 mixed (5 female plus 5 male) speakers as the one used in Figs. 5.10 and 5.11, the mean m_{FPE} and standard deviation σ_{FPE} obtained from the five methods as a function of SNR are plotted in Figs. 5.16 and 5.17, respectively, as an illustration of the accuracy of the proposed pitch estimation method. As expected, the estimation accuracy of the proposed PCELS-ES method is reduced in comparison to that of the case of the white noise corruption. Although its performance is comparable only to CEP and NRCEP methods at high SNR equal to or above 5 dB, the accuracy remains considerably better than that provided by the other four methods for lower

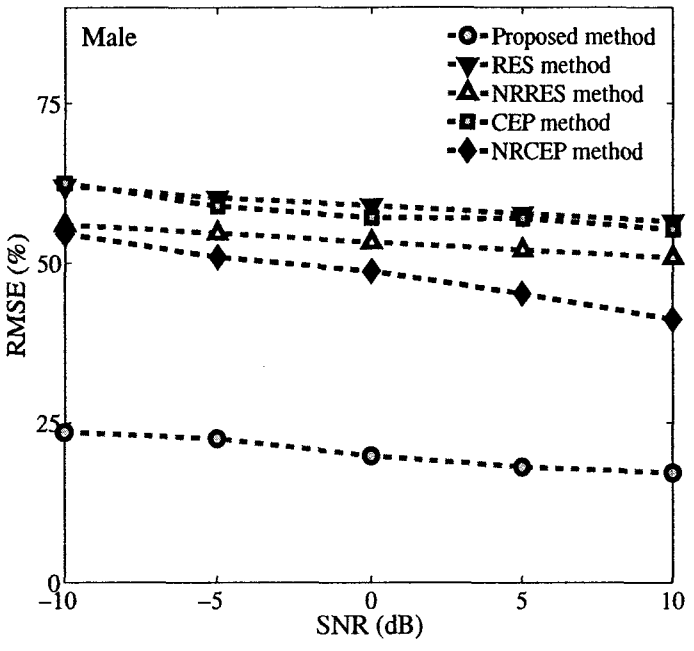


Figure 5.15: RMSE (%) as a function of SNR for male speaker group in babble noise.

SNR levels as low as −10 dB.

Fig. 5.18 shows a reference pitch contour accompanied by the spectrogram of a 1.4-second excerpt of the clean speech of a male speaker from the reference database. In the same figure, the pitch contours from other pitch estimation methods are also overlaid on the spectrograms of the white noise-corrupted speech. It is seen from the figure that, in contrast to the other methods, the pitch contour resulting from the proposed method is comparatively smoother even at SNR of −10 dB. Similarly, Fig. 5.18 illustrates a comparison of the pitch contours resulting from the five methods for female speech corrupted by babble noise at SNR −10 dB. From Fig. 5.18, it is noted that the proposed method yields a smoother contour even in the presence of babble noise. The pitch contours obtained from the five methods and shown in Figs. 5.18

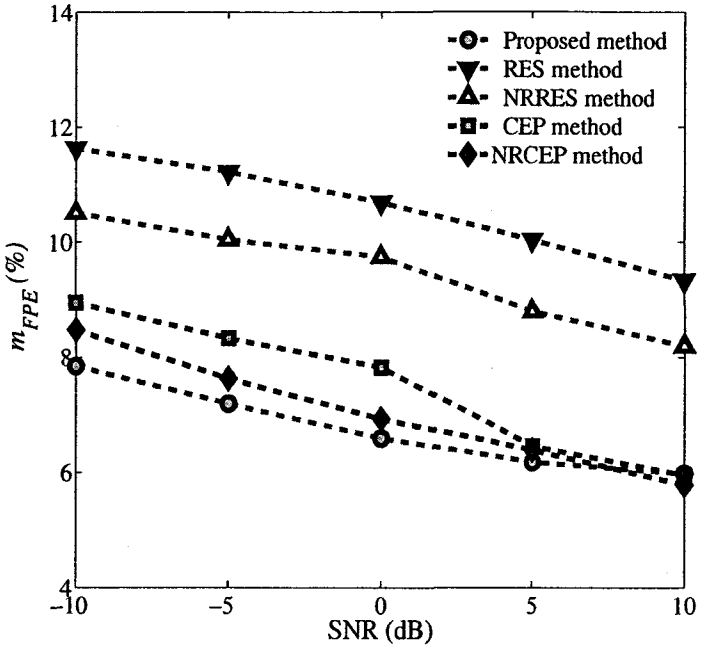


Figure 5.16: m_{FPE} (%) as a function of SNR for all female and male speakers in babble noise.

and 5.19 have convincingly demonstrated that the double and half-pitch errors in the proposed method can be significantly reduced by the use of the proposed prior noise reduction scheme.

Thus, the noise reduction applied prior to the proposed PCELS-ES method renders its suitability for practical applications involving severely noise-corrupted speech.

5.7 Conclusion

In this Chapter, to make use of the superiority of the ELS to the speech itself in highlighting the pitch information even at a very low SNR, a new method using the ELS obtained from the noise-reduced speech has been presented for the estimation of pitch from severely noise-corrupted speech.

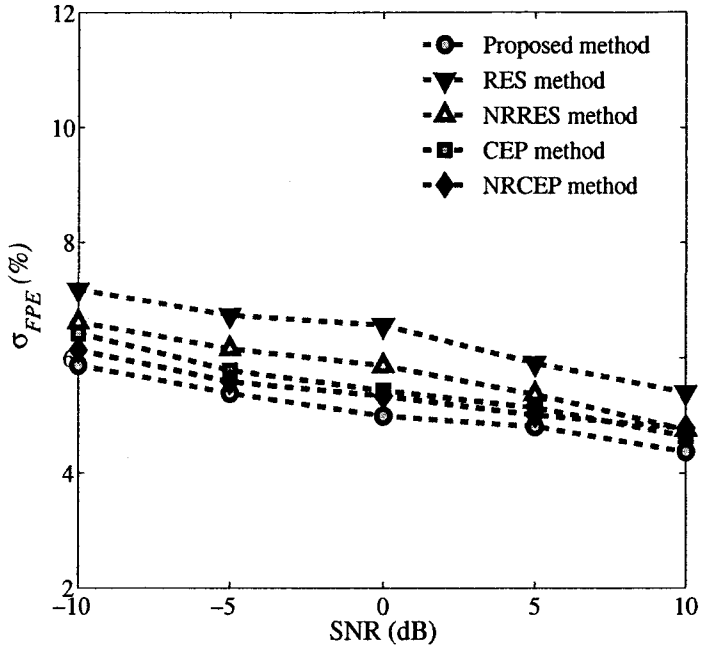


Figure 5.17: σ_{FPE} (%) as a function of SNR for all female and male speakers in babble noise .

The significant features of the introduced new frequency-domain noise-reduction scheme, prior to pitch estimation, lie in that it takes into account the cross-correlation between speech and noise and offers the advantages of noise being updated with time and adjusted at each frame. As a result of using the prior noise reduction and homomorphic deconvolution for identification of the VTS parameters, the RS produced by employing the estimated VTS parameters does not need a separate noise compensation. It has then been shown that an SHE of the RS can sufficiently represent the ELS. Considering that the conventional cepstrum widely used in speech processing has limited capability of handling noise, a more robust method based on a time-frequency domain pseudo-cepstrum of the ELS has been introduced to estimate the pitch at a very low SNR.

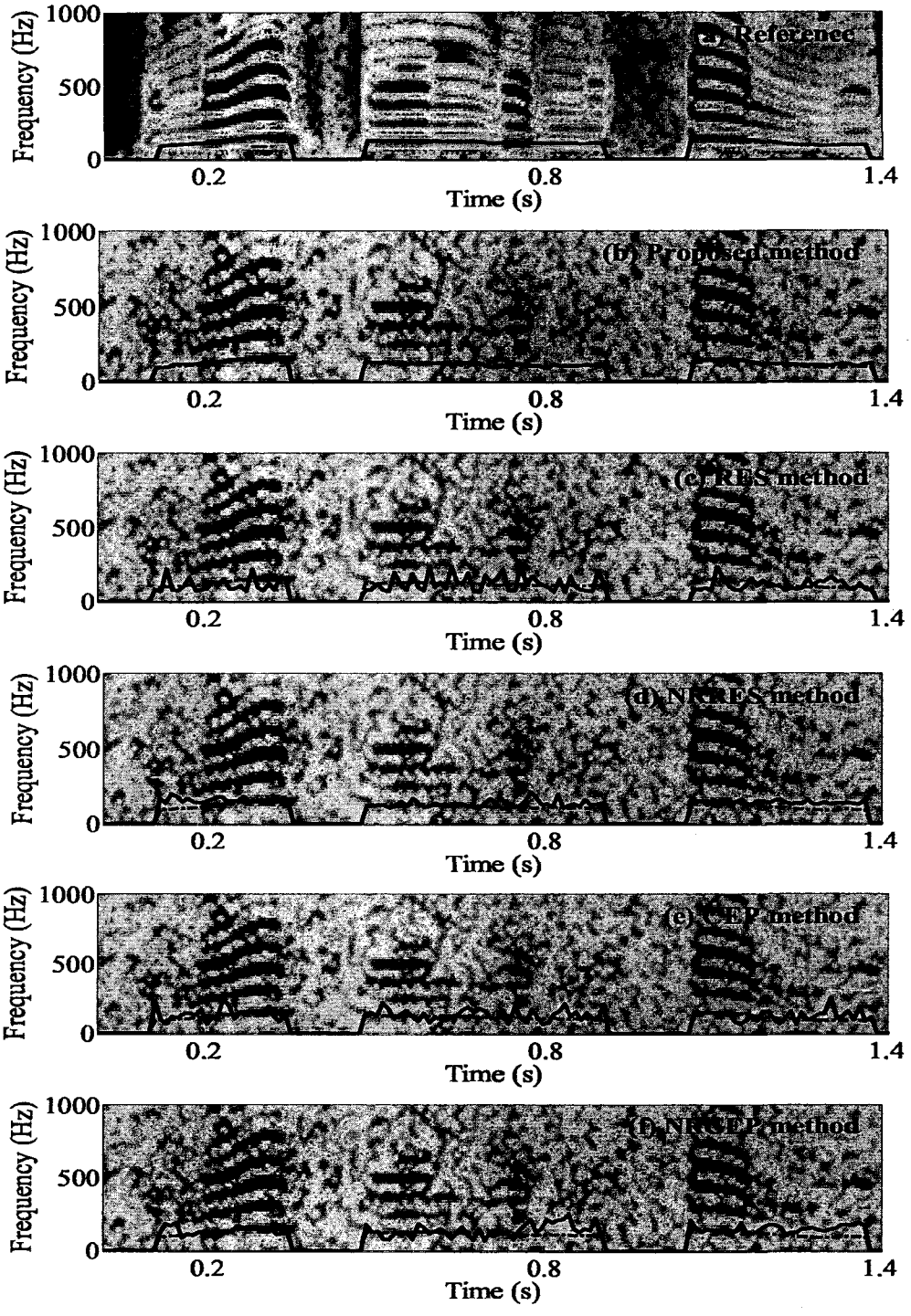


Figure 5.18: Pitch contours of different methods at SNR $= -10$ dB in white noise.

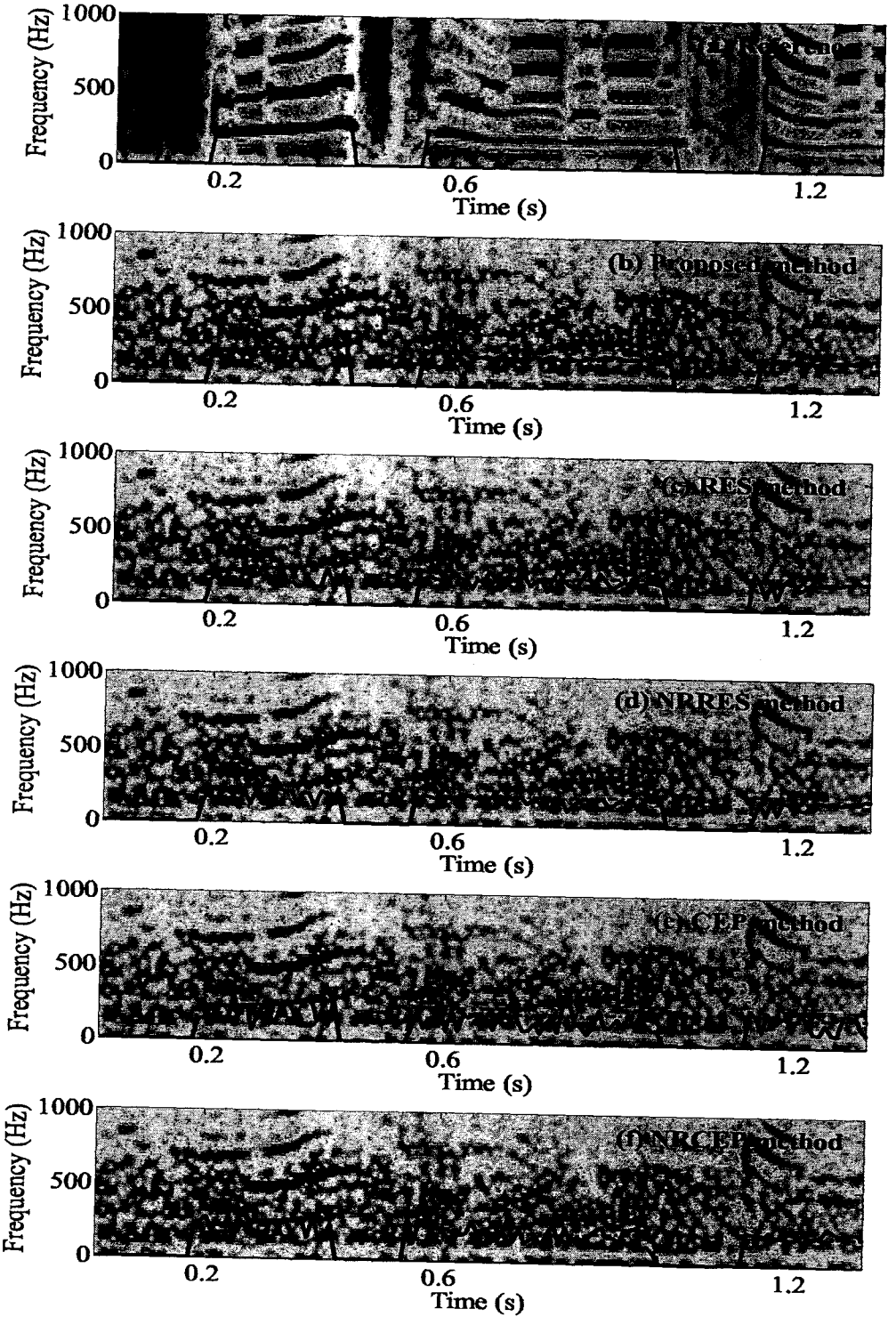


Figure 5.19: Pitch contours of different methods at SNR = -10 dB in Babble noise.

Extensive simulations have been carried out to demonstrate the performance of the proposed time-frequency domain method, referred to as PCELS-ES. It has been shown that the new method outperforms other existing methods in comparison. The pitch estimation in the presence of multi-talker babble noise has also been attempted and a quite accurate pitch estimate obtained by the proposed method illustrates its applicability in real-life uses.

Chapter 6

Conclusion

6.1 Concluding Remarks

Pitch estimation from given noisy speech observations plays an eminent role in many practical applications in speech communication, such as speaker recognition, speech synthesis, coding, enhancement and articulation training for the deaf. As far as real-life applications are concerned, pitch estimation using only the noise-corrupted speech observations under a very low SNR is a very difficult problem; yet there is a strong demand for noise-robust pitch estimation methods. Most of the existing methods deal with pitch estimation of clean speech. Only a few research results are available in the literature for pitch estimation from noisy speech, especially in the presence of real-world noise. In this dissertation, some effective methodologies for pitch estimation, which are able to estimate pitch accurately from the noise-corrupted speech observations under very low levels of SNR, have been developed.

The new pitch estimation methods developed in this thesis are classified into two approaches. The first approach uses directly the noisy speech for pitch estimation whereas, in the second approach, an excitation-like signal (ELS), generated from the noisy speech or its noise-reduced version is employed for developing the new time-

frequency domain methods in order to tackle the pitch estimation problem even in severe noisy conditions, where the SNR is very low.

In the first approach, a new harmonic cosine autocorrelation (HCAC) model of clean speech has been introduced and then employed in a least-squares autocorrelation-fitting optimization technique to estimate a pitch-harmonic (PH) from noisy speech observations. By exploiting the extracted PH along with the power spectrum of the noisy speech, a harmonic measure is proposed to deduce a set of harmonic numbers corresponding to the estimated PH and then a harmonic number matching (HNM) scheme based on a harmonic-to-noise power ratio (HNPR) is developed to determine the appropriate harmonic number that leads to a pitch estimate. A smoothed pitch contour is obtained by employing a pitch tracking scheme based on dynamic programming (DP).

Unlike the HCAC model derived from the statistical estimator by neglecting cross-product terms of different harmonics, a harmonic sinusoidal autocorrelation (HSAC) model of clean speech is derived from the conventional correlation estimator. The PH estimation from noisy speech observations is then carried out by employing a least-squares HSAC model-fitting optimization technique. Considering the advantageous attributes of the DCT over the FFT in the case of real signals, a smoothed DCT power spectrum is adopted in the HSAC model-based optimization to obtain an initial estimate of the PH. It has been shown that, once the PH is estimated exploiting the accurate HSAC model, the associated harmonic number required for pitch estimation can be successfully estimated at a very low SNR using the proposed symmetric average magnitude sum function (SMSF) based Impulse-train Matching (SIM) technique. The employment of a DP based pitch tracking scheme yields a smoothed pitch contour.

In the second approach, a new method employing an excitation-like signal (ELS)

obtained from the noisy speech has been presented for pitch estimation. In order to overcome the limitation of the conventional linear prediction (LP) residual, a homomorphic deconvolution (HD) based scheme is devised in the time-frequency domain to extract the VTS parameters, giving a better residual signal (RS). A squared Hilbert envelope (SHE) is then generated as an ELS by using the noise-compensated ACF of the RS. A new symmetric normalized magnitude difference function (SNMDF) of the ELS is finally employed for pitch estimation and pitch tracking.

In order to enhance the superiority of the ELS to the speech itself in pronouncing the pitch information even under a heavier noisy condition, another ELS based pitch estimation method is developed using the noise-reduced speech observations. A new frequency domain noise-reduction scheme, which takes into account the possible cross-correlation between speech and noise, and offers an additional advantage of noise updating and adjusting from time to time, is employed prior to pitch estimation. It has been shown that once the VTS parameters are obtained from the enhanced frame via HD, the ELS can be generated by employing a squared Hilbert envelope (SHE) of the RS without any noise compensation. In order to tackle the adverse effect of noise on the ELS, and considering the fact that the conventional cepstrum commonly used in speech processing is capable of handling clean speech only, a new method based on the time-frequency domain pseudo-cepstrum of the ELS of the enhanced speech, carrying the information of both magnitude and phase spectra of the ELS, has been devised for pitch estimation at a very low SNR.

The performance of the proposed pitch estimation methods has been studied through extensive experimentations under very noisy conditions, and the results have demonstrated a good robustness against the noise and superior estimation accuracy and consistency of the methods.

Some of the key contributions of the investigation undertaken in this thesis can be summarized as follows:

1. All the pitch estimation methods have been developed with a target to obtain accurate and consistent estimates of pitch from speech signals under very low levels of SNR. The estimation accuracy of the pitch estimation methods available in the literature, including those proposed to handle a noisy environment, deteriorates drastically with the increase in the level of noise. The pitch estimation methods proposed in this thesis provide a much superior performance under heavy noisy conditions.

2. Unlike most of the existing pitch estimation methods, the proposed approaches are capable of performing well even when the speech observations are corrupted by real world noise also. The performance evaluation and effectiveness of the proposed methods in a multi-talker babble-noise environment are an illustration of their suitability for real-life applications.

3. The smooth pitch contours obtained by employing DP based pitch tracking schemes, whenever applicable to the proposed methods, support that the obtained contours are readily applicable for practical use.

4. Because of a better noise immunity of the new HCAC and HSAC models developed in the correlation domain, the two methods proposed in Chapters 2 and 3 using these models in the least-squares model fitting framework provide an accurate pitch-harmonic, which is an important quantity to determine pitch in the presence of heavy noise.

5. Due to our use of homomorphic deconvolution in the process of generating an

excitation-like signal (ELS) that has the desirable impulse-like characteristic at the GC instants, the two methods developed in Chapters 4 and 5 using the ELS makes them suitable not only for low-pitched males but also for high-pitched female speakers.

6.2 Scope for Further Work

The research work undertaken in this thesis can be extended in several aspects. One interesting area of investigation could be the possible use of the excitation like signal (ELS). In one of the two approaches adopted in this thesis, an accurate estimate of pitch is obtained from the ELS, which consists of the significant excitation instants at the moment of glottal closures (GC) in a voiced speech. As an excitation source, the ELS can play an important role in several speech analysis-by-synthesis methods where a residual signal is often used. New methods proposed in this thesis are capable of extracting precisely the ELS even under severe noisy conditions. Thus, the proposed ELS estimation schemes can be investigated in detail in applications like prosodic speech analysis. It is to be mentioned that pitch is found to be highly correlated with prosodic features such as lexical stress, tone, and sentence intonation, which provide important perceptual cues to human speech communication.

It has been shown that the proposed methods can accurately estimate pitch even under very noisy conditions. This important feature creates room for its wide applications in various speech processing tasks, such as natural language synthesis, speaker recognition, emotion detection, and many possible musical applications like sound transformations. Incorporation of pitch tracking data helps a recognition system to increase the recognition accuracy and provide a more natural-like synthesized voice. Therefore, a noise-robust pitch estimator is in a great demand. However, in these

applications, pitch estimation is done with a prior V/UV detection, which we assumed known. Many criteria and functions proposed in this thesis can be potentially employed to detect voiced frames and genders. Therefore, an overall system incorporating pitch estimator, and voicing and gender detector could be an interesting topic of future investigation.

Bibliography

[1] D. O'Shaughnessy, *Speech Communications Human and Machine*, 2nd ed. NY: IEEE Press, 2000.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Upper Saddle River, NJ: Prentice-Hall, 2002.

[3] N. Roman and D. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 458–469, July 2006.

[4] B. S. Atal, "Automatic speaker recognition bsed on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 1, pp. 1687–1697, 1972.

[5] A. E. Rosenberg and M. R. Sabmur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 169–176, 1975.

[6] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception.* New York: Springer-Verlag, 1972.

[7] D. W. Griffith and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223–1235, 1988.

[8] M. Ostendorf and K. Ross, *A multi-level model for recognition of intonation labels.* New York: Springer-Verlag, 1997, pp. 291–308.

[9] E. Chang, J. Zhou, S. Di, C. Huang, and K.-F. Lee, "Large vocabulary mandarin speech recognition with different approaches in modeling tones," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Beijing, China, 2000, pp. 983–986.

[10] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, UT, May 2001, pp. 125–128.

[11] H. Levitt, "Speech processing aids for the deaf: an overview," *IEEE Trans. Audio Electroacoust.*, vol. 21, pp. 269–273, 1973.

[12] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of the F0 contours for computer aided intonation teaching," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, Berlin, Germany, 1993, pp. 1003–1006.

[13] S. A. Zahorian, A. Zimmer, , and B. Dai, "Personal computer software vowel training aid for the hearing impaired," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seattle, WA, May 1998, pp. 3625–3628.

[14] W. J. Hess, *Pitch Determination of Speech Signals.* New York: Springer-Verlag, 1993.

[15] L. R. Rabiner, M. J. Cheng, A. H. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 5, pp. 399–417, 1976.

[16] P. Veprek and M. S. Scordilis, "Analysis, enhancement and evaluation of five pitch determination techniques," *Speech Commun.*, vol. 37, pp. 249–270, 2002.

[17] N. Miller, "Pitch detection by data reduction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 72–79, 1975.

[18] 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.

[19] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 6, pp. 562–570, 1975.

[20] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 12, pp. 1805–1815, Dec. 1989.

[21] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1966.

[22] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, pp. 829–834, 1968.

[23] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust*, vol. 20, pp. 367–377, 1972.

[24] R. J. McAuley and T. F. Quartieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Albuquerque, NM, Apr. 1990, pp. 249–252.

[25] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing.* Upper Saddle River, NJ: Prentice-Hall, 2001.

[26] D.-J. Liu and C.-T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 9, pp. 609–621, 2001.

[27] B. Resch, M. Nilsson, A. Ekman, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 3, pp. 813–822, March 2007.

[28] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 2–8, 1976.

[29] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 24–33, 1977.

[30] J. C. Brown and M. S. Puckette, "Calculation of a narrowed autocorrelation function," *J. Acoust. Soc. Amer.*, vol. 85, pp. 1595–1601, 1989.

[31] A. D. Chevengne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1930, 2002.

[32] M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. 16, pp. 262–266, 1968.

[33] M. Ross, H. Schafer, A. Cohen, R. Freuberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 22, pp. 353–362, 1974.

[34] O. Deshmukh, J. Singh, and C. E.-Wilson, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," *IEEE Trans. Speech Audio Processing*, vol. 13, pp. 776–786, 2005.

[35] L. Hui, B.-Q. Dai, and L. Wei, "A pitch detection algorithm based on AMDF and ACF," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toulouse, France, May 2006, pp. 377–380.

[36] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, 1988.

[37] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Orlando, FL, May 2002, pp. 333–336.

[38] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. 1343–1346.

[39] H. Friedman, "Pseudo-maximum-likelihood speech pitch extraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 213–221, 1977.

[40] C. Un and S.-C.Yang, "A pitch extraction algorithm based on LPC inverse filtering and AMDF," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 565–572, 1977.

[41] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. NY: Wiley, IEEE Press, 1999.

[42] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, Feb. 1970.

[43] G. S. Ying, L. H. Jamieson, and C. D. Michell, "A probabilistic approach to AMDF pitch detection," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Philadelphia, PA, Oct. 1996, pp. 1201–1204.

[44] K. A. Oh and C. K. Un, "A performance comparison of pitch extraction algorithms for noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Diego, CA, Mar. 1984, pp. 18B4.1–18B4.4.

[45] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 319–329, 1991.

[46] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 727–730, 2001.

[47] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, Quebec, Canada, May 2004, pp. 109–112.

[48] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3690–3700, 2004.

[49] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Pitch estimation based on harmonic cosine autocorrelation model and frequency-domain matching," *submitted to IET Signal Process.*

[50] ——, "Robust pitch estimation at very low SNR exploiting time and frequency domain cues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Mar. 2005, pp. 389–392.

[51] ——, "A pitch extraction algorithm in noise based on temporal and spectral representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Mar. 2008, pp. 4477–4480.

[52] S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, "A novel technique for the identification of ARMA systems under very low levels of SNR," *IEEE Trans. Circuits Syst. I*, vol. 55, no. 7, pp. 1988–2001, Aug. 2008.

[53] Z. Lin, R. A. Goubran, and R. M. Dansereau, "Noise estimation using speech/non-speech frame decision and subband spectral tracking," *Speech Commun.*, vol. 49, pp. 542–557, 2007.

[54] L. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing.* New Jersey: Prentice-Hall, Inc., 1975.

[55] S. M. Kay, *Modern Spectral Estimation, Theory and Application.* Englewood Cliffs, NJ: Prentice-Hall Ltd., 1988.

[56] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A spectral matching method for pitch estimation from noise-corrupted speech," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Taipei, Taiwan, May 2009, pp. 1413–1416.

[57] ——, "A method for pitch estimation from noisy speech signals based on a pitch-harmonic extraction," in *Proc. IEEE Int. Conf. Neural Networks, Signal Process. (ICNNSP)*, Zhenjiang, China, Jun. 2008, pp. 120–123, (**Received the Best Student Paper Award**).

[58] ——, "On the estimation of pitch of noisy speech based on time and frequency domain representations," in *Proc. IEEE Canadian Conf. on Elect. and Comp. Eng. (CCECE)*, Niagara Falls, Canada, May 2008, pp. 1819–1822.

[59] D. Talkin, "A robust algorithm for pitch tracking," in *Speech Coding and Synthesis*, Eds. Amsterdam, The Netherlands, 1995, pp. 495–518.

[60] R. Schwartz and Y.-L. Chow, "The N-best algorithm: an efficient and exact procedure to finding the N most likely sentence hypothesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Albuquerque, NM, Apr. 1990, pp. 81–84.

[61] J.-K. Chen and F. Soong, "An N-best candidates-based discriminative training for speech recognition applications," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 206–216, 1994.

[62] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech Language Processing*, vol. 15, pp. 34–43, 2007.

[63] G. Meyer, F Plante and W. A. Ainsworth. Keele Pitch Database. Contact: Dr G.F. Meyer (georg@liv.ac.uk). [Online]. Available: http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html.

[64] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, Madrid, Spain, 1995, pp. 827–840.

[65] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.

[66] A. V. Oppenheim, R. Schafer, and J. R. Buck, *Discrete-Time Signal Processing.* Upper Saddle River, NJ: Prentice-Hall, 1999.

[67] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Pitch estimation based on harmonic sinusoidal autocorrelation model and time-domain matching," *submitted to IEEE Trans. Audio, Speech, Language Process.*

[68] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A dct domain pitch extractor for speech signals under noisy conditions," *Presented in the Centre for Advanced Systems and Communications (SYTACom) research workshop, Montreal, Canada*, May 2008.

[69] G. Strang, "The discrete cosine transform," *SIAM Rev.*, vol. 41, pp. 135–147, 1999.

[70] F. Ferdousi, A. T. Connie, M. Sharmin, and M. R. Khan, "System identification at an extremely low SNR using energy density in DCT domain," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 289–292, Apr. 2005.

[71] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A robust pitch estimation algorithm in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Honolulu, HI, Apr. 2007, pp. 1073–1076.

[72] ——, "A temporal matching method for pitch determination from noisy speech signals," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Knoxville, TN, Aug. 2008, pp. 938–941, (**selected as among the top ten finialists in the student paper contest**).

[73] A. T. Connie, F. Ferdousi, M. Sharmin, and M. R. Khan, "Identification of AR parameters at a very low SNR using estimated spectral distribution in DCT domain," *IEE Proc. Vis. Image Signal Process.*, vol. 153, no. 2, pp. 95–100, Apr. 2006.

[74] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 4, pp. 309–319, Aug. 1979.

[75] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, pp. 762–765, 2007.

[76] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[77] J. Makhoul and J. I. Wolf, "Linear prediction and the spectral analysis of speech," Bolt, Beranek, and Newman Inc., Cambridge, MA, Technical rept. 2304, 1972.

[78] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[79] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[80] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Commun.*, vol. 38, pp. 141–160, 2002.

[81] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A time-domain pitch extraction scheme for noisy speech signals," *Canadian Acoustics*, vol. 35, no. 3, pp. 114–115, Oct. 2007.

[82] ——, "A residual-cepstrum method of pitch estimation from noisy speech," *to be published in Canadian Acoustics*, Oct. 2009.

[83] ——, "An approach for voiced/unvoiced decision of colored noise-corrupted speech," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, New Orleans, LA, May 2007, pp. 3944–3947.

[84] ——, "A bifeature voiced/unvoiced discrimination algorithm for speech signals in the presense of noise," in *Proc. IEEE Int. Northeast Workshop Circuits Syst. (NEWCAS)*, Montreal, Canada, Aug. 2007, pp. 89–92.

[85] ——, "A robust pitch estimation approach for colored noise-corrupted speech," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, kobe, japan, May 2005, pp. 3143–3146.

[86] ——, "A multifeature voiced/unvoiced decision algorithm for noisy speech," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Island of Kos, Greece, May 2006, pp. 25–28.

[87] ——, "A new technique for the estimation of jitter and shimmer of voiced speech signal," in *Proc. IEEE Canadian Conf. on Elect. and Comp. Eng. (CCECE)*, Ottawa, Canada, May 2006, pp. 2112–2115.

[88] ——, "Pitch estimation based on magnitude difference function of an excitation-like signal obtained from noisy speech," *to be submitted to Speech Commun.*

[89] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio and Electroacoustics*, vol. 16, pp. 221–226, 1968.

[90] F. Wang and P. Yip, "Cepstrum analysis using discrete trigonometric transform," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 538–541, Feb. 1991.

[91] B. J. Shannon and K. K. Paliwal, "Spectral estimation using higher-leag autocorrelation coefficients with applications to speech recognition," in *in Proc IEEE int. symp. Signal Process. and its Applications*, Sydney, Australia, Aug. 2005, pp. 599–602.

[92] S. A. Fattah, W.-P. Zhu, and M. O. Ahmad, "Identification of autoregressive systems in noise based on a ramp cepstrum model," *IEEE Transaction on Circuits and Systems II*, vol. 55, no. 10, pp. 1051–1055, Oct. 2008.

[93] Y. T. Chan and R. P. Langford, "Spectral estimation via the high-order yule-walker equations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 5, pp. 689–698, Oct. 1982.

[94] J. S. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.

[95] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1235–1238, 1984.

[96] W. Verhelst and O. Steenhaut, "A new model for the shorttime complex cepstrum of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 43–51, 1986.

[97] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on melfrequency cepstra for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Mar. 2008, pp. 4041–4044.

[98] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. Audio, Speech Language Processing*, vol. 15, no. 8, pp. 2257–2269, 2007.

[99] H. Quast, O. Schreiner, and M. R. Schroeder, "Robust pitch tracking in the car environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, May 2002, pp. 353–356.

[100] K. Ymagisawa, K. Tanaka, and I. Yamaura, "Detection of the fundamental frequency in noisy environment for speech enhancement of a hearing aid," in *Proc.IEEE Int. Conf. Contr. Applications (ICCA)*, 1999, pp. 1331–1335.

[101] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A robust pitch estimation scheme for speech enhancement," *Presented in the Centre for Advanced Systems and*

Communications (SYTACom) research workshop, Quebec City, Canada, May 2007.

[102] J. S. Lim, *Speech Enhancement.* Englewood Cliffs, NJ: Prentice-Hall Ltd., 1983.

[103] D. O'Shaughnessy, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Comm. Mag.*, vol. 27, no. 2, pp. 46–52, Feb. 1989.

[104] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[105] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

[106] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[107] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A spectro-cepstral method of pitch estimation from noisy speech," *Presented in the Centre for Advanced Systems and Communications (SYTACom) research workshop, Montreal, Canada*, Apr. 2009, (**Selected as one of 15 finalists in research presentation competition**).

[108] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A cepstral-domain algorithm for pitch estimation from noise-corrupted speech," *Canadian Acoustics*, vol. 36, no. 3, pp. 80–81, Sep. 2008.

[109] ——, "A spectro-temporal algorithm for pitch frequency estimation from noisy observations," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Seattle,WA, May 2008, pp. 1704–1707.

[110] ——, "A pitch detection method for speech signals with low signal-to-noise ratio," in *Proc. IEEE Int. Symp. Signals, Systems Elect. (ISSSE)*, Montreal, Canada, Aug. 2007, pp. 399–402.

[111] ——, "On extracting pitch from noisy speech signals based on spectral and temporal enhancement," in *Proc. IEEE Int. Northeast Workshop Circuits Syst. (NEWCAS)*, Montreal, Canada, Jun. 2008, pp. 77–80.

[112] ——, "A robust pitch detection algorithm for speech signals in a practical noisy environment," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Montreal, Canada, Aug. 2007, pp. 385–388.

[113] ——, "An approach for pitch estimation from noisy speech," in *Proc. IEEE Canadian Conf. on Elect. and Comp. Eng. (CCECE)*, Vancouver, Canada, Apr. 2007, pp. 1590–1593.

[114] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Processing Mag.*, vol. 21, no. 5, pp. 95–106, Sep. 2004.

[115] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Signal Processing Mag.*, vol. 12, no. 3, pp. 218–233, May 2004.

[116] J. Xu, J. Cheng, and Y. Wu, "A cepstral method for analysis of acoustic transmission characteristics of respiratory system," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 5, pp. 660–664, May 1998.

[117] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 2, pp. 425–434, Feb. 2006.

[118] S. Young, "A review of large-vocabulary continuous speech recognition," *IEEE Signal Processing Mag.*, pp. 45–57, 1996.

[119] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[120] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall Ltd., 1993.

[121] H. K. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 435–446, Sept. 2003.

[122] C. Shahnaz, W.-P. Zhu, and M. O. Ahmad, "A pitch detector for noise-corrupted speech," *Presented in the Centre for Advanced Systems and Communications (SYTACom) research poster session, Montreal, Canada*, Jul. 2008.

[123] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Pitch estimation using pseudo cepstrum of an excitation-like signal obtained from enhanced speech," *to be submitted to Eurasip Journal Speech Audio Music Processing.*

[124] F. Charpentier, "Pitch detection using the short-term phase spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Tokyo, Japan, Apr. 1986, pp. 113–116.

[125] K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, Geneva, Switzerland, 2003, pp. 2117–2120.

[126] K. Paliwal and B. Atal, "Frequency-related representation of speech," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, Geneva, Switzerland, 2003, pp. 65–68.

[127] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Hong Kong, Apr. 2003, pp. 68–71.

[128] R. Smiths and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 9, pp. 325–333, 1995.

[129] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 609–619, 1999.

[130] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, Yokohama, Japan, 1994, pp. 1182–1185.

[131] J. Shon, N. S. Kim, and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Seattle, WA, May 1998, pp. 365–368.

Appendix A

Derivation of the SNMDF of the ELS in Noise

By introducing an equality term $\Gamma_{\widetilde{\psi}}(m) \geq 0$, (4.36) can be easily written as

$$\widetilde{\psi}(m) = \frac{\sum_{n=0}^{Q-1} |E_r(l) - E_r(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|} + \frac{\sum_{n=0}^{Q-1} |E_w(l) - E_w(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|} - \Gamma_{\widetilde{\psi}}(m), \quad m \in [0, 1, \ldots, m_s]. \tag{A.1}$$

The first term in the right hand side can be expressed as

$$\begin{aligned} \frac{\sum_{n=0}^{Q-1} |E_r(l) - E_r(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|} &= \frac{\sum_{n=0}^{Q-1} |E_r(l) - E_r(n)|}{\sum_{n=0}^{K_m} |E_r(n)| + \sum_{n=0}^{K_m} |E_w(n)| - \sum_{n=0}^{K_m} \varepsilon_r(n)} \\ &= \psi(m) - \Gamma_{\psi}(m), \end{aligned} \tag{A.2}$$

where $\sum_{n=0}^{K_m} \varepsilon_r(n) \geq 0$, $\psi(m)$ is the SNMDF of $E_r(n)$ as defined in (4.31) and $\Gamma_{\psi}(m)$ can be derived as

$$\Gamma_{\psi}(m) = \frac{\left[\sum_{n=0}^{K_m} |E_w(n)| - \sum_{n=0}^{K_m} \varepsilon_r(n)\right] \psi(m)}{\sum_{n=0}^{K_m} |E_r(n)| + \sum_{n=0}^{K_m} |E_w(n)| - \sum_{n=0}^{K_m} \varepsilon_r(n)}. \tag{A.3}$$

Similarly, the second term in the right hand side of (A.1) can be expressed as

$$\frac{\sum_{n=0}^{Q-1} |E_w(l) - E_w(n)|}{\sum_{n=0}^{K_m} |E_r(n) + E_w(n)|} = \psi_w(m) - \Gamma_w(m), \tag{A.4}$$

where

$$\psi_w(m) = \frac{\sum_{n=0}^{Q-1} |E_w(l) - E_w(n)|}{\sum_{n=0}^{K_m} |E_w(n)|}, \quad m \in [0, 1, \ldots, m_s] \tag{A.5}$$

$\psi_w(m)$ is the SNMDF of $E_w(n)$ which is introduced in (4.30) and $\Gamma_w(m)$ can be derived as

$$\Gamma_w(m) = \frac{\left[\sum_{n=0}^{K_m} |E_r(n)| - \sum_{n=0}^{K_m} \varepsilon_r(n)\right] \psi_w(m)}{\sum_{n=0}^{K_m} |E_r(n)| + \sum_{n=0}^{K_m} |E_w(n)| - \sum_{n=0}^{K_m} \varepsilon_r(n)}. \tag{A.6}$$

By using (A.2) and (A.4), (A.1) can be expressed as

$$\begin{aligned} \widetilde{\psi}(m) &= \psi(m) + \left[\psi_w(m) + \left(-\Gamma_{\widetilde{\psi}}(m) - \Gamma_{\psi}(m) - \Gamma_w(m)\right)\right], \quad m \in [0, 1, \ldots, m_s] \\ &= \psi(m) + \psi_w(m) + \Gamma_s(m). \end{aligned} \tag{A.7}$$

Denoting $\widetilde{\Gamma}(m) = [\psi_w(m) + \Gamma_s(m)]$, where $\Gamma_s(m)$ represents the sum of equality terms, (A.7) can be written as (4.37).