Brunel
UNIVERSITY
LONDON

MODELLING AND EXTRACTION OF FUNDAMENTAL FREQUENCY IN SPEECH SIGNALS

A thesis submitted for the degree of

Doctor of Philosophy

by

ALIPAH PAWI

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

BRUNEL UNIVERSITY

17 MAY 2013

Abstract

One of the most important parameters of speech is the fundamental frequency of vibration of voiced sounds. The audio sensation of the fundamental frequency is known as the pitch. Depending on the tonal/non-tonal category of language, the fundamental frequency conveys intonation, pragmatics and meaning. In addition the fundamental frequency and intonation carry speaker gender, age, identity, speaking style and emotional state.

Accurate estimation of the fundamental frequency is critically important for functioning of speech processing applications such as speech coding, speech recognition, speech synthesis and voice morphing.

This thesis makes contributions to the development of accurate pitch estimation research in three distinct ways: (1) an investigation of the impact of the window length on pitch estimation error, (2) an investigation of the use of the higher order moments and (3) an investigation of an analysis-synthesis method for selection of the best pitch value among *N* proposed candidates.

Experimental evaluations show that the length of the speech window has a major impact on the accuracy of pitch estimation. Depending on the similarity criteria and the order of the statistical moment a window length of 37 to 80 ms gives the least error. In order to avoid excessive delay as a consequence of using a longer window, a method is proposed

where the current short window is concatenated with the previous frames to form a longer signal window for pitch extraction.

The use of second order and higher order moments, and the magnitude difference function, as the similarity criteria were explored and compared. A novel method of calculation of moments is introduced where the signal is split, i.e. rectified, into positive and negative valued samples. The moments for the positive and negative parts of the signal are computed separately and combined. The new method of calculation of moments from positive and negative parts and the higher order criteria provide competitive results.

A challenging issue in pitch estimation is the determination of the best candidate from N extrema of the similarity criteria. The analysis-synthesis method proposed in this thesis selects the pitch candidate that provides the best reproduction (synthesis) of the harmonic spectrum of the original speech. The synthesis method must be such that the distortion increases with the increasing error in the estimate of the fundamental frequency. To this end a new method of spectral synthesis is proposed using an estimate of the spectral envelop and harmonically spaced asymmetric Gaussian pulses as excitation. The N-best method provides consistent reduction in pitch estimation error.

The methods described in this thesis result in a significant improvement in the pitch accuracy and outperform the benchmark YIN method.

Statement of copyright

Table of Contents

List of Figure

List of Table

List of Abbreviations

ACF	Autocorrelation Function
AMDF	Average Magnitude Difference Function
ARMA	Autoregressive Moving Average
BW	Bandwidth
CCF	Cross-Correlation Function
dB	Decibel
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
FIR	Finite Impulse Response
FPE	Fine Percentage Error
FWHM	Full Width at Half Magnitude,
F$W\alpha$M	Full Width α Maximum
FWQM	Full Width Quarter Maximum
GMM	Gaussian mixture model
GPE	Gross Percentage Error
GSM	Global System for mobile
HD	Harmonicity Distance
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HOM	Higher Order Moment
Hz	Hertz
IF	Instantaneous Frequency (IF)
IPA	International phonetic Association
LF	Liljencrants-Fant
LMS	Least mean square
LP	Linear Prediction
LPC	Linear Prediction Coefficient
LSE	Least Squared Error

LSF	Line Spectral Frequency
MA	Moving Average
MAP	Maximum A-Posteriori
MHOM	Modified Higher Order Model
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
Ms	Millisecond
MSE	Mean Squared Error
NCCF	Normalized Cross-Correlation Function
p.d.f	Probability Density Function
PCHIP	Piecewise Cubic Hermite Interpolation Polynomial
Sec	Second
SMDF	Squared Magnitude Difference Function
SNR	Signal to Noise Ratio
STFT	short-term Fourier transform
Std.	Standard deviation
VoIP	Voice over Internet Protocol
w.r.t.	With respect to
WHD	Weighted Harmonicity Distance
WMMSE	Weighted Minimum Mean Squared Error
WSNR	Weighted Signal to Noise Ratio

List of Symbols

Chapter	Notation	Description
2,5,6	F_0	Fundamental Frequency
2,5,6	T_0	Period of one cycle, pitch
2,6	t	Time index
2	m	Discrete index
2,6	T	Period of time or time lag
2	F	Frequency domain
2	N	Number of sample
2,6	$x(t)$	Speech time signal
2,3	$x(m)$	Discrete speech signal
2,5	N_h	Number of harmonics
2,3,5	$a_k(m)$	Time-varying amplitude, LP parameter
2	$\theta(m)$	Time varying phase
2	$v(m)$	Non-periodic noise
2,5	$F_0(m)$	Time-varying fundamental frequency
2	kF_0	Integer multiples of harmonic
2	$r_{xx}(T)$	Autocorrelation function
2	$x(m-T)$	Time lag of T samples
2	σ_x^2	Signal variance
2	$E(T)$	Power in time
2	T_{min}	Minimum time period
2	T_{max}	Maximum time period
2	$d(T)$	AMDF function
2	$C_k(T)$	General expression of K^{th} order moment
2	Z	Zero-crossing expression

Chapter	Notation	Description
2	$E(F)$	Power spectrum
2	K	Number of order moments
2	$X(kF)$	Integer multiple of speech spectrum
2	$C(t)$	Cepstrum expression of $x(t)$
2	$\varphi(f)$	Instantaneous frequency spectrum
2	$\frac{d\theta(t)}{dt}$	The derivative of the phase with respect to time
2	$H(z)$	Transfer function of comb filter
2	$\alpha = [.]$	LP model coefficients
2	$ML(\hat{F}_0)$	Maximum likelihood pitch estimation
2	e	The residue
2	$f(F_0)$	Prior
2,3	$f(x\|F_0)$	Likelihood of x
2	$f(F_0\|x)$	Posteriori pdf
3	$V_{LF}(t)$	The Liljencrants-Fant (LF) model
3	ω	Angle frequency
3	$E_0 e^{\alpha t}$	An exponential envelope
3	$E(f)$	Excitation spectrum
3	G	Gain factor
3	$G/A(f)$	Spectrum of LP model
3	$\varepsilon(m)$	Speech excitation, noise model
3	$A(z)$	Inverse linear predictor filter
3	ω_i	LSF parameter
3	c_k	Linear prediction model coefficients

Chapter	Notation	Description
3	ee^T	Squared error function
3	S	Matrix of sine and cosine function
3	c	Vector of amplitude of the harmonic
3	H_k	Harmonicity of the speech signal
3	$X(f)$	Discrete Fourier transform of the speech signal
3	$M(f)$	Gaussian shaped function
3	$N(f)$	Rayleigh distribution random variable
3	$f(F_0\|x,a)$	Posterior probability
3	a, A	Spectral envelop, spectral envelop vector
3	$MAP(\hat{F}_0)$	Maximum a Posterior (MAP)
3	$ML(\hat{F}_0)$	Maximum likelihood
3	$E_h(\hat{F}_0)$	Harmonic of the excitation signal
3	E_n	Noise of the excitation signal
3	$C(\hat{F}_0, F_0)$	Cost function
4	$[\hat{F}_0(m)]$	The sequence of pitch estimates within each utterance
4	$\bar{F}_0(m)$	The prediction of the pitch value at frame *m* of utterance unit
4	$\tilde{F}_0(m)$	The pitch prediction error
4	$g(e(m))$	The limiting or influence function
4	$\delta(t)$	Impulse function
4	$p(t)$	Pulse function
4	$h(k)$	Impulse response of a filter

Chapter	Notation	Description
4	$\Delta(f)$	The Fourier transform of the impulse function
4	$n_i(m)$	Impulse noise sequence
4	$b(m)$	Binary-valued random sequence model
4	$n(m)$	Continuous-valued random process model of impulse amplitude
4	$\delta(.)$	The Kronecker delta function
4	$P_B(b(m))$	The probability mass function of a Bernoulli
4	$\{a_{ij}\}$	The Markovian state transition probability
4	$\{b_{ik}\}$	The state observation probability
4	S_0, a_{00} ; S_1, a_{11}	Self-loop transition probabilities
4	$b[\,,]$	FIR filter coefficients
4	y	Influence function
5	$\hat{F}_0(m)$	The pitch estimation value
5	N_L	Large window length sample
5	N_T	Maximum number of period
5	$x_+(m)$	Positive-amplitude part
5	$x_-(m)$	Negative-amplitude part
5	$\Theta_k(m)$	Time-varying phase
5	*E*	Average percentage estimated pitch error
5	*Fmax*	Maximum sample frequency
5	*Fmin*	Minimum sample frequency
5	*K*	Dimension of the moment
5	*M*	Discrete-time
5	*N*	Number of frames or segments
5	*Ns*	Current frame
5	*T*	Samples a period

Chapter	Notation	Description
5	*Tmax*	Maximum sample period
5	*Tmin*	Minimum sample period
5	*x(m)*	Discrete time signal
5	*x(m-□)*	Delayed discrete time signal
5	$E(\cdot)$	General function
5	$fn[.]$	General function
5	$v(m)$	Non-periodic component
5	*X(t)*	Speech signal in time
5	a	Proportion of locally pitch value
6	F_{0i}	*N*-best pitch candidates
6	*Fs*	Sampling frequency (Hz)
6	T_{0i}	N-best peak candidates
6	*N*	Number of peak candidates
6	f	Frequency (Hz)
6	*x(t)*	Speech signal in time
6	*X (f)*	Speech spectrum
6	*X*	Actual signal spectrum
6	X_s	Synthesised signal spectrum
6	k^{th}	Order moment
6	*M(T)*	K^{th} order moment
6	*Tmin*	Minimum period of time
6	*Tmax*	Maximum period of time
6	*Fmin*	Minimum period in frequency
6	*Fmax*	Maximum period in frequency
6	*BW*	Bandwidth of speech signal
6	X_h	Synthesized harmonic speech
6	$\hat{X}_{hi}$	Synthesized harmonic speech candidate
6	$\hat{X}$	Synthesized speech
6	N_h	Number of harmonics

Chapter	Notation	Description
6	$A(f)$	Spectral envelope
6	$G(f)$	Gaussian pulse
6	$G_k(f)$	Gaussian pulse of harmonic excitation
6	*V(f)*	Non-harmonic speech signal
6	E_h	Harmonic excitation series
6	δ	Search region
6	Δ	Deviation of k^{th} harmonic
6	*D*	Distortion measure
6	*W(k)*	Weight function

Acknowledgement

First and foremost, I want to thank the Almighty GOD for giving me the patient and strength to achieve this milestone.

I would like to express my highest and deepest gratitude to my supervisor Professor Saeed Vaseghi, for the continued guidance, support, motivation, endurance, and encouragement throughout my PhD research. I have no words to thank him enough, and his unconditional support is the most important lesson I have ever received. This thesis could not have been accomplished without his commitment and supervision. It has been a privilege for me to work with him and I hope that we continue our collaboration in the future.

I also would like to extend my sincere gratitude to my mother Tuminah Hj Nor and my family for all their love, support and encourage. It makes me very proud to know that I can count on them, in whatever I may be and at any time.

I am also would like to thankful to my fellow research at Brunel University, Ben Milner at University of East Anglia, and all my friends in University of Kuala Lumpur British Malaysian Institute, Malaysia.

I wish to thank all my friends and colleges at Brunel University especially who despite neglecting them for some time, never stop their encouragement and enthusiasm.

Alipah Pawi

1

INTRODUCTION

1.1 INTRODUCTION

Speech is composed of the spoken words and sentences; it is the main form of human communication and interaction; this is particularly true in a historical sense when the great majority of humans did not have the benefit of learning to write and read using the textual forms of communication. Even now for many individuals and for most very young people below the age of five, speech is practically the only form of communication.

The ability to use speech comes naturally and develops during the early years of a person's life at such a rate that by the age of six on average a child has a vocabulary of 13000 words, these increases to an average of 60,000 words vocabulary for a high school graduate [1].

Speech is exhaled random air fluctuations from the lung that is time and frequency modulated and temporally-spectrally shaped along the way from lung and out from the lips. At the glottis the rate of openings and closings of the vocal folds determines the fundamental frequency of the speech and the time-variation of the fundamental frequency primarily determines the intonation of speech. The fundamental frequency is perceived as pitch level, a low value of the fundamental frequency is perceived as a low pitch and a high value of fundamental frequency is perceived as a high pitch, and the intonations are the trajectory of changes in the pitch level.

Speech signals are multilayered in that they simultaneously contain different forms of information, i.e. segmental and supra-segmental, conveyed mainly by the spoken form of word sequence and by the intended intonation [2].

At a word-sequence level layer, speech is composed of phrases, sentences and paragraphs each of which is composed of a number of words. The words themselves are combinations of elementary phonetic units; the arrangement of the words follows the constraints set by the rules of the grammar of the language. Speech is spoken as a sequence of connected and co-articulated words, where the articulation of each word is affected by the context of the previous and succeeding words. The degree of co-articulation of words depends on the personal style of speaking, the speaking rate, the accent and the emotional state of the speaker.

At the supra-segmental level that is at the pitch intonation level speech signals convey phrase/sentence boundaries, punctuations, pragmatics, intent, emotion, accent intonation and the state of health and mind of the speaker.

The function of the pitch intonations varies with languages. In tonal languages, such as Chinese and some African languages and to some extent Japanese, a change in the tone or the pitch of a word can completely change the meaning of the word; i.e. seemingly same sounding words (particularly to a foreign or non-native speaker), with different tones may have entirely different meanings [3].

In non-tonal languages, such as the English language, the function of the pitch is to convey supra-segmental information. For example a change of pitch may:

- Distinguish between a question or a statement,
- Signal the intent of the speaker,
- Be used to stress/emphasis a part of speech,
- Signal reactions such as approval, surprise or boredom,
- Signal emotions such as anger, happiness, contentment, indifference etc.,
- Signal phrase, sentence boundaries.

Speech processing methods, employed in voice communication technology, deal mainly with three broad application areas of speech coding, speech synthesis and speech recognition. Speech processing methods are concerned with the accurate estimation and efficient representation and reconstruction of speech parameters.

Speech is composed of a time-varying mixture of voiced (i.e. quasi-periodic) and unvoiced (i.e. noise like) signals. There are two main models of speech in current use; source-filter model and harmonic plus noise model [4]. In the source filter model, speech is composed of a mixture of a random noise plus harmonic (i.e. periodic) excitation that is input to a linear prediction model of vocal tract. In the harmonic plus noise model, speech is simply

modelled as the sum of a harmonic periodic part (modelled by a Fourier series representation) plus a noise part.

One of the most important speech parameters is the fundamental frequency, perceived by the human auditory system as the pitch level and commonly referred to as the pitch of voice. The fundamental frequency of speech models the rate of repetition of the periodic, voiced part, of speech. It is in fact the rate of the opening and closing of vocal folds during voiced excitation.

The estimation of the fundamental frequency of speech has proved a very challenging problem that continues to be a subject of research interest and the focus of this PhD thesis.

1.2 THE RANGE OF PITCH VARIATIONS IN VOICE

The range of values of fundamental frequency, pitch, varies substantially from very low values of 20 Hz to very high values of above 3000 Hz in young children. Variations and changes of voice pitch can be due to the following reasons:

1) Physiological function of age, gender and speaker characteristics
2) Intended intonation
3) Induced by emotion, style, accent or singing.

Associated with aging, the changes of the fundamental frequency from childhood to adulthood is significant, a child due to smaller and younger vocal organs can have double or more the pitch of an adult [5] - [6]**.** Changes in pitch as an adult person ages are slight increases for males aged 70 years or older. For females there is a decrease in the pitch associated with the aging process which is most obvious when comparing females in the

30s-40s age range with females in their 20s [7]. There may also be some tremor or shakiness in pitch with advanced aging process.

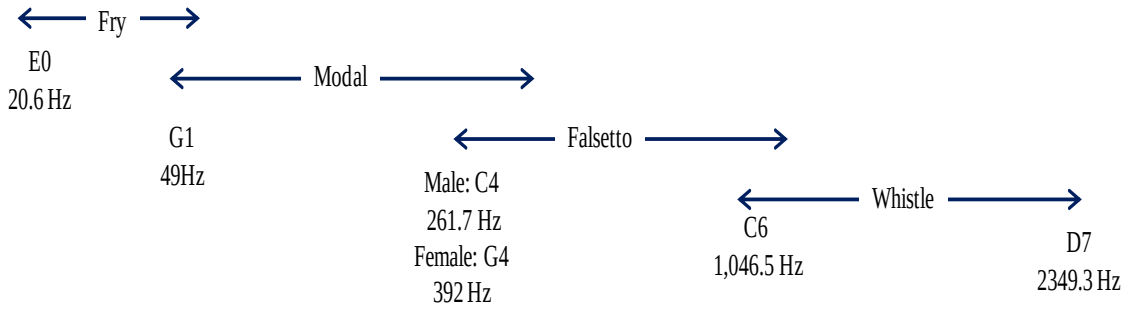


Figure 1.1- General vocal register in speech and singing.

The range of pitch is particularly wide well described for singing voice where the pitch range is categorised into four categories of pitch/vocal registers [8] namely

1) whistle (very high pitch),
2) falsetto (high pitch),
3) modal (normal pitch range) and
4) fry (low pitch) .

Whistle is the highest pitch register mainly used in singing voice or by children. In most singers the whistle begins above the soprano "high C", C6 or 1,046.5 Hz, and extends more than an octave to above D7 or 2349.3 Hz. Young Children can produce high frequency whistle phonation that go beyond G7 or 3136.0 Hz.

Falsetto is the vocal register in the high frequency range just above the modal voice register, overlapping with it by approximately one octave, and below the whistle register. It is produced by the vibration of the ligament edges of the vocal cords.

Modal is the vocal register used normally in speech and singing. It is also the term used in linguistics for the most common phonation of vowels. In linguistics, modal voice is the only phonation found in the vowels and other sonorant (consonants such as *m, n, l,* and *r).*

Vocal fry is the lowest vocal register produced through a loose glottal closure which will permit air to bubble through slowly with a popping or rattling sound of a very low frequency.

Figure 1.1 shows a depiction of the approximate range of the vocal registers from the lowest pitch (Fry) to the highest pitch (Whistle).

1.3 THE IMPORTANCE OF PITCH IN SPEECH COMMUNICATION TECHNOLOGY

Pitch is an important speech parameter for a wide range of applications such as speech coding, text to speech synthesis, speech recognition, gender identification, accent identification, accent synthesis, speaker identification, speaker verification and speech morphing.

In mobile speech coders, such as GSM, the fundamental frequency is estimated at a rate of once every five ms (i.e. 200 times a second) and around 20% of the total coding bit resources are allocated to the quantization of the fundamental frequency parameter [9].

For text to speech synthesis one of the major issues is the reproduction of the correct pitch intonation that fits the intended expression, style and emotion. Note that here; the issue is not the estimation of the fundamental frequency but that of the determination of the correct context dependent time-variation of the fundamental frequency at the phrase and sentence levels.

Pitch intonation is also a major indicator of an accent's characteristics identity, for example in Southern UK English accent, the pitch intonation trajectory falls at the end of a declarative statement and raises at the end of a question. In contrast in some Northern UK English accents the pitch intonation rises at the end of a statement.

The gender identity is carried mostly by the pitch, in that pitch appears to be the most distinctive parametric indicator of the gender. Typical adult males have a mean pitch value of around 120 Hz (with a range of 85 Hz to 180 Hz) and typical adult females have a mean pitch value of around 210 Hz (with a range of 165 Hz to 255 Hz) [10].

Speaker identity and speaking style may be parameterised by the pitch variations, formants and spectral features such as cepstral features. In particular the speaking style is significantly impacted by the variations of the pitch trajectory, i.e. intonation style and habits, as well as the speaking rate [7].

1.4 THE CHALLENGES OF PITCH ESTIMATION

Despite more than 50 years of research and development, the estimation of the fundamental frequency of voiced speech, pitch, remains a challenging and nontrivial problem as there does not exists a closed-form solution, or an error- free method of any form, for calculation of the pitch values and the correct pitch values have to be estimated and tracked from a number of likely candidates obtained at the maxima or minima of a similarity criterion.

The factors that contribute to the challenging nature of pitch estimation may be listed as follows.

1) *The time-varying nature of pitch;* implies that the period, or its inverse the fundamental frequency, estimated from a speech frame is at best the average value of the period, or the fundamental frequency, within the frame. The actual period can vary substantially over a frame or it may oscillate within a frame depending on the emotional state of the voice.

2) *Indeterminate nature of some quasi-periodic speech;* for transient speech segments, in particular at the onset and at the end of a voiced segment, it is difficult, even for an expert, to visually determine the correct pitch value as the period sometimes changes erratically or drastically. In particular, at the end of a phrase/sentence the fundamental frequency and the number of significant harmonics may decrease substantially. Furthermore, sometimes within a voiced segment the signal amplitude and or its harmonic to noise ratio can drop significantly. Many of the most challenging errors, that the author has observed, occur when the speech analysis window contains transient speech with substantial variations of fundamental frequency and spectral content within the analysis window.

3) *Missing fundamental;* the fundamental frequency may coincide with a trough (anti-resonance) of the spectral envelop such that the first observable harmonic is the second or a higher harmonic.

4) *Half and double pitch estimation*; a periodic signal with a period of T exhibits peak correlations at integer multiples of T. This may lead to ‘half pitch’ estimation error, i.e. an estimate that is an octave below the actual pitch value, in cases where the similarity measure at $2T$ is stronger than at T. For some speech segments periodicity is also exhibited at half period leading to ‘double pitch’ estimation, i.e. an octave

above the actual pitch value, in cases where the similarity measure is a stronger at $T/2$ than at *T*. Note in addition there can be large errors which are not necessarily an integer multiple of the actual pitch.

5) *Voicing errors;* for the purpose of pitch estimation, speech is broadly composed of two states; a voiced state with a harmonic structure and an unvoiced state with a noise-like structure. The error in detection of voiced/unvoiced states affects the accuracy of pitch estimate.

6) *Harmonic to noise ratio* (*harmonicity*)*;* generally voiced signals are composed of a mixture of harmonic and noise components. Pitch estimation accuracy improves with the increasing harmonic to noise ratio and degrades as the harmonic to noise ratio drops, for example, for breathy, creaky or hoarse voice [11].

7) *Noise;* pitch estimation, particularly for mobile communication environment, can be affected by background noise, as in all signal estimation methods, the accuracy of pitch estimation drops with increasing background noise or decreasing signal to noise ratio.

8) *Speech disorders and impediment*; speech impediments can complicate pitch estimation and result in increased estimation errors. Examples of speech disorders are: stuttering, apraxia (trouble sequencing the sounds in syllables and words), dysarthria (paralyse of speech muscles), voice disorder impairment, cluttering (abnormal speech delivery rate) and muteness [11] - [12].

1.5 RESEARCH AIMS, OBJECTIVES, AND MOTIVATIONS

The broad aims of this research thesis are:

1) The development of digital signal processing models and methods for accurate estimation of the fundamental frequency, also commonly referred to as the pitch, of voiced speech signals.
2) In particular a main focus of this research work is development of methods that are more robust in that they are less prone to suffer from large pitch estimation errors such as double and half pitch estimation.
3) Comparative evaluation of the proposed pitch estimation methods using actual fundamental frequency as the reference.

In order to achieve the aim of more accurate and robust pitch estimation, the objectives of this research thesis are defined as follows:

1) A critical study of the existing pitch estimation methods in particular regarding to the objective criteria employed and the impact of the choice of the values for such parameters as the window length.
2) Exploring, implementation and evaluation of novel similarity objective criteria for pitch estimation.
3) A study of the methods used for resolving the ambiguity when the objective criteria yield several closely competing possible solutions.
4) Exploring and implementation of novel methods for selection of the best candidate among several proposals from the objective criteria.

5) Comparative evaluation of the proposed pitch extraction methods and a state of the art method using as the reference, or as the 'ground truth', the values of the fundamental frequency extracted from the laryngograph signal.

The motivation for this work is twofold;

1) The central role practical importance of accurate pitch estimates in all speech processing technology applications for coding, synthesis and recognition.
2) The continuing need for the development of pitch estimators that is more accurate and robust particularly with regard to large pitch errors.

1.6 THE THESIS STRUCTURE

Figure 1.2 shows the structure of the thesis whereas shown the research work presented in this thesis is organized in 7 chapters; each chapter starts with a brief introduction of its subject, highlights the main contributions and provides an overview of that chapter. At the end of each chapter, a brief conclusion is presented. The structures of this thesis are as follows:

- **Chapter 1** provides a brief descriptive introduction to the fundamental frequency of speech, perceived as pitch, and explains the challenges in pitch estimation. The chapter sets out the aims, objectives and contribution of the thesis and includes a description of the thesis structure.
- **Chapter 2** describes an overview of the literature review of the established methods for estimation of the fundamental frequency or pitch. First, the general introduction to pitch extraction systems is presented. Then, the pitch extraction are categorised in different approaches such as in time domain, frequency domain, time-frequency

domain, and statistical model based approach and examples of their working principles applied to fundamental frequency or pitch of the speech segment are presented.

- **Chapter 3** presents the signal processing methods for production and synthesis of the human speech excitation signal. The physiology of human speech production model where the mechanism of human speech system and the role of the various speech organs involved are described. Then, the source-filter model of speech production and the method of synthesis of speech production including linear prediction filter model and harmonic plus noise ratio model are explained.
- **Chapter 4** provides the statistical theory of the variation of pitch curve and includes the smoothing process. Firstly, the general theory of the method of the pitch trajectory of the speech signal such as finite-state Markovian-Gaussian and linear prediction models are introduced. Then algorithms that would limit the errors regarding the smoothness and the continuity of pitch trajectory are introduced.
- **Chapter 5** presents the detailed development of proposal pitch estimation method in time domain using the concept of modified higher order moments. Firstly, the general introduction of pitch estimation and the impact of varying segment or frame length of speech signal to the pitch estimation are described and evaluated. Then, the theoretical concept of the higher order moment (HOM) of the pitch estimation method is presented. The evaluation of the proposed pitch extraction method is performed by measuring the mean of percentage errors for three categories of errors: an average error, the fine error (less than 20%) and the gross error (more than 20%). The population errors for fine and gross errors are also calculated. Finally, the analysis from the obtained evaluation results is discussed and conclusion presented.

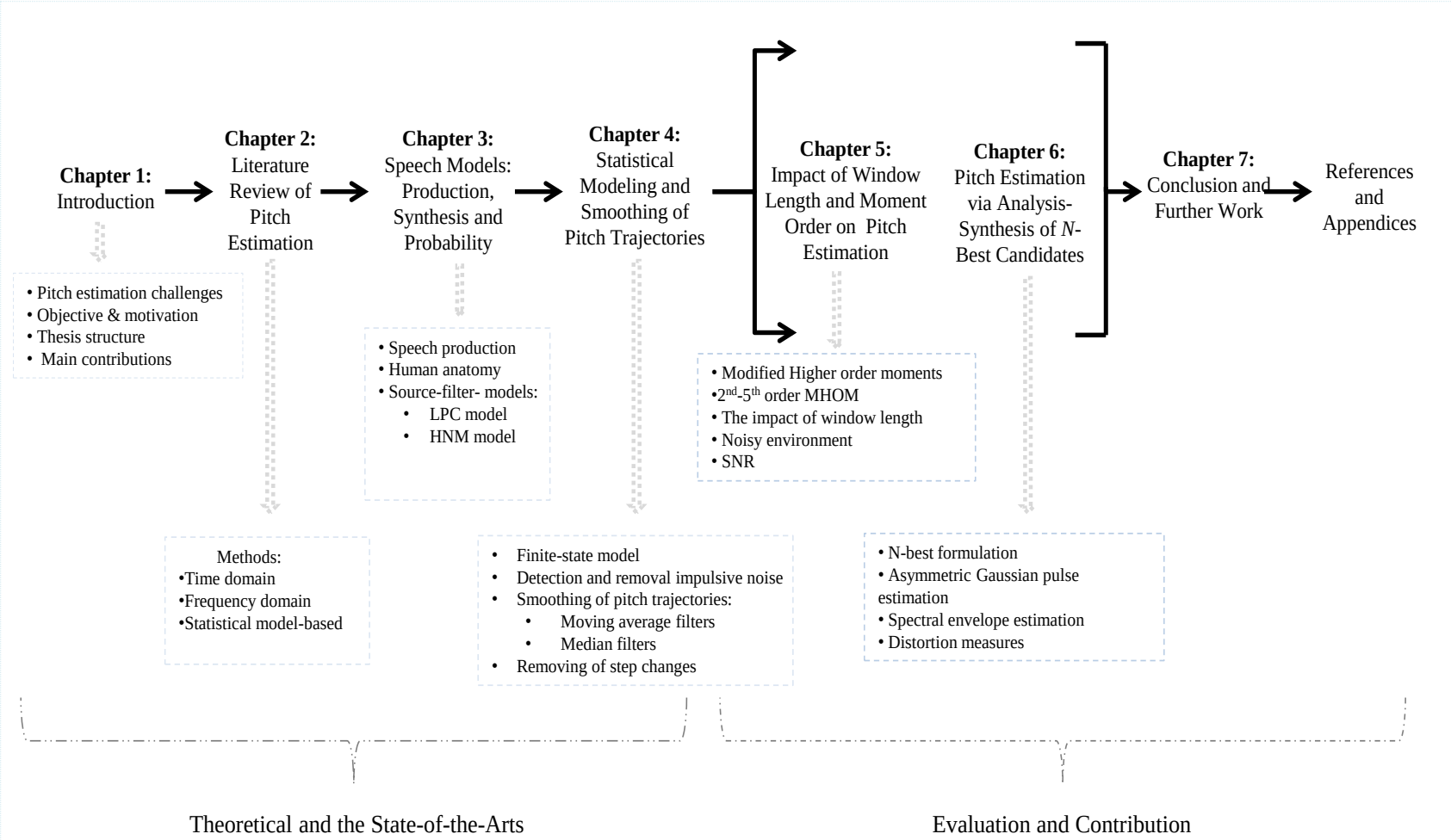


Figure 1.2 - The Structure of the thesis.

- **Chapter 6** describes the detail of the determination of the best pitch value given a selection of N-best pitch candidates. In this work, the N-best pitch candidates are obtained from the positions of the N extrema of the similarity criterion. The determination of which candidate is the best one, is based on the fidelity by which the harmonic part of the speech spectrum can be synthesised using the candidates, i.e. the candidate that provides for the best reconstruction of harmonic part of speech spectrum is selected as the best pitch. The speech synthesis is performed in the frequency domain using an analysis-synthesis method based on a harmonic plus noise model of speech. The synthesis of speech involved extraction of the spectral envelop and synthesis of a periodic excitation composed of asymmetric Gaussian pulses positioned at the harmonic frequency series determined by the pitch candidate. The selection of the best pitch candidate is based on evaluation of the distortion of the speech synthesised using each pitch candidate and as such a distortion measure that rewards for good spectral match and penalises spectral mismatch is required. The different distortion measure approaches evaluated are signal-to-noise ratio (SNR), minimum mean squared error (MMSE), and the mean Harmonicity (MH). Lastly the obtained results are analysed and conclude the evaluation.
- **Chapter 7** contains the summary of the research work, the conclusions, and the suggestions for the further work.
- **References** and **citation.**

1.7 THE MAIN CONTRIBUTIONS OF THE RESEARCH

In this thesis significant issues concerning the estimation of the fundamental frequency or pitch have been considered and the proposed solutions evaluated. The main contributions of this research work are summarised as follows.

1) A modified higher order moments (MHOM) are proposed and evaluated as alternatives to the conventional second order moment model. The MHOM version proposed in this work includes a new method of calculation of the MHOM and the evaluation of the third order, fourth order and fifth order criteria. The MHOM criteria compare favourably and often outperform the conventional methods (published in ICASSP2010 and INTERSPEECH2011).

2) A novel N-best strategy for determination and selection of the best pitch value given N pitch candidates, the N pitch candidates are the top N extrema of a similarity moment criteria. For each proposal, the harmonic part of the signal spectrum is synthesised and subtracted from the original spectrum to determine the proposal that can best reproduce the harmonics content of speech.

3) Spectral signal synthesis methods using a combination of a new spectral envelop estimation method and a novel asymmetric Gaussian model of excitations. The spectral synthesis is employed in N-best pitch estimation for determination of the pitch candidate that can facilitate the best synthesis the original speech spectrum.

4) A set of t spectral distortion measures are proposed to determine the similarity between an original signal and a synthesised signal version. Several different criteria were considered including combinations of weighted SNR, weighted MSE and

harmonicity, the latter is a measure of harmonic strength of each harmonic partial of the signal spectrum.

5) Cost functions are proposed that combine different spectral distortions and apply a cost penalty for impulsive and step changes in estimation trajectory of pitch.

2

A LITERATURE REVIEW OF PITCH ESTIMATION METHODS

This chapter provides a literature review of the speech pitch extraction methods, describes the major advances made, the current challenges and the state of the art of pitch extraction methods. The literature review is described in terms of the main categories of methodologies. The descriptions of the methods are arranged in each class the major contributions are described in the chronological order. The methods are compared in terms of their complexity and performance.

2.1 INTRODUCTION

The aim of pitch estimation (aka pitch extraction or pitch detection) algorithms is to detect and measure the time-varying period of repetition, or equivalently the fundamental frequency, of voiced speech.

Pitch can be defined as the auditory sensation of the fundamental frequency, F_0 of a periodic audio signal; whereas the fundamental frequency of a periodic signal is a numerical quantity that may be accurately measured and assigned a value by an electronic instrument, pitch is the perception of the 'frequency level 'or the 'frequency tone' of a signal by a human audio sensory system [4], [13] - [14].

The fundamental frequency, F_0, of a periodic signal defined as the number of repetitions or cycles per second, is the inverse of the duration or the period of one cycle T_0 and is defined in units of Hz as

$$F_0 = \frac{1}{T_0} \quad Hz \tag{2.1}$$

Pitch is a significant parameter in speech, music and generally audio signal processing systems because a significant proportion of audio signals are often composed of quasi-periodic components, examples are voiced signals or music signals generated by string or brass music instruments.

Reliability and high accuracy in estimation of pitch, from the raw speech signals, are essential and necessary for accurate and/or high quality output in most speech communication application including in speech recognition, speech synthesis, and speech coding [15].

Generally, speech signals are combination of a quasi-periodic voiced (harmonic) and a non-periodic unvoiced (coloured noised) signals and silence segments [16]. The term quasi-periodic implies the signal is seemingly, but not strictly, periodic because the period varies over time. The characteristic features of the quasi-periodic voiced segments is a harmonic spectral structure, a relatively higher overall energy compared to unvoiced signal and a greater concentration of the spectral energy in the lower frequency part of the spectrum (less than 2 kHz). In contrast, unvoiced signals are random/aperiodic signals, have a lower overall energy and most of their spectral energy is concentrated at higher frequencies (above 2 kHz) [17] - [18].

The harmonic plus noise model (HNM) of speech may be expressed as

$$x(m) = \sum_{k=1}^{N_h} a_k(m) cos(2\pi k F_0(m) m + \varphi(m)) + v(m) \qquad (2.2)$$

where $F_0(m) = 1/T_0(m)$ is the time-varying fundamental frequency at discrete-time m, $a_k(m)$and $\varphi_k(m)$ are the time-varying amplitude and phase of the k^{th} harmonic of the signal respectively, N_h is the number of harmonics up to the bandwidth and $v(m)$ is the non-periodic (noise) component of the signal [19]- [20]. Note that the harmonic part of Equation (2.2) is essentially a Fourier series expansion of the periodic voice signal.

The pattern of time-variation of the pitch, $F_0(m)$, known as intonation, conveys such information as pragmatics of speech, intent, style, emotion and accent [4],[13]. In English language pitch does not affect the word identity, however in tonal languages, such as Chinese and some African languages, the word identity can change with the pitch

intonation; these changes can be subtle presenting particular challenges to automatic speech recognition for tonal languages [21].

Although, over more than 50 years, numerous papers and ideas in pitch extraction methods have been published, with several significant contributions, nevertheless pitch estimation remains a nontrivial problem that continues to be a research challenge in development of speech processing systems contributions.

Reliable and accurate pitch period estimation is the most important objective in pitch estimation methods in all pitch extraction approaches. There are several different approaches in pitch period estimation which will be reviewed in this section.

2.2 AN OVERVIEW OF PITCH ESTIMATION METHODS

Figure 2.1 shows a categorisation of pitch estimation methods into four broad categories of time domain, frequency domain, time-frequency and statistical methods. Within each category, several prominent methods are described. Note that each of the first three methods, namely time, frequency and time-frequency can be combined with and expressed in terms of a statistical probabilistic framework. Note also that most models/methods, such as moments, harmonic plus noise model, linear prediction model etc. can be described alternatively in time or in frequency domains.

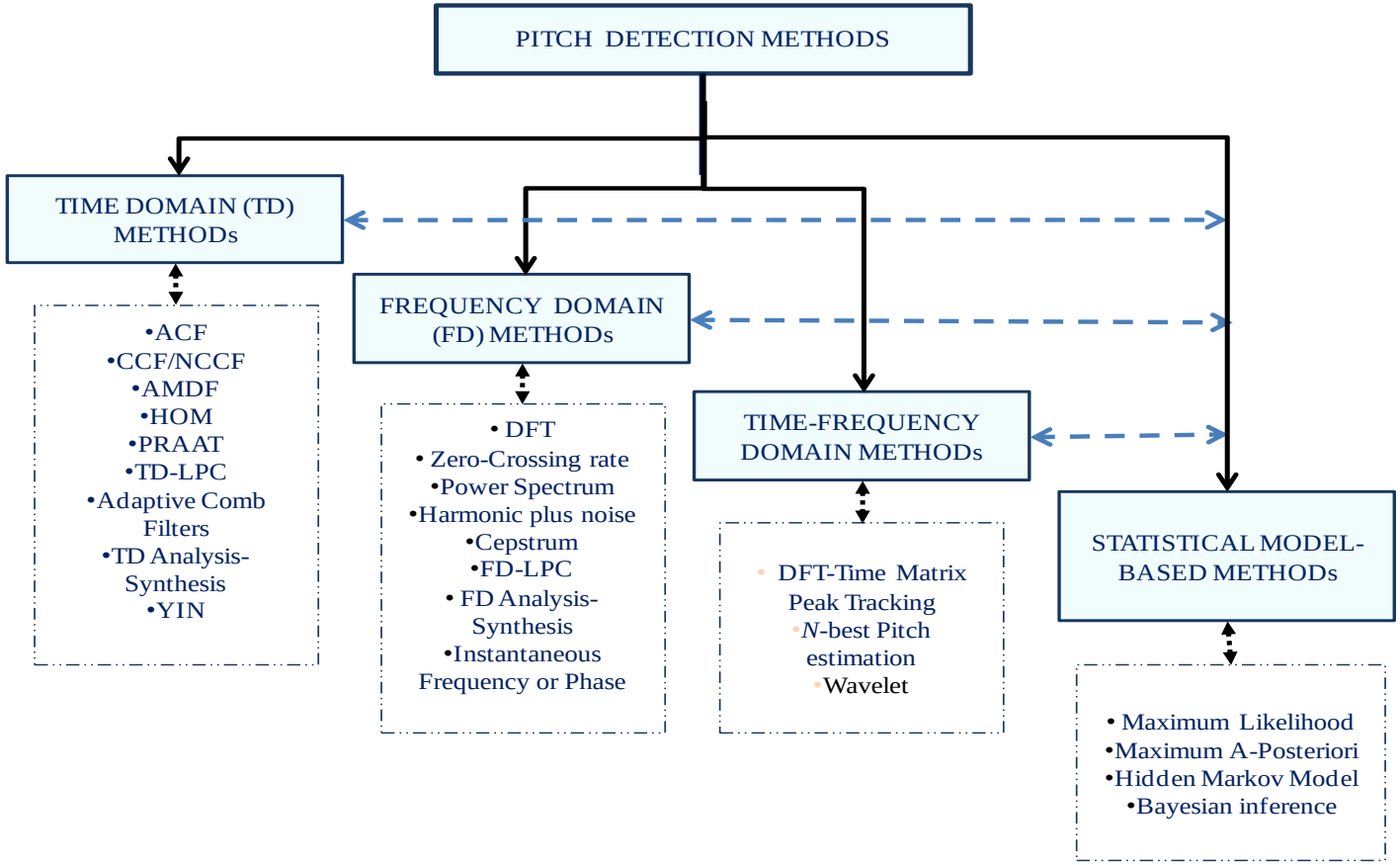


Figure 2.1- Illustration of the categories of the pitch estimation.

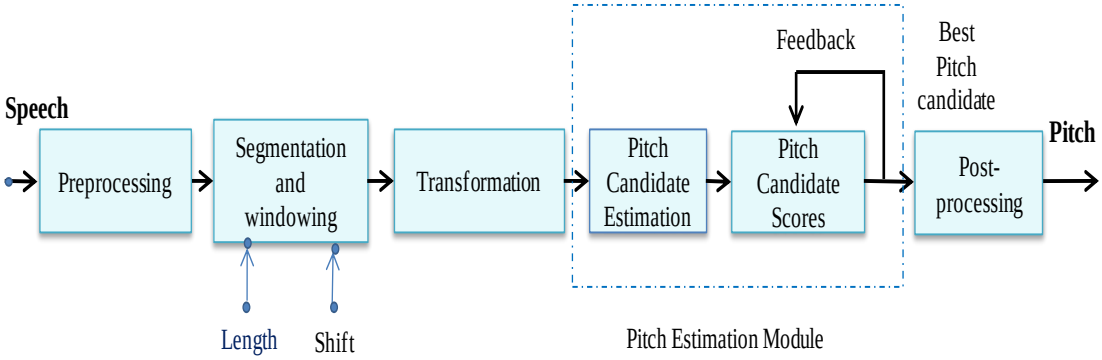


Figure 2.2 - The generic block diagram of pitch estimation systems.

Figure 2.2 shows the generic block diagram of a pitch estimation signal processing system. As shown most pitch estimation algorithms may have about six stages of signal processing:

1) Pre-processing or signal conditioning.
2) Segmentation and windowing.
3) Signal transformation.
4) Estimation of pitch candidates (core module).
5) Selection of the best pitch candidate.
6) Post-processing.

2.2.1 Pre-Processing of Speech Signal

The pre-processing or signal conditioning stage removes the signal components that are unwanted or detrimental for the purpose of pitch estimation [22]- [23], such as;

1) The mean/dc signal values are removed by a simple mean subtraction process; this may be done for the whole signal or on a segment by segment basis. This may also be done as part of bandpass filter.
2) Very low frequencies that are below the minimum expected value of the pitch (typically less than 60 Hz), these components are removed by a lowpass filter with a cutoff frequency of 40 Hz.
3) Higher frequencies with low harmonic to non-harmonic ratio (i.e. low harmonicity), typically above 2000 Hz, are removed by a high pass filter, note a band pass filter may be used to simultaneously remove the unwanted low and high frequency parts of the signal.
4) The effect of vocal tract resonances at formants may be removed by an inverse linear prediction filter.
5) Noise and interference can be reduced by a noise reduction method; care must be taken such that the noise reduction process does not have an adverse impact on harmonic structure of the signal [24].

2.2.2 Segmentation and Windowing

For a non-stationary signal such as speech, with a time-varying spectral composition and time-varying fundamental frequency, a short-term similarity measurement (or other pitch estimation function) capable of tracking the trajectory of speech parameters is desirable. For this reason, pitch estimation measures are applied to relatively short segments (in the order of several 10's of ms) of windowed speech segments. The segmentation and windowing are described here.

1) ***Segmentation*** or dividing the speech signal into frames is an integral part of all digital speech processing systems. Note the two words, *segment* and *frame*, are used interchangeably [4], [25]. The segment/frame length is constrained by: (a) the non-stationary character of speech and (b) the maximum allowable delay in voice communication systems. The International Telecommunication Unit-T (ITU-T) allows the delay in voice communication systems is 300ms under G.114 [9]**.**

2) Hence in mobile communication systems the standard segment length may have significant time variations. Hence in mobile phones the pitch values is set to a value of 20 ms. However, even within a short segment of 20 ms the pitch values are actually updated four times per frame equivalent to a pitch update rate of once every 5 ms (or 200 Hz). Note the actual window used for pitch estimation in mobile phones spans one and half window duration (equivalent to 30 ms), however, the centre of the window is shifted every 5 ms.

3) The choice of the segment length is an important issue in pitch estimation; in general for calculation of the similarity criteria a segment should contain at least two or three pitch periods. If the segment is too short, the variance of the similarity criteria will be large and the pitch estimation method will not be able to estimate the pitch period accurately, likewise, if the segment is too long, then the pitch method will not be able to detect the non-stationary variation in the length of the pitch period from period to period [26] - [27].

4) *Windowing* is commonly applied in speech processing. It can have two benefits: (i) reduce the end-point discontinuity and the consequent spectral leakage and (ii) shape the signal envelop such that it places more emphasis on a particular part (or subframe) in the current frame. This latter property is useful when the current frame is concatenated with one or more previous frame for the purpose of pitch

estimation as is the practice in mobile phone speech processing when one and a half frame length is used. One can then shape the window such that the current frame has a greater relative weight than the previous frames concatenated to it. Examples of popular windows used in speech processing include Hamming and Hanning window functions [17], [28].

2.2.3 Signal Transformation

Pitch estimation can be achieved in the time domain, in the frequency domain, in the time-frequency domain or in the frequency-scale domain such as wavelets. Hence, the transformation module may be Discrete Fourier Transform (DFT) [25], [29], a wavelet transforms [30] - [31] or for the case of time domain pitch extraction the transformation will be an identity matrix.

The most commonly applied pitch estimation methods are based on time domain using the correlation function as the similarity criterion. However, as correlation and power spectrum (or squared magnitude spectrum) are Fourier transform pairs, one can apply correlation-based pitch estimation method in the frequency domain using the power spectrum as the similarity criterion [32] - [33].

In this thesis a time domain approach to pitch estimation using higher order moment methods is described in Chapter 5. In Chapter 6 a frequency domain approach is used to select the best pitch estimate among N candidates.

2.2.4 Pitch Estimation Module: Estimation of the Pitch Candidates

The different similarity criteria used for pitch estimation have a common feature: they all measure the turning points at which the extrema of similarity reinforcements (e.g. as with the correlation criterion) or similarity cancellations (e.g. as with the magnitude difference function criterion) occur. Usually, for periodic signals, the similarity criteria have a number of significant extrema points and hence yield more than one likely pitch candidate.

As explained above depending on the method employed, pitch estimation may be categorized into four distinct approaches:

1) Time-domain pitch extraction methods often employ a moment-based similarity criterion, these include; correlation higher order moments (HOM) and magnitude difference; the latter includes the benchmark YIN method [34] as an implementation. The most commonly used similarity measurements include the peaks of the moments or the troughs of the differences of the signal as a function of the proposed period T. For example, correlation-based pitch extraction methods estimate the period as the value of T for which the average of the product of $x(m)x(m-T)$ over a frame of speech samples, known as the short-time correlation, attains a maximum value. Magnitude-difference-based pitch extraction methods estimate the period as the value of T for which the average magnitude difference $|x(m)-x(m-T)|$ over a frame of speech samples attains a minimum [16], [23], [27], [34] - [35].
2) Frequency domain pitch extraction methods first transform speech segments into the frequency domain using a discrete Fourier transform or some variant of it. Frequency domain methods are based on the observation that the energy/power of

periodic signals is concentrated in a set of narrow bands of frequencies around the fundamental frequency $F_0 = 1/T_0$ and its integer multiples kF_0; the harmonics, The fundamental frequency is obtained by detecting a set of spectral peaks and then, from this set, processing the frequency positions of the harmonically related components of the signal. Frequency-transformation methods include, magnitude/power spectrum-based methods cepstrum, zero-crossing and instantaneous frequency method [13], [25] - [26], [36]- [37].

3) Time-Frequency methods employ an expansion of the speech signal in time and frequency domains. This may take the form of a DFT-based spectrogram matrix of the speech signal combined with a peak tracking algorithm that tracks the time-varying positions of the fundamental frequency of speech and its harmonics. Wavelets may also be used as a pitch extraction method employing frequency and scale [38].

4) Model-based methods which may include the use of both: (a) a generative method such as an autoregressive moving average (ARMA) filter [39] - [40], a linear prediction model or a harmonic plus noise model of speech and (b) a statistical, probability, model of the speech and pitch signals [41] - [42].

A model-based comb filter, using an adaptive ARMA structure, with poles and zeros, can track the time variation of the fundamental frequency of a periodic signal. The criterion used is the minimisation of the energy (MMSE) of the filter output.

In Chapter 6 a model-based method using linear prediction and a harmonics model is described for synthesis of periodic part of speech.

2.2.5 Pitch Estimation Module: Selection of the Best Pitch Candidate

The selection of the best pitch value from a set of likely candidates is achieved by the use of scoring algorithm which calculates the least costly choice among the proposed candidates output by the pitch similarity function.

Over time, the use of N-best candidate gives rise to a trellis of pitch trajectories, with different costs associated with different path in the trellis. The most common method of estimation of the best pitch is the computation of the most likely, or the minimum error, trajectory using a dynamic programming method. The Viterbi algorithm is often used to obtain the minimum cost path [43].

A new method described in Chapter 6 directly synthesises a harmonic signal for each pitch candidate. The best candidate is obtained as the one that produces the best reconstruction of the original signal.

2.2.6 Post-processing of Pitch Trajectory Estimates

After selection of the best pitch candidate, a number of techniques may be used to improve the pitch estimates including fine-tuning estimation, impulsive noise removal filters and trajectory smoothing filters.

1) *Fine-tuning*, in some application after the calculation of the initial pitch value from a window length of L samples, a constrained recalculation/refining of the pitch value over a shorter window length, and within a relatively small region centred about the initial pitch estimate, may be performed in order to increase the time-resolution of pitch estimation [34], [44].

2) Impulsive noise removal is achieved by one a of a number of alternatives, (a) median filters, (b) removal of abnormally large differences in the residuals of a linear prediction (LP) filter, this involves processing of the residues of low-order LP model of p-itch tracks, (c) detection of sudden changes, samples corrupted by an impulse are removed and interpolated from the past values [24], [45].

3) Smoothing of random fluctuations of pitch estimates; constitutes the use of a moving average, low pass, filter for smoothing of the pitch trajectory [13], [17], [24], [34].

2.3 PITCH EXTRACTION METHODS

2.3.1 Moment Based Pitch Estimation

Pitch extraction methods based on the autocorrelation function (ACF) are the most well-known moment-based method for estimation of the period of an audio signal. Other moment based methods considered in this section include cross-correlation function (CCF), average magnitude difference function (AMDF) and higher order moment methods (HOMs).

2.3.1.1 Autocorrelation Method (ACF)

Autocorrelation based method is the method most commonly used to estimate period, T_0 or its inverse the fundamental frequency F_0 of speech signal. The autocorrelation function is taken as a mathematical measure of the similarity of a signal with itself as a function of time lag *T*. The autocorrelation function (ACF) of *N* samples of a signal $x(m)$, for a time lag of *T* samples is defined as

$$r_{xx}(T) = \frac{1}{N}\sum_{m=0}^{N-1} x(m)x(m-T) \tag{2.3}$$

A normalised version of autocorrelation as showed in Figure 2.3 may be defined as

$$r_{xx}(T) = \frac{1}{N\sigma_x^2}\sum_{m=0}^{N-1} x(m)x(m-T) \tag{2.4}$$

where the signal variance σ_x^2 is the normalization factor of the autocorrelation function.

The signal period T_0 may be estimated as the value of the lag T corresponding to the maximum of the ACF in the range T_{min} to T_{max}

$$T_0 = \arg\max_T r_{xx}(T) \qquad T_{min} < T < T_{max} \tag{2.5}$$

where T_{min} and T_{max} are the expected minimum and maximum values of the period [16] - [17],[23],[35],[46]-[48]. Since the ACF of a periodic signal is itself periodic, a time-domain energy maximising function that utilises the periodic energy peaks of the ACF is defined as

$$E(T) = \frac{1}{N_T}\sum_{k=1}^{N_T} r_{xx}(kT) \qquad T_{min} < T < T_{max} \tag{2.6}$$

where $N_T = fix(N/T)$ is the maximum number of periods T that can be fitted within the N samples length of a speech frame [49]. The estimate of the period T_0 is obtained as

$$T_0 = \arg\max_T E(T) \qquad T_{min} < T < T_{max} \tag{2.7}$$

In autocorrelation pitch extraction method, the perception of pitch is strongly related to the periodicity in the waveform in the time domain.

Some pre-processing methods have been combined with the autocorrelation method including centre-clipping technique and pre-whitening to flatten the spectrum of the speech signal [16], [48]**, [**50**] - [**51**].**

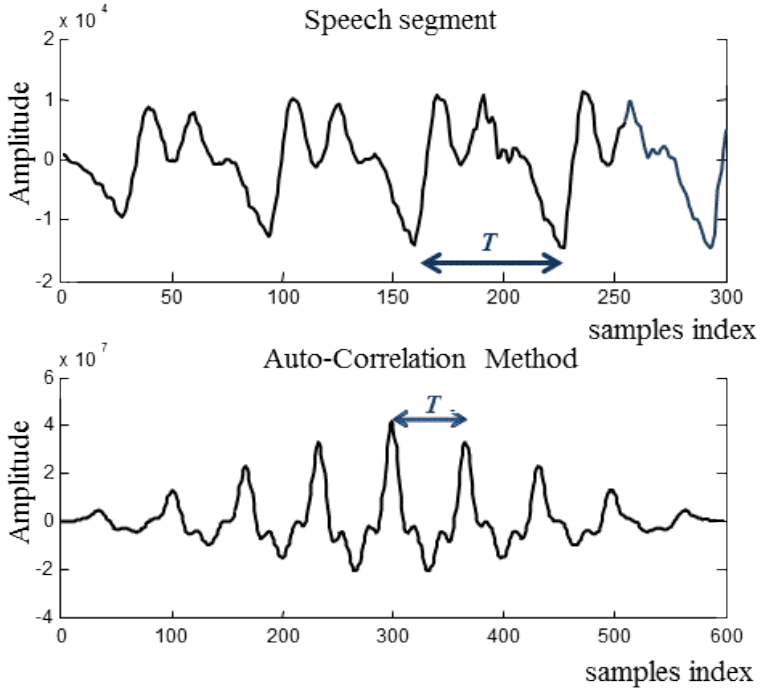


Figure 2.3 - Illustration of normalized autocorrelation method, ACF.

2.3.1.2 Normalized Cross-Correlation Function (NCCF) Method

Whereas the autocorrelation method finds the correlation of the samples, $x(m)$ and $x(m-k)$, of the same speech segment, the cross-correlation methods obtains the correlation between samples of different speech segments (or frames) [27], [52]

$$r_{x_i x_j}(k) = \frac{1}{N\sigma_x^2} \sum_{m=0}^{N-1} x_i(m) x_j(m-k) \tag{2.8}$$

where $x_i(m)$ and $x_j(m-k)$ are samples of different segments of speech as shown in Figure 2.4. The different speech segments may be overlapping successive segments or one segment, e.g. x_j may be a sub-segment of another, e.g. x_i.

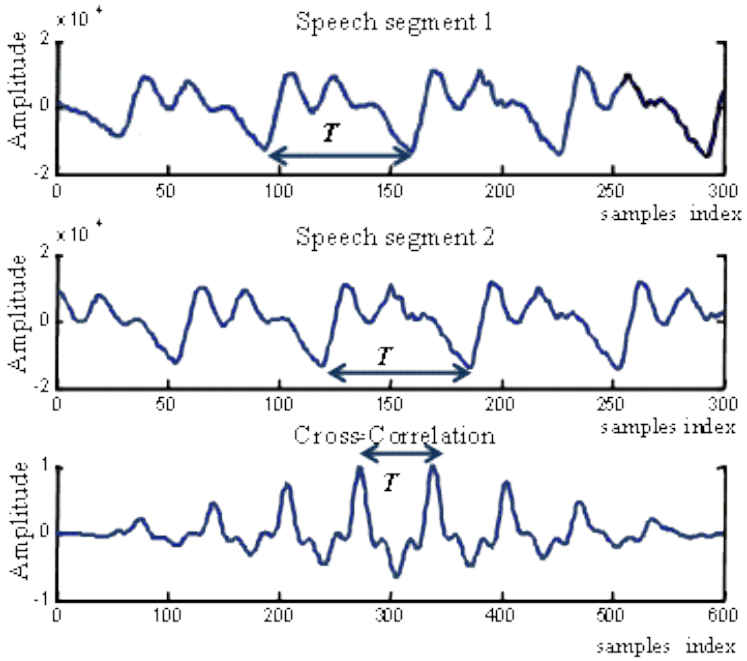


Figure 2.4 - Illustration of Normalized Cross-correlation method, NCCF.

Note in this method the period is obtained as the difference between the positions of the two most prominent peaks. This method is similar to autocorrelation function, it is claimed to follow the rapid variation in pitch and the amplitude of the speech signal. The NCCF overcomes most of the shortcoming of autocorrelation based algorithm at a slight increase in computation complexity. NCCF is better suited for pitch detection than the normal autocorrelation where the peak corresponding to the pitch period are more prominent and less affected by the rapid variations in the signal amplitude [23],[52] - [54].

2.3.1.3 Average Magnitude Difference Function (AMDF) Method

The Average Magnitude Difference Function (AMDF), like the autocorrelation function, measures the degree of similarity between two signals as shown in Figure 2.5 [38], [55]-[56]. The general form of the AMDF criterion for pitch extraction may be defined as

$$d(T) = \frac{1}{N}\sum_{m=0}^{N-1} |x(m) - x(m-T)|^{\alpha} \tag{2.9}$$

where for $\alpha = 1$ we have the AMDF function and for □=2 we have the squared magnitude difference function (SMDF) [34]. For a periodic signal, the AMDF/SMDF attain a minimum at the period T and its integer multiples, kT, when $x(m)$ has a value similar to $x(m + kT)$.

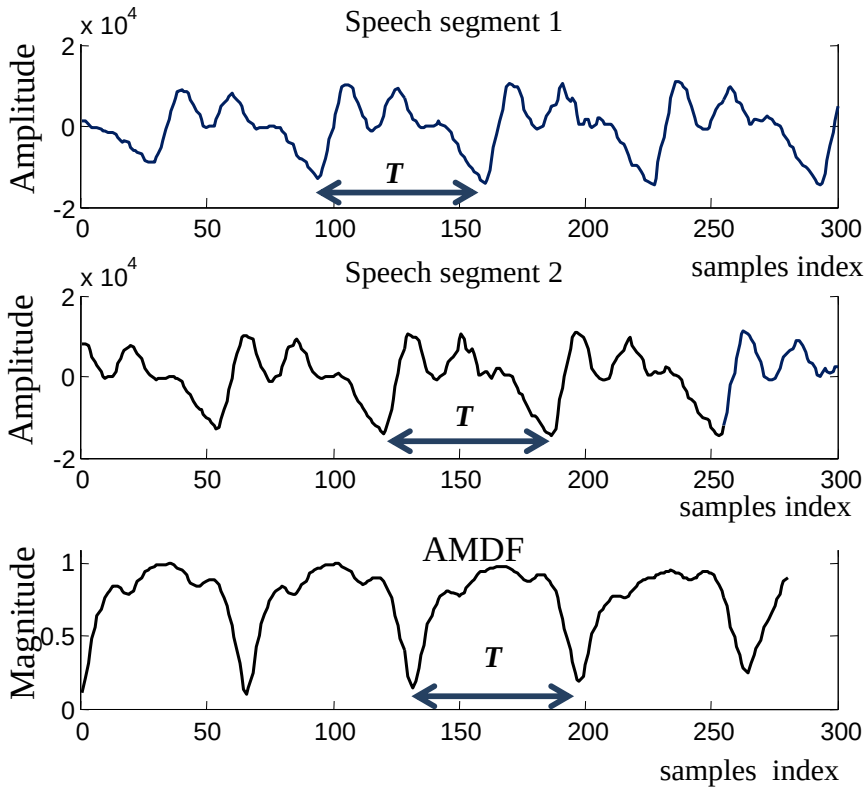


Figure 2.5 - Illustration of Normalized Average Magnitude Difference Function (NAMDF).

A combination of AMDF and ACF pitch estimation methods has been introduced to take the advantage of the AMDF and ACF complementary natures [34], [46], [57].

2.3.1.4 YIN Method

Among the recent and established pitch estimation method in speech and music signals in time domain, one can refer to, is the YIN approach [34]. The YIN method is introduced by A. de Cheveigne and H. Kawahara in 2002, which uses an expansion of the Squared Magnitude Difference Function (SMDF) criteria, Equation (2.9) with □=2, as

$$d(T) = r_{xx}(0) + r_{xx,T}(0) - 2r_{xx}(T) \tag{2.10}$$

where $r_{xx}(0)$ and $r_{xx,T}(0)$ are time-varying autocorrelations at lag zero, calculated at times zero, and T , respectively and $r_{xx}(T)$ is autocorrelation at lag T. This technique yields better result than the autocorrelation function method; it is less sensitive to changes in signal amplitudes, being thus less prone to F_0 estimation error. About 80% decreases in pitch error is reported when the SMDF criteria of Equation (2.9) is used instead of the conventional ACF criteria of Equation (2.3).

The YIN method includes several signal processing steps namely the autocorrelation method (ACF), the difference function, the cumulative mean normalized difference function, the absolute threshold function, the parabolic interpolation and the best local estimate. Consequently, the YIN method is named based on the oriental of the yin-yang philosophical principal of balance between autocorrelation and cancellation in the algorithm.

2.3.1.5 Higher Order Moments Methods (HOM)

Correlation (2nd order moment) methods utilise the similarity between two samples, e.g. $x(m)$ and $x(m-T)$, the higher order methods exploit the similarity between three (e.g. $x(m)$, $x(m-T)$, $x(m-2T)$) or more samples. For pitch extraction, where the intent is to estimate the period *T*, the general expression for the K^{th} order moment can be defined as

$$C_K(T) = \frac{1}{N-(K-1)T} \sum_{m=0}^{N-(K-1)T} [x(m)x(m-T) \dots x(m-(K-1)T)] \qquad (2.11)$$

where $K = 2, 3$ ....

The primary advantage of the higher order moments methods is a greater reinforcement obtained from the product of a higher number of similar samples; $x(m)x(m-T) \dots x(m-(K-1)T)$ as shown in Figure 2.6. The main disadvantage is a larger window length required to average samples that are $2T$ or more apart. The contributions to higher order method of pitch extraction found by the authors include a number of conference papers: [58] - [59].

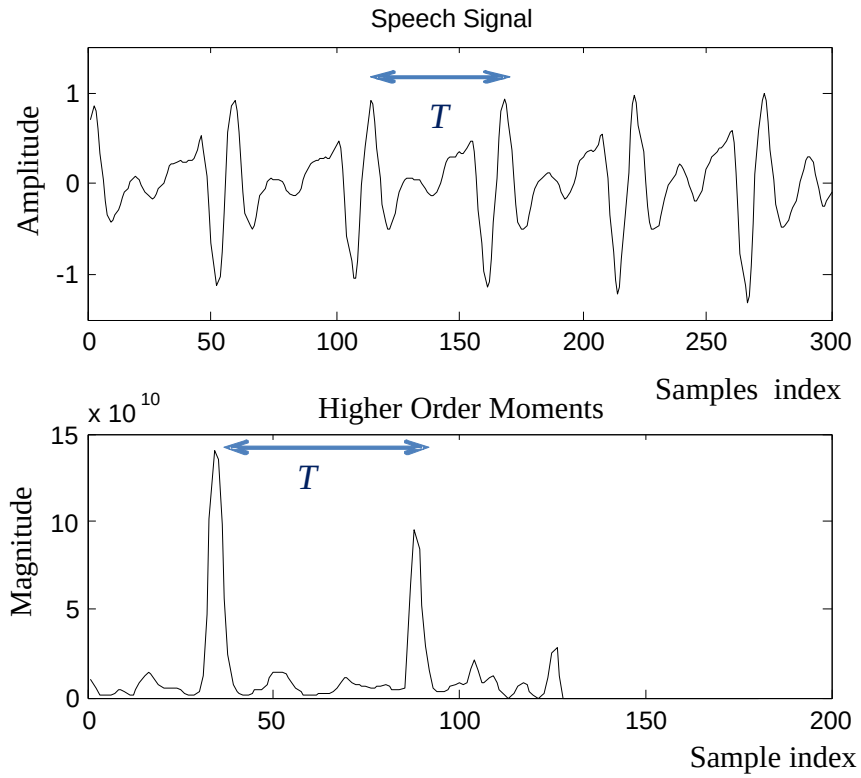


Figure 2.6 - Illustration of higher order moments function (HOM).

2.3.2 Frequency-Domain Transformation Methods

Frequency domain pitch analysis is an alternative approach for estimation of the fundamental frequency of speech. These methods are based on the observation that the energy of a periodic signal is concentrated in narrow bands of frequencies at the fundamental frequency $F_0 = 1/T_0$ and at its harmonics kF_0. Hence the signal is first transformed into the frequency domain and then one of several strategies can be used to estimate the value of F_0 which a cumulative function of the signal energy at harmonics attains its extrema. Like time domain methods frequency domain methods can suffer from problems of half and double pitch estimation.

2.3.2.1 Zero–Crossing Function (ZCF) Method

We have classified zero crossing as a frequency domain method because it gives a direct estimate of the fundamental frequency. A simple early approach for estimation of the fundamental frequency F_0, or its inverse, the period T_0, is measuring the distance between the zero crossing points of the signal, the inverse of which is the zero-crossing rate (ZCR) [13,25], [60] - [61]. Zero Crossing rate (ZCR) is a measurement of how often the waveform crosses zeros per unit time.

Zero-crossing occurs in a speech signal every time the waveform crosses the time axis and provides information about the spectral content of the waveform. Each cycle of a sinusoid signal has two zero-crossing per period as shown in Figure 2.7, therefore the long-time average rate of zero-crossings can be related to the fundamental frequency as

$$Z = 2/T_0 = 2F_0 \tag{2.12}$$

Hence from an estimate of zero-crossings the fundamental frequency can be obtained as

$$F_0 = {}^{Z}/_{2} \tag{2.13}$$

Since speech signals are broadband signals, the pitch estimation can be obtained using representation based on the short-time average zero-crossing rate, which has same properties as the short-time energy and the short-time average magnitude. This method provides an intuitive approach in estimating the frequency of a sine wave. This approach is accurate for narrowband signals which may not have higher harmonics.

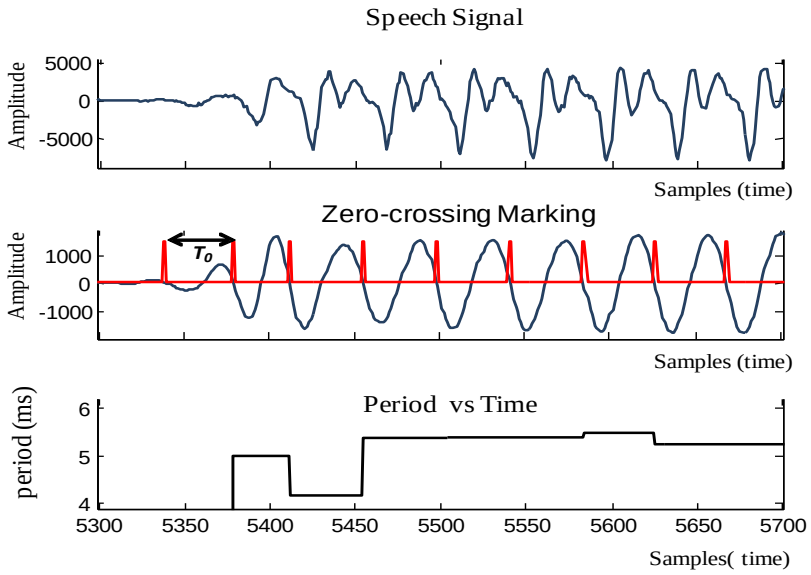


Figure 2.7 - Illustration of zero-crossing functions (ZSF).

The advantage of this method is in its simplicity and computational efficiency; however their main drawback is that for most waveforms do not have only one pair of zero crossing per cycle due to the possible existence of strong second or higher harmonics.

2.3.2.2 Magnitude (power) Spectrum Method

Since autocorrelation and squared magnitude spectrum of a signal are discrete Fourier transform (DFT) pairs, a frequency-domain form of Equation (2.6), a frequency-domain energy maximising function that utilises the uniformly spaced (with spacing of fundamental frequency) harmonic energy peaks of the power spectrum can be defined as

$$E(F) = \frac{1}{N_F}\sum_{k=1}^{N_F}|X(kF)|^2 \qquad F_{min} < F < F_{max} \tag{2.14}$$

where $N_F = fix(N/F)$ is the maximum number of harmonics of a proposed fundamental frequency F that can be fitted within the $N/2$ frequency bins of the DFT of

speech frame of length *N* samples [49]. The estimate of the fundamental frequency F_0 is obtained as shown in Figure 2.8 and may be defined as

$$F_0 = \arg\max_F E(F) \qquad F_{min} < F < F_{max} \tag{2.15}$$

The research contributions on magnitude and power spectrum include the following works: [26], [28, 38], [62] - [64].

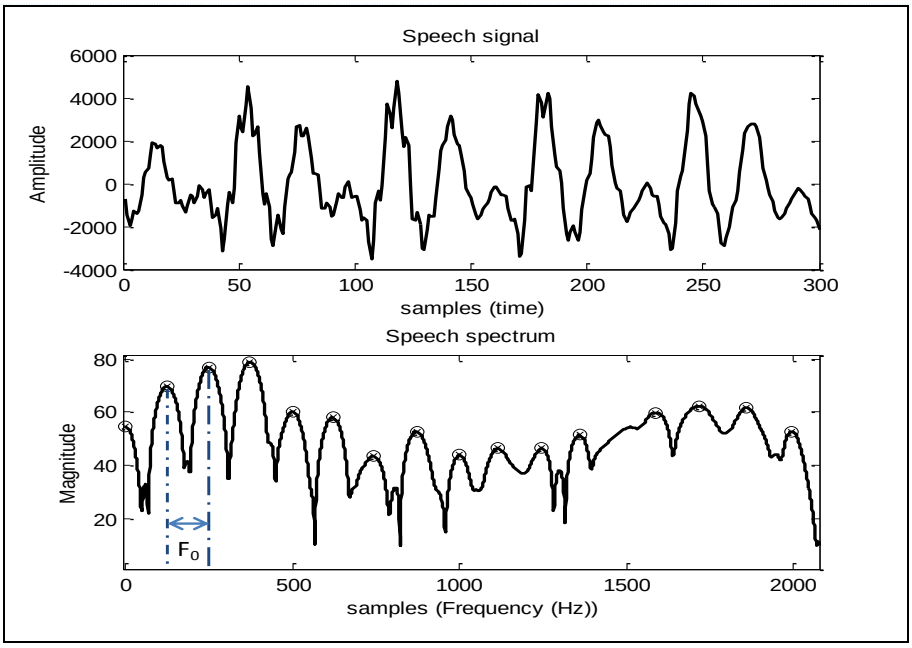


Figure 2.8 - Illustration of pitch estimation based on peak-picking of the frequency spectrum of a signal.

2.3.2.3 Cepstrum Method

Cepstrum analysis is based on the Fourier transform of the log of the magnitude spectrum of a signal. A speech signal $x(t)$ may be modelled as the convolution of a vocal tract $v(t)$ and a glottal input *e*(*t*); $x(t) = e(t) * v(t)$. The cepstrum of $x(t)$ may be defined as the cosine transform of its log magnitude spectrum as

$$C(t) = \int_0^{\infty} log\,|X(f)|^2 cos(2\pi f t)df$$

$$= \int_0^{\infty} log\,|V(f)|^2 cos(2\pi f t)df + \int_0^{\infty} log\,|E(f)|^2 cos(2\pi f t)df \qquad (2.16)$$

where $X(f)$, $V(f)$ and $E(f)$ are the spectra of speech, vocal tract and excitation signals respectively. The cepstrum index is known as the quefrency (an anagram of frequency) as depicted in Figure 2.9. In cepstrum method the vocal tract is mainly confined to the lower quefrency indices whereas the pitch value appears as a distinct peak at a higher quefrency relative to that of the vocal tract [36] - [37], [65] - [70].

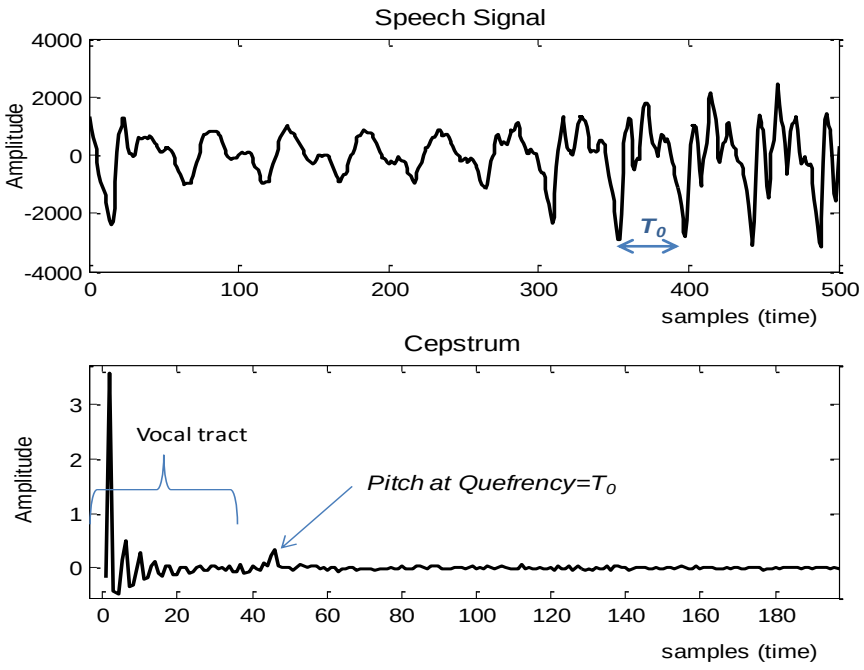


Figure 2.9 - Illustration of decomposition of vocal tract and pitch function using cepstrum function.

2.3.2.4 Instantaneous Frequency (IF) Method

Instantaneous frequency is a fundamental concept that can be found in many disciplines such as communications, speech, and music processing. In the continuous time-frequency domains the instantaneous frequency spectrum is defined as the first derivative of the phase spectrum as

$$\varphi(f,t) = \frac{d\theta(f,t)}{dt} \tag{2.17}$$

where $\theta(t)$ is the phase term. Consider a harmonic model of speech signal with the k^{th} harmonic component of the signal defined as

$$x_k(t) = a_k cos\big(\theta_k(t)\big) \tag{2.18}$$

where

$$\theta_k(t) = 2\pi k F_0 t + \varphi_k(t) \tag{2.19}$$

Now taking the derivative of the phase of the k^{th} harmonic w.r.t. time gives

$$\frac{d\theta_k(t)}{dt} = 2\pi k F_0 + \frac{d\varphi_k(t)}{dt} \tag{2.20}$$

Assuming that $\frac{d\varphi_k(t)}{dt}$ is relatively small

$$\frac{d\theta_k(t)}{dt} \approx 2\pi k F_0 \tag{2.21}$$

From Equation (2.21), the frequency of a harmonic is proportional to the derivative of phase. For discrete-time speech signals in [63] a method is proposed for extraction of IF from the differences between the phase spectrum of the short-term Fourier transform (STFT) of two consecutive speech frames; the basic idea can be expressed as

$$\Delta\theta(k,m) = \theta(k,m) - \theta(k,m-1) = arg[X(k,m) - X(k,m-1)] \quad (2.22)$$

where the variables k and m are the discrete frequency and the frame index respectively, $X(k,m)$ is the time-varing DFT spectrum and arg[·] is the phase function. The contributions in IF-based pitch extraction include the followings [63], [71] - [74].

2.3.2.5 Adaptive Comb Filters for Pitch Estimation

A comb filter is defined as a pole-zero filter whose zeros are on the unit circle and whose poles, at the same frequency as the zeros are used to adjust the selectivity, bandwidth of the filter [61], [75] - [81]. The transfer function of a comb filter with fundamental frequency F_0 is defined as

$$H(z) = \prod_{k=1}^{N_h} \left(\frac{\left(1 - e^{-j2\pi jkF_0} z^{-1}\right)\left(1 - e^{-j2\pi jkF_0} z^{-1}\right)}{(1 - re^{-j2\pi jkF_0} z^{-1})(1 - re^{-j2\pi jkF_0} z^{-1})} \right) \quad (2.23)$$

where N_h is the number of harmonics of the speech signal.

The main parameter to adjust is the fundamental frequency of the comb filter F_0. This parameter can be optimised by noting that for the corrected value of F_0 when the comb notch values coincide with the peaks of the harmonics the filter output energy is minimised. Hence an energy minimisation approach such as the gradient descent based adaptive least mean squared (LMS) error algorithm can be used to find the value of F_0 for which the comb filter output energy is at a minimum [67], [82]. In [78] an adaptive lattice notch filter is described where optimisation is performed by minimisation of forward and backward residues of error, and in [83] pitch detection in musical sounds using a number of parallel comb filters is described.

Figure 2.10 (a - b) are illustration of the poles and zeros and the frequency response of a comb filter.

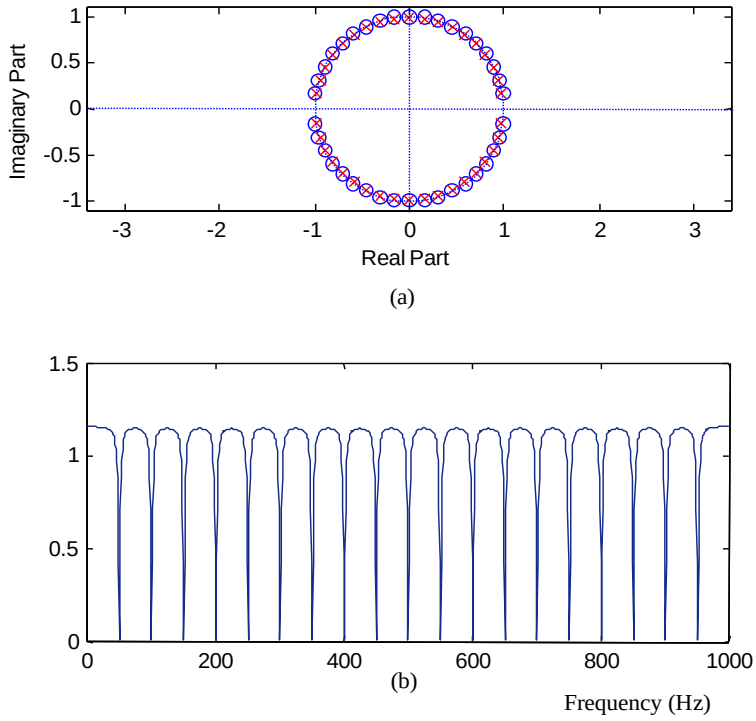


Figure 2.10 - Illustration of the (a) poles and zeros parameter, and (b) the frequency response curve of the adaptive comb filter.

2.3.3 Model-Based Pitch Extraction Methods

The term 'model-based methods' is a generic term that refers to variety methods that as a common feature include a parametric model of the description of the signal. Model-based methods may include:

1) A synthesis or generative model of speech signals such as a harmonic plus noise model (HNM) or a linear prediction (LP) model. Such models would have an input signal, an output signal and a set of filter model coefficients and a parametric

model of the fundamental frequency [20], [33], [39], [84]- [85]. For example GSM mobile phone employs an LP model of speech [ITU-T GSM standard].

2) A statistical model of the probability distribution of the signal such as a Gaussian model [39], [86], a Gaussian mixture model, hidden Markov-model (HMM) which involves with finite-state model [41] - [42], [87] - [88].

3) A combination of a generative/synthesis and a statistical model. For example, LP or HNM model may be formulated within a maximum likelihood framework.

As described in Chapter 3, the two common forms of parameterisation of speech are linear prediction model (LP) and harmonic plus noise model (HNM). A linear prediction model of a segment of *N* samples of speech can be described compactly in Chapter 3 as

$$x = Xa + e \tag{2.24}$$

where $a = [a1, a2, \cdots, aP]$ are the LP model coefficients. Similarly a harmonic plus noise model of a segment of *N* samples speech can be described compactly as

$$x = Ca + e \tag{2.25}$$

where C is a matrix of sine and cosine functions and a is vector containing the weights assigned to sine and cosine components and e is the noise component of HNM method. Note the similarity of form between Equations (2.24) and (2.25), in both cases the solutions may be derived from a least mean squared error or a maximum likelihood optimisation method. Also note that it is assumed that the pitch is given or that it can be estimated using a separate pitch estimation method.

2.3.3.1 Maximum Likelihood (ML) of Pitch Estimation

Maximum likelihood pitch estimation uses a statistical approach to find the most likely parameters which model a segment of speech signal. The most common approach to ML method is to calculate the likelihood of a residual signal obtained as the difference between the speech signal and an estimate of the periodic component of speech. The ML approach can be formulated as

$$\mathrm{ML}(\hat{F}_0) = \hat{F}_0^{ML} = arg\max_{\hat{F}_0}(f(x|F_0)) = \arg\max_{\hat{F}_0}(f(e = f(x)|F_0)) \tag{2.26}$$

where e is the residue from an operation on x such as linear prediction inverse filter output $e = x - AX$ or it may be the difference between the spectrum of the actual signal and the synthesised harmonic based on the pitch proposal F_0. The maximisation of the ML function (coinciding with minimisation of the error residue) can be performed on the log probability as

$$\hat{F}_0^{ML} = \arg\max_{\hat{F}_0}\left(log(f(x|F_0))\right) \tag{2.27}$$

The commonly used least squared error criterion may be considered as a special case of ML estimation where the signal has a Gaussian probability distribution. Furthermore when the likelihood function $f(x|F_0)$ is Gaussian then the ML estimate is the same as the least squared error (LSE) estimate due to the fact that maximisation of a Gaussian function is equivalent to minimisation of the exponent, e.g. $(x - \hat{x}(F_0))^2/\sigma^2$, of the Gaussian function [89].

Hence, Bayesian minimum mean squared error (MMSE) is a probabilistically weighted MMSE method [39], [41] - [42], [60], [87], [90].

2.3.3.2 Maximum a Posterior (MAP) of Pitch Estimation

Maximum a-Posteriori (MAP) is a special case of Bayesian function when the cost is $C(\hat{F}_0, F_0) = 1 - \delta(\hat{F}_0, F_0)$, then the Maximum a positerior method is obtained

$$\text{MAP}(\hat{F}_0) = \hat{F}_0^{MAP} = arg\max_{\hat{F}_0}\left(\underbrace{f(F_0|x)}_{Posterior\ pdf}\right) = arg\max_{\hat{F}_0}\left(\underbrace{f(x|F_0)}_{Likelihood}\ \underbrace{f(F_0)}_{Prior}\right) \quad (2.28)$$

Where δ is the Kronecker delta function.

The maximisation of the MAP function can be performed on the log probability as

$$\hat{F}_0^{MAP} = \text{arg}\max_{\hat{F}_0}\left(log(f(x|F_0)) + log(f(F_0))\right) \quad (2.29)$$

Note that the main difference between ML and MAP pitch estimators is that the latter employs a prior knowledge of the distribution of pitch F_0 in the form of the probability function $f(F_0)$ [89], [91].

3

SPEECH MODELS: PRODUCTION, SYNTHESIS AND DISTRIBUTION

This chapter describes the principles of human speech production mechanism; including the function and models of the acoustic articulator components. The voiced and unvoiced speech production and the commonly used speech models such as the source-filter linear prediction model and the harmonic plus noise model are described. The histograms of pitch distributions are shown and the probability distribution models of the pitch such as the Bayesian pitch estimation models including the maximum likelihood and the maximum a posterior model are presented.

3.1 INTRODUCTION

Speech signals are the natural and primary form of human communication, where acoustic signals (i.e. variations of air vibrations) are employed to convey words, sentences and intonations from which the listener deduces meaning, expression, intension, emotion and accent.

Human speech signals are produced by vibrating vocal organs mainly the vocal cords and the vocal tract and moving the articulators mainly the jaws, the tongue and the lips. The air vibrations coming out of the mouth/lips set the surrounding air into motion which, as the communication channel/medium, facilitates transmission of the speech from the source speaker to the receiver listener.

Speech sounds are auditory sensations of air pressure vibrations produced by air exhaled from the lungs. The air flows through the larynx, which contains the vocal cords, to the pharynx (throat cavity) and then goes through the oral cavity and the lips and also via the nasal cavity and the nostrils. Both the oral cavity and the nasal cavity can be closed. The tube leading from the larynx to the pharynx and from there on to the oral and nasal cavities is called the vocal tract [13**,**25]**,** [92].

The speech signal production may be modelled as the convolution of the excitation waveform produced by the glottis and the impulse of the vocal tract. Speech sounds are modulated and spectrally shaped by the frequency and mode of vibrations of the glottal cords and the frequency response and resonances of the vocal tract and the anti-resonances of the nasal cavity as the air is pushed out through the lips and nose.

Speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration) to convey information about words, speaker identity, accent, expression, style of speech, emotion and the state of health of the speaker. Most of this information is conveyed primarily within the traditional telephone bandwidth of 4 kHz. The speech energy above 4 kHz mostly conveys audio quality and sensation and some of the information of unvoiced.

3.2 THE PHYSIOLOGY SPEECH PRODUCTION MODEL

Acoustic speech output is produced by exhaling air from lungs through trachea, larynx, vocal cords (vocal valve), and epiglottis, oral and nasal cavities and finally out through the mouth opening and lips as illustrated in Figure 3.1.

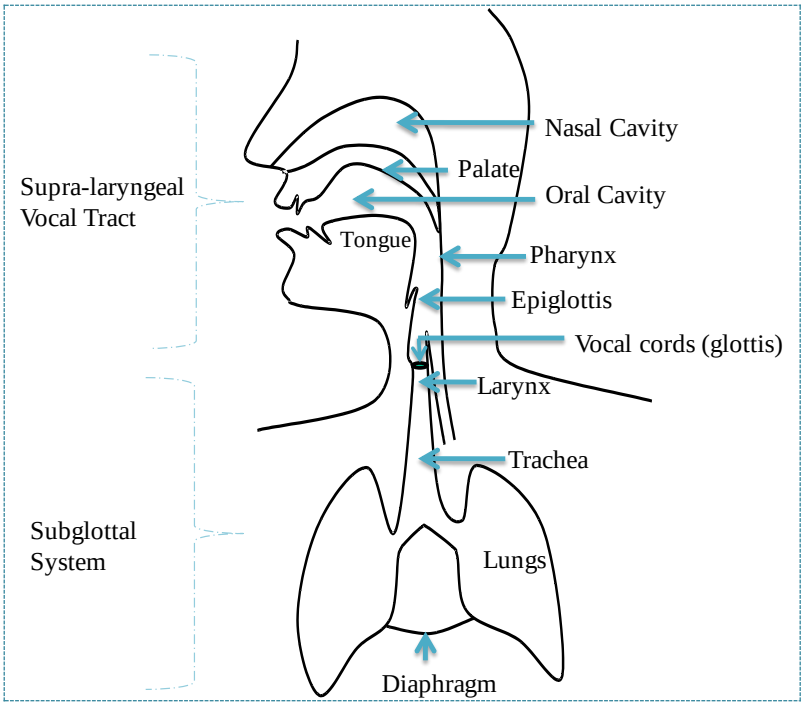


Figure 3.1 - Illustration of anatomy of speech production.

3.2.1 The Anatomy of Speech Production System

The act of production of speech begins with exhaling (inhaled) air from the lung. The exhaled air is continuously replenished. Without the subsequent modulations and spectral shaping, by the voice production system, this air will sound like a low energy random noise, not dissimilar to that from a deflating balloon, with no information imprinted upon it.

The information is first modulated onto the passing air by the manner and the frequency of vibrations of the closing and opening of the glottal folds, the resonance frequencies of the vocal cavity, the passage/blocking of air through the nasal cavity and the shape and the opening of mouth and lips.

The output of the glottal folds, in the form of a sequence of pulses for voiced sounds and air turbulence for unvoiced sounds, is the excitation signal to the vocal tract which is further shaped by the resonances of the vocal tract and the effects of the openings to the nasal cavities and the teeth and the shape of lips.

Figure 3.1 shows an outline of the anatomy of speech production system. The speech production system consists of the lungs, larynx, vocal tract cavity, nasal cavity, teeth, lips, and the connecting tubes.

1) *Lung*- The lung acts as the source of the air that is exhaled and in passing through the vocal system, is shaped and modulated with speech information to result in the acoustic form of speech output from the lips. The exhaled air is continuously replenished through inhalation. In fact efficient use of the exhaling and inhaling process, during speech production, is an important aspect of speech production.

2) The total volume of air that an average adult can hold in male/female lungs is around 6-7 litres. However, only part of this air can be actually used for speech production and physical activities; around 2 litres of air is always present in the lungs, and is called residual volume. This residual air could not be expelled unless the lungs collapse. The remaining volume of 4-5 litres, called the tidal volume, is usable for respiration or voice use. However, 10-15% of the tidal volume of air in the lungs is used in speech production [93]. The rest is held in reserve for more demanding physical activities, such as physical exercise or singing, which can demand our entire tidal volume.

3) *Bronchi and Trachea* -The lungs are connected via left and right bronchi tubes to the trachea tube which goes up to the vocal cords (the trachea or windpipe). The trachea is made of smooth muscle tissue along the back wall with 16 to 20 C-shaped bands of cartilage running along its length. The air at the point that the air exits lung, in its way to the larynx, it is entirely noise like and random and of very little loudness as evident from individuals who had laryngectomy surgery [94].

4) *Larynx* - The larynx, commonly known as voice box, is an organ in the production of sound that allows changes in pitch and volume of sound and also serves to protect the upper part of the trachea. The larynx houses the vocal folds or vocal cords and is shaped like a funnel. These components are connected to each other by muscles and ligament. The movement of these cartilages alters the tension of the vocal folds, which changes the pitch of the sound emitted by the vocal folds when they vibrate [95] - [96].

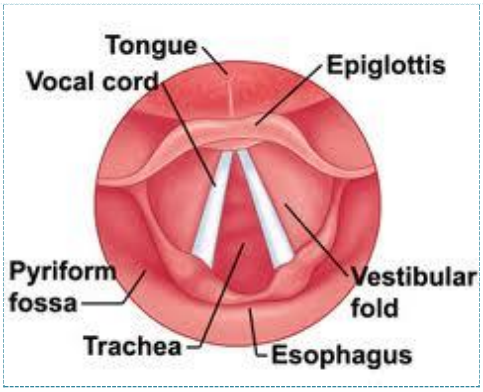


Figure 3.2 - Illustration of vocal cords (taken from [97])

5) *Vocal cords/vocal folds*- The shape of the opening and closing and the frequency of vibrations of the vocal cords affect the voiced/unvoiced character of sounds as well as the trajectory of the pitch and intonation in speech. The space between the vocal cords is called the glottis. The term glottis includes both the space between the vocal folds, called the membranous glottis, and the space behind the vocal folds between the arytenoid cartilages, called the cartilaginous glottis. The vocal cords are composed of two strings of muscle (mucous membrane) that form a V-shape stretched horizontally across the larynx in the respiratory tract as shown in Figure 3.2. Producing speech causes the vocal cords tighten together but then air from the lungs forces its way between the two vocal cords. The air causes the vocal cords to vibrate which, in turn, creates sound [98].

6) Male and female have different vocal fold sizes. The male vocal folds are between 17.5 - 25 mm in length, and the female vocal folds are between 12.5 - 17.5 mm. The average adult female pitch is around 210 Hz compared to the average adult male pitch of 120 Hz [98] - [99].

7) *Vocal tract*- The vocal tract consists of the laryngeal cavity, the pharynx, the oral cavity, and the nasal cavity. The vocal tract begins at the opening between vocal cords or glottis, and ends at the lips. The vocal tract acts as a time-varying filter with a set of around five time-varying resonances whose position and shape convey phonetic label/identity and speaker information. The estimated average length of the vocal tract in adult male humans is 16.9 cm and 14.1 cm in adult females. Assuming a closed tube model (resonance frequency = speed of sound / 4 × length of tube), these vocal tract lengths correspond to an average fundamental resonance frequency of 503 Hz for male and 603 Hz for female **[**100].

3.2.2 Production of Voiced/Unvoiced Excitation Signals

3.2.2.1 Voiced Speech Excitation Signal

Speech sounds are produced by either a periodic, or a non-periodic, sequence of opening and closing of the vocal folds. For voiced speech sounds with a periodic character, such as vowels or voiced consonants, the output of the larynx, which is the voicing input to the vocal tract, is composed of a quasi-periodic series of air pulses resulting from periodic cycles of opening and closing of vocal folds at a rate that is determined by the pitch and intended intonation of the speaker.

Figure 3.3 shows sketches of the gradual opening (3.3 (a-e)) and the gradual closing (3.3(f-j)) of the vocal cords during one cycle of production of voiced speech.

Voiced sounds are produced by a repeating sequence of opening and closing of glottal folds with a frequency of between 40 Hz (e.g. for a male with a low pitch and a heavy/coarse voice) to 600 Hz (e.g. for young children's voice) cycles per second (Hz)

depending on the speaker, the phoneme and the linguistic and emotional/expressional context. For an adult male, with a fundamental frequency of 100 cycles/second or 100 Hz, the duration of the average period of a single glottal cycle of the opening and closing of vocal cords is in the region of $1/100_{th}$ of second. This rate is too fast for the human ear to discriminate each individual opening/closing of the vocal cords. However, the overall rate of vibration is perceived as the pitch of the voice, "pitch" being the perceptual correlate of acoustic frequency.

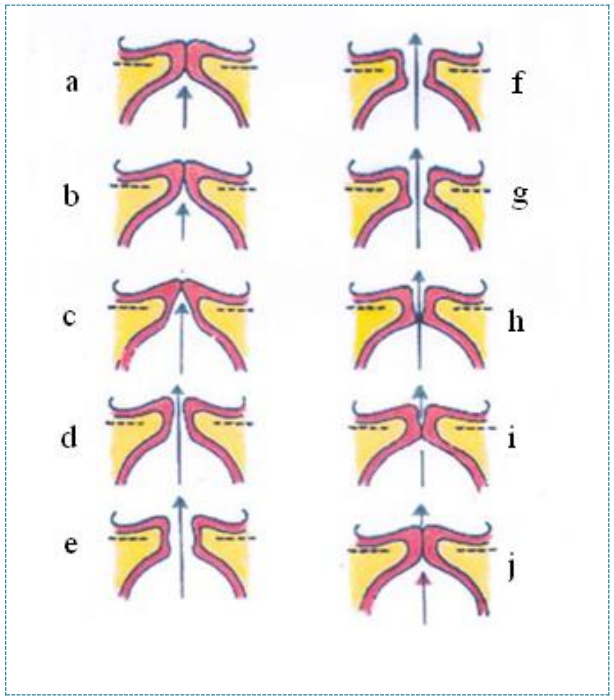


Figure 3.3 - Illustration of vocal cords activities during voiced speech production (taken from [101]).

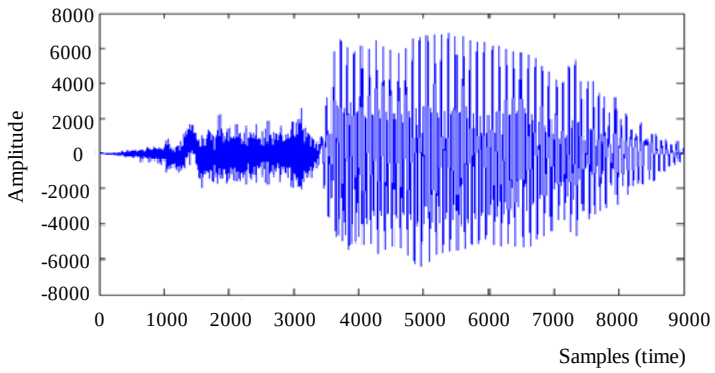
The periodicity of the glottal pulses determines the fundamental frequency, F_0 of the source signal and contributes to the perceived pitch of the sound. The time-variations of glottal pulse period convey the style, the intonation, the stress and emphasis in speech signals. In normal speech the fundamental frequency (pitch) changes constantly, providing linguistic clues and speaker information, as in the different intonation patterns associated

with questions or statements, or information about the emotional content, such as differences in speaker mood e.g. calmness, excitement, sadness etc.

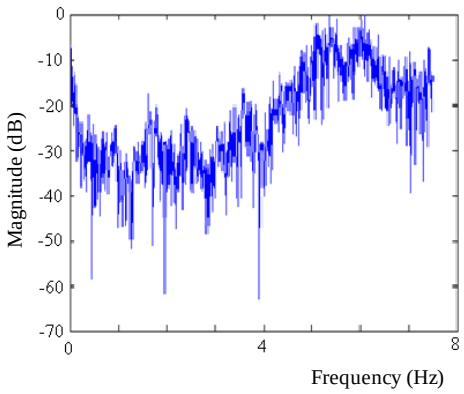
The manner and the duration of the opening and closing of the glottal folds in each cycle of voice sounds and contribute to the perception of the voice quality and the speaker's identify. The quality of voice and its classification into such types as normal (modal), creaky, breathy, husky, tense etc. depends on the glottal pulse shape.

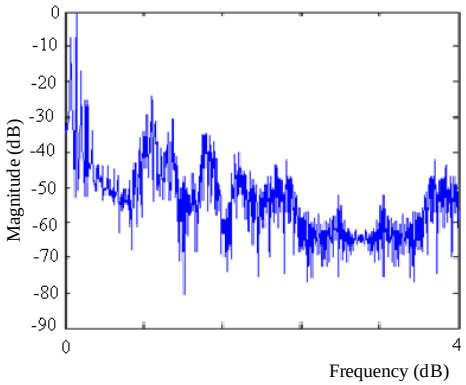
3.2.2.2 Unvoiced Speech Excitation Signal

For speech signal with a non-periodic character, such as unvoiced consonants and stops, the output of the larynx is non-periodic and may take one of several forms depending on whether the sound is nasal, fricative or stop. For fricatives the air flow input to the vocal/nasal cavities is formed by a small opening of the vocal fold constriction.



(a)





(b) (c)

Figure 3.4 - a) Acoustic production of the word ***sea*** (pronounced *s-iy*), (b) spectrum of the unvoiced segment "*s*", and (c) spectrum of the voiced speech segment "*iy*"[Taken from [4].

For unvoiced sounds air is passed through some obstacle in the mouth (e.g. when pronouncing 'S'), or is let out with a sudden burst (e.g. when pronouncing 'P'). The position where the obstacle is created depends on which speech sound (i.e. phoneme) is produced. During transitions, and for some mixed-excitation phonemes, the same air stream is used twice: first to make a low-frequency hum with the vocal cords, then to make a high-frequency, noisy hiss in the mouth [4].

If the vocal cords are held apart, air can flow between them without being obstructed, so that no noise is produced by the larynx. In voiceless fricatives such as /f/, /s/, /c/,/x/ the vocal cords are held apart. If there is a sufficiently high rate of airflow through the open glottis, a quiet disruption of the air, whisper, results. The glottal fricative /h/ has whisper phonation, as do whispered vowels, and the aspiration portion of voiceless aspirated stops such as English /p/, /t/, or /k/ in pre-vocalic position. For stops and fricatives, on the other hand, there are separate letters for voiced and voiceless sounds, e.g. /b/ (voiced) vs. /p/ [4].

Figure 3.4 shows an example of a speech segment containing an unvoiced sound "*s*" and a voiced sound "*iy*". Note that the spectrum of voiced sounds is shaped by the resonance of the vocal tract filter and contains the harmonics of the quasi-periodic glottal excitation, and has most of its power in the lower frequency bands, whereas the spectrum of unvoiced sounds is non-harmonic and usually has more energy in higher frequency bands. The shape of the spectrum of the input to the vocal tract filter is determined by the details of the opening and closing movements of the vocal cords, and by the fundamental frequency of the glottal pulses.

3.3 SOURCE-FILTER MODEL OF SPEECH

Speech sounds result from a combination of a source of sound energy from the lung modulated by time-varying openings and closings of vocal cords and spectrally shaped by the transfer function filter of vocal articulators determined by the shape and size of the vocal tract and nasal cavity. This results in a shaped spectrum with broadband energy peaks. This combination model is known as the source-filter model of speech production as shown in Figure 3.5. This model is at a common component of many speech analysis methods and also drives ideas in speech perception research. The source-filter model is a significant model of speech production, as an outline of the anatomy of the human speech production system as shown in Figure 3.1.

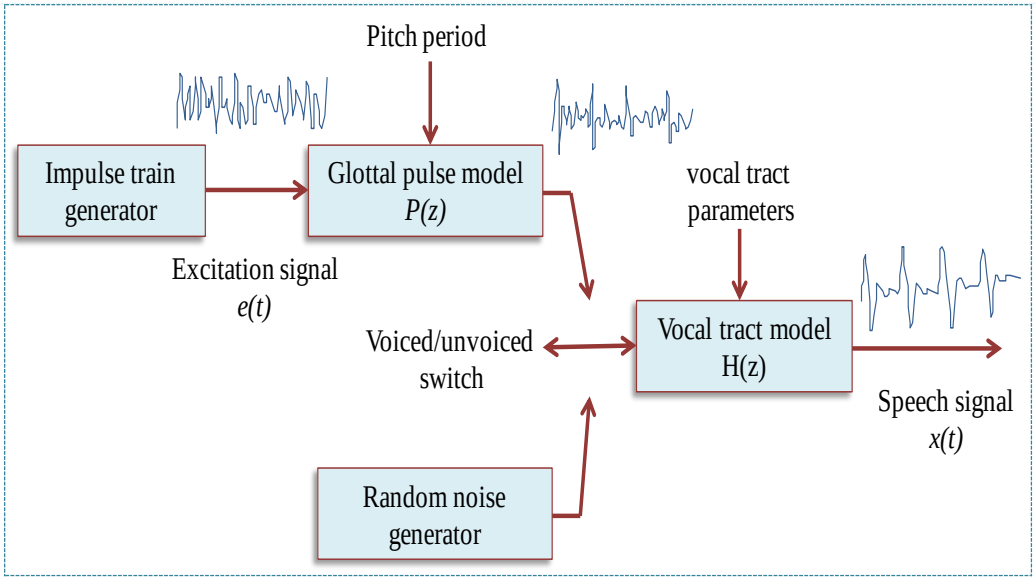
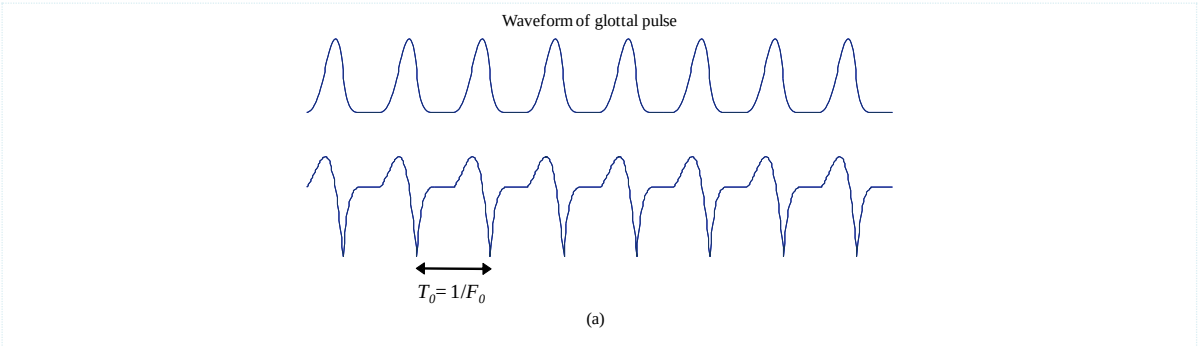


Figure 3.5 - A discrete-time source-filter model of speech production.

The source-filter theory describes speech production as a two stage process involving the generation of an excitation sound source, with its own spectral fine structure which is then filtered and spectrally shaped by the resonant properties of the vocal tract [102].

3.3.1 Voiced Source Signal Model

For voiced sounds, the source signal, the airflow from the lungs, is shaped into a quasi-periodic sequence of air pulses by the opening and closing vibrations of the vocal fold. Figure 3.5 shows a model of glottal pulse with an impulse train input and a filter model of the spectrum of the vocal folds.



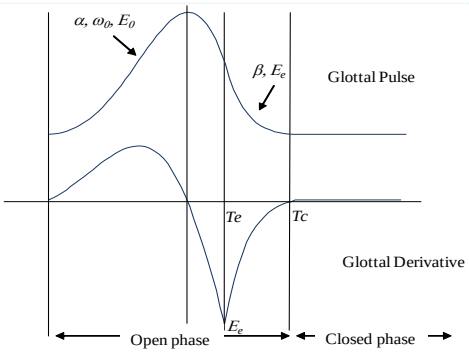


Figure 3.6 - Illustration of (a) the Liljencrants-Fan (LF) model of a sequence of glottal pulses a glottal pulse and (b) its derivative. (Taken from [4]).

Figure 3.6 (a) shows a model of a sequence of glottal pulses based on the Liljencrants-Fant (LF) model [103]. Figure 3.6 (b) shows a single LF glottal pulse and its derivative. The glottal pulse consists of an open phase, during which a pulse or puff of air is let through, and a closed phase. The open phase of the cycle itself is composed of an opening phase which culminates in the maximum opening of the glottal folds and a closing phase. The maximum negative value of the derivative of the pulse is reached at the point of the fastest rate of closing of the glottal folds.

The duration of each cycle is called the (duration of the) glottal pulse or pitch period length. We represent the length in time of the glottal pulse or pitch period length.

The LF model of the derivative of the glottal pulse is defined as

$$V_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin \omega t & 0 \le t < T_e \\ E_1\left(e^{-\beta(t-T)} - e^{-\beta(T_c - T_e)}\right) & T_e \le t < T_c \\ 0 & T_c \le t < T_0 \end{cases} \quad (3.1)$$

where a composition of a segment of less than ¾ of a period of a sine wave, with a frequency of ω and an exponential envelop $E_0 e^t$, is used to model the derivative of the glottal pulse up to the instance T_e where the derivative of the pulse reaches the most

negative value which corresponds to the fastest rate of change of the closing of the glottal folds. The final part of the closing phase of the glottal folds, the so called return phase, is modelled by an exponentially decaying function in the second line of Equation 3.1. In Equation 3.1, T_0 is the period of the glottal waveform; $F_0 = 1/T_0$ is the fundamental frequency (pitch) of speech harmonics, and T_c is the instance of closing of the glottal fold. The parameters E_0 and E_1 can be described in terms of the most negative value of the pulse E_e at the instant T_e; $E_0 = E_e/e^{\alpha T_e} \sin \omega T_e$ and $E_1 = E_e/\left[1 - e^{-\beta(T_c - T_e)}\right]$. The modelling and estimation of the glottal pulse is one of the ongoing challenges of speech processing research [4].

3.3.2 Unvoiced Source Signal Model

Unvoiced source of sound is generally modelled by a random noise sequence, such as a Gaussian noise. This noise is subsequently filtered by a vocal tract whose spectral shape determines the perception and identity of the unvoiced sound. Due to its randomness unvoiced signals has higher entropy than the more predictable voiced signals as a result in speech coding most of the coding bit resources are allocated to the encoding of unvoiced part of speech.

3.3.3 The Vocal Tract Filter Model

Whereas the source model describes the fine-detailed structure of speech spectrum, the filter model describes the spectral envelope of speech. The resonance characteristic of the physical space, such as the combination of vocal and nasal tracts, through which a sound wave propagates, changes the spectrum of sound and its perception.

The vocal tract space composed of the oral and the nasal cavities and the airways can be viewed as a time-varying acoustic filter that amplifies and filters the sound energy and shapes its frequency spectrum. The resonance frequencies of the vocal tract are called the formants. The identities of the acoustic realisation of phonemes are conveyed by the resonance frequencies at formants. Depending on the phoneme sound and the speaker characteristics there are about 3 to 5 formants in voiced sounds.

Formants are dependent on the phonemes but are also affected the overall shape, length, volume and reverberation characteristics of the vocal space and the vocal tract tissues and the associated parts i.e. nasal cavities, tongues, teeth and lips. The detailed shape of the filter transfer function is determined by the entire vocal tract serving as an acoustically resonant system combined with losses including those due to radiations at the lips.

3.3.3.1 Linear Prediction (LP) Model

Linear prediction model is a source-filter model of speech production. A LP model is defined as

$$x(m) = \sum_{k=1}^{P} a_k x(m-k) + \varepsilon(m) \tag{3.2}$$

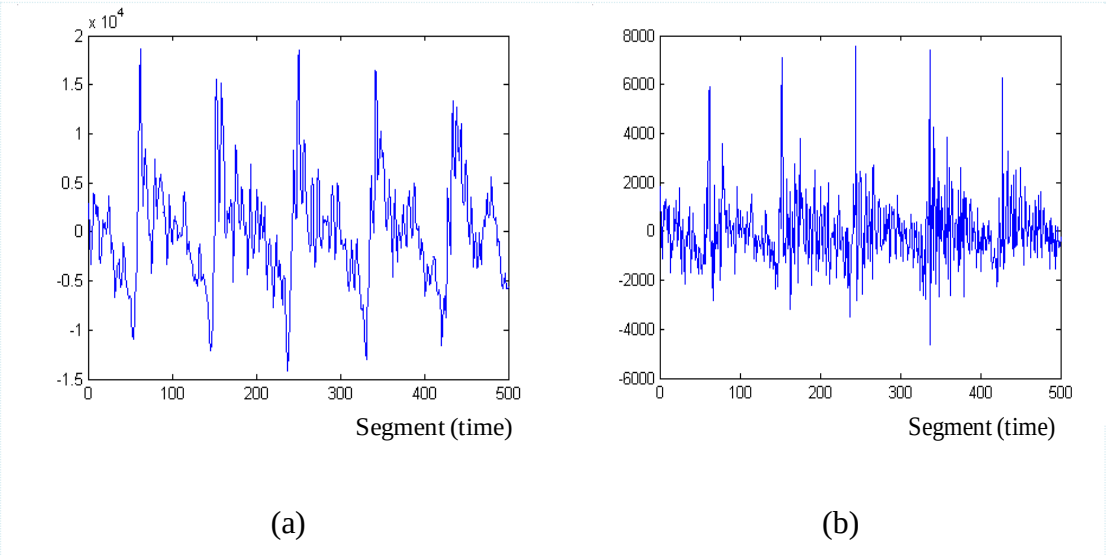
where $x(m)$ is a speech signal, a_k is LP parameters, and $\varepsilon(m)$ is speech excitation. The coefficients a_k model the correlation of each sample with the previous P samples whereas $e(m)$ models the part of speech that cannot be predicted from the past P samples.

Equation (3.2) could be represented in frequency domain as

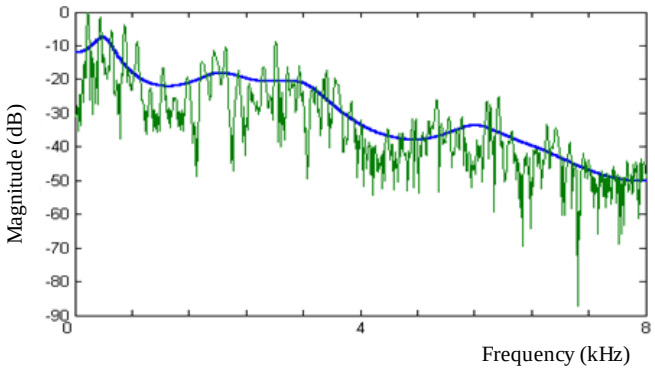
$$X(f) = \frac{E(f)}{1 - \sum_{k=1}^{P} a_k e^{-j2\pi fk}} = \frac{E(f)}{A(f)} = \frac{G.U(f)}{A(f)} \tag{3.3}$$

where $X(f)$ is the speech spectrum, $E(f)$ is the spectrum of excitation, $U(f)$ is the normalised power of excitation, G is a gain factor and $G/A(f)$ is the spectrum of the LP model of the combination of vocal tract and nasal cavities and lips as well as the spectral slope due to glottal pulse. In a source-filter LP model of speech, the spectral envelope of speech is modelled by the frequency response of the LP model $G/A(f)$, whereas the finer harmonic and random noise-like structure of the speech spectrum is modelled by the excitation (source) signal $E(f)$.

The model parameters $\{a_k\, k = 1, \cdots P\}$ of the speech spectral can be factorised and described in terms of a set of complex conjugate and real roots, and called poles of the model $\{\rho_k\, k = 1 \cdots P\}$. The poles are related to the resonance or formants of speech. Figure 3.7 shows the frequency response of a linear prediction model of a speech sound.



(a) (b)



(c)

Figure 3.7 - Illustration of (a) a segment of the vowel 'ay', (b) its glottal excitation, and (c) its magnitude Fourier transform and the frequency response of a linear prediction model of the vocal tract [taken from [4]].

3.3.3.2 Line Spectral Frequency (LSF) Model

The line spectral frequencies (LSF) are an alternative representation of linear prediction parameters. LSFs are used in speech coding, and in the interpolation and extrapolations of LP model parameters, for their good interpolation and quantisation properties. LSFs are derived as the roots of the following two polynomials:

$$P(z) = A(z) + z^{-(P+1)}A(z^{-1})$$
$$= 1 - (a_1 - a_P)z^{-1} - (a_2 - a_{P-1})z^{-2} - \cdots - (a_P - a_1)z^{-P} + z^{-P+1} \quad (3.4)$$

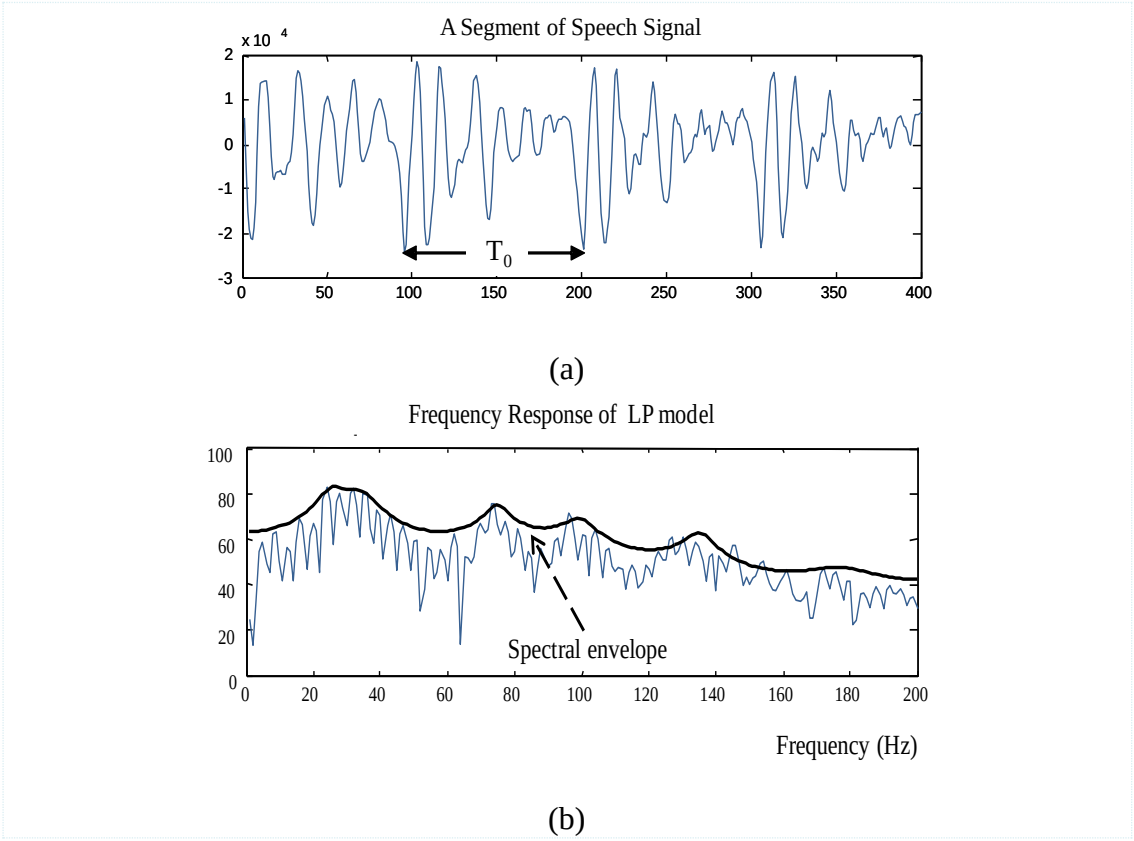
$$Q(z) = A(z) - z^{-(P-1)}A(z^{-1})$$
$$= 1 - (a_1 + a_P)z^{-1} - (a_2 + a_{P-1})z^{-2} - \cdots - (a_P + a_1)z^{-P}z^{-P+1} \quad (3.5)$$

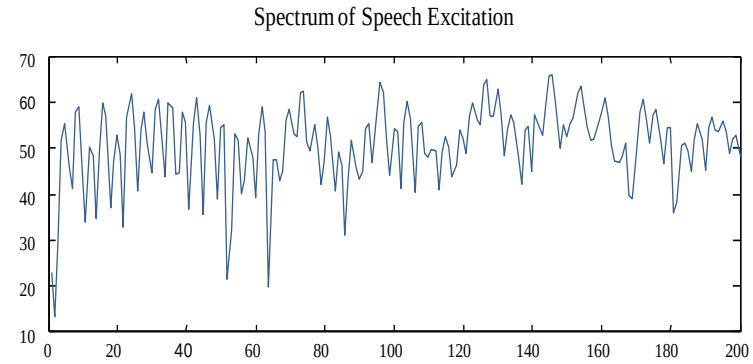
where $A(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_P z^{-P}$ is the inverse linear predictor filter. Clearly $A(z) = [P(z) + Q(z)]/2$. The polynomial Equations (3.4) and (3.5) can be written in factorised form as

$$P(z) = \prod_{i=1,3,5,\cdots} (1 - 2cos\omega_i z^{-1} + z^{-2}) \tag{3.6}$$

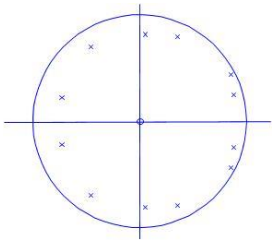
$$Q(z) = \prod_{i=2,4,6,\cdots} (1 - 2cos\omega_i z^{-1} + z^{-2}) \tag{3.7}$$

where ω_i are the LSF parameters. It can be shown that all the roots of the two polynomials have a magnitude of one and they are located on the unit circle and alternate each other.

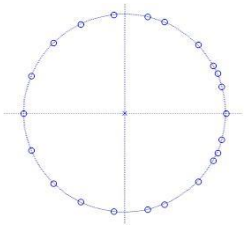




(c)



(d)



(e)

Figure 3.8 - (a) A segment of speech signal, (b) its FFT and LP spectra, (c) the spectrum of its excitation, (d) the poles of its LP model, (e) the roots of $P(z)$ and $Q(z)$ LSF polynomials.

Hence in LSF representation the parameter vector $[a_1, a_2, \ldots\ a_P]$ is converted to LSF vector $[\omega_1, \omega_2, \ldots\ \omega_P]$.

Figure 3.8, shows a segment of voiced speech together with poles of its linear predictor model and the LSF parameters.

3.4 HARMONIC PLUS NOISE MODEL (HNM) OF SPEECH

Harmonic plus noise, as the name implies, models speech as a combination of a harmonic component, modelled by a Fourier series, and a noise component

$$x(m) = \underbrace{\sum_{k=1}^{M} a_k \cos(2\pi k F_0 m) + b_k \sin(2\pi k F_0 m)}_{Harmonic\ Model\ (Fourier\ Series)} + \underbrace{\varepsilon(m)}_{Noise\ Model} \tag{3.8}$$

where F_0 is the fundamental frequency, a_k and b_k are the amplitudes of the sine and cosine components of the k^{th} harmonic, M is the number of the harmonics up to the bandwidth, and $\varepsilon(m)$ is the noise-like random components that model the fricatives and noise contents of speech.

The spectral shape of noise-like signal component of speech $\varepsilon(m)$ is often modelled by linear prediction model as

$$\varepsilon(m) = \sum_{k=1}^{P} c_k \varepsilon(m-k) + ge(m) \tag{3.9}$$

where c_k are the Linear Prediction (LP) model coefficients, $e(m)$ the unit variance random process and g is a gain factor. Hence the parameters vector of the HNM is $\{a, b, c, g$ and $F_0\}$. The harmonic part of the model is a Fourier series representation of the periodic component of the speech signal.

A segment of N samples of speech can be expressed in a vector-matrix notation as

$$\begin{bmatrix} x(m) \\ x(m-1) \\ x(m-2) \\ \vdots \\ x(m-N-1) \end{bmatrix} = \begin{bmatrix} cos2\pi F_0 m & \cdots & cos2\pi M F_0 m & sin2\pi F_0 m & \cdots & sin2\pi M F_0 m \\ cos2\pi F_0(m-1) & \cdots & cos2\pi M F_0(m-1) & sin2\pi F_{0(}m-1) & \cdots & sin2\pi M F_0(m-1) \\ cos2\pi F_0(m-2) & \cdots & cos2\pi M F_0(m-2) & sin2\pi F_0(m-2) & \cdots & sin2\pi M F_0(m-2) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ cos2\pi F_0(m-N-1) & \cdots & cos2\pi M F_0(m-N-1) & \sin 2\pi F_0(m-N-1) & \cdots & sin2\pi M F_0(m-N-1) \end{bmatrix}$$

$$\begin{bmatrix} a_1 \\ \vdots \\ a_M \\ b_1 \\ \vdots \\ b_2 \end{bmatrix} + \begin{bmatrix} v(m) \\ v(m-1) \\ v(m-2) \\ \vdots \\ v(m-N-1) \end{bmatrix} \tag{3.10}$$

In compact notation Equation (3.10) can be written as

$$x = Sc + v \tag{3.11}$$

where $\boldsymbol{x}$ is the vector of discrete-time speech samples, S is a matrix of sine and cosine functions, $c = [a\ b]$ is the vector of amplitudes of the harmonics and v is the noise component of the speech model. The harmonics amplitude vector c can be obtained from a least squared error minimisation process. Define an error vector as the difference between speech and its harmonic model as

$$e = x - Sc \tag{3.12}$$

The squared error function is given by

$$ee^T = (x - Sc\ )\ (x - Sc\ )^T \tag{3.13}$$

Minimization of Equation (3.13) with respect to the amplitudes vector $\boldsymbol{c}$ yields

$$c = [S^T S]^{-1} S^T x \tag{3.14}$$

The frequency domain representation of Equation (3.8) is defined as

$$X(f,t) = \underbrace{\sum_{k=1}^{M} A_k(t) M(f - kF_0)}_{Harmonic\ Model\ (Fourier\ Series)} + \underbrace{\epsilon(f,t)}_{Noise\ Model} \tag{3.15}$$

where t is the frame index, $M(f)$ is the Gaussian-shaped function.

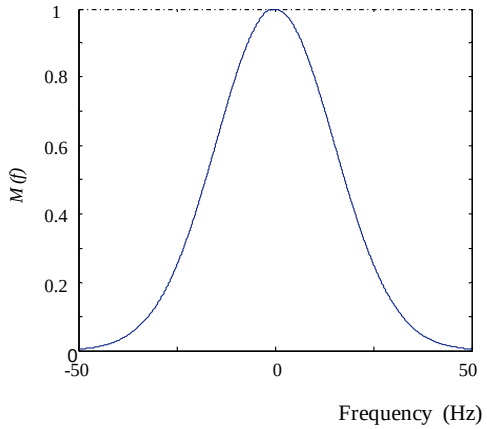


Figure 3.9 - Gaussian-shaped function *M*(*f*) is used for modeling harmonics.

3.4.1 A Harmonicity Model of Excitation

The proportion of harmonic and noise, at each harmonic, in voiced speech depends on a number of factors including: the speaker characteristics; the speech segment character (e.g. voice/unvoiced) and the harmonic frequency; the higher frequencies of voiced speech have a higher proportion of noise-like components. The ratio of the harmonic energy to the noise energy in each sub-band can be calculated as the level of *harmonicity* of that sub-band defined as:

$$H_k = 1 - \frac{\int_{-F_0/2}^{F_0/2} [|X(kF_0)|M(f - kF_0) - |X(f - kF_0)|]^2 \, df}{\int_{-F_0/2}^{F_0/2} |X(f)|^2 df} \tag{3.16}$$

where H_k is the harmonicity of the speech signal in the k^{th} band, $X(f)$ is the discrete Fourier transform (DFT) of the speech signal and $M(f)$ is a Gaussian shaped function shown in Figure 3.9 and defined as:

$$M(f) = \exp\left(-(f/\alpha)^2\right) \tag{3.17}$$

where typically a value of $\alpha = 45.5$ is used. The signal around each subband frequency of each frame is then reconstructed as:

$$\left|\hat{X}(f)\right| = X(kF_0)\left(\frac{H_k M(f-kF_0)}{\sqrt{\int M^2(f)df}} + \frac{(1-H_k)N(f)}{\sqrt{\int N^2(f)df}}\right) \quad for\ \ kF_0 - \frac{F_0}{2} < f < kF_0 + \frac{F_0}{2} \tag{3.18}$$

where *N(f)* is the noise component of the excitation. *N(f)* is a Rayleigh distributed random variable to comply with the assumption of the Gaussian distribution model of the speech DFT [4]. Figure 3.10 illustrates the excitation of a sample frame together with the harmonicity values of each band.

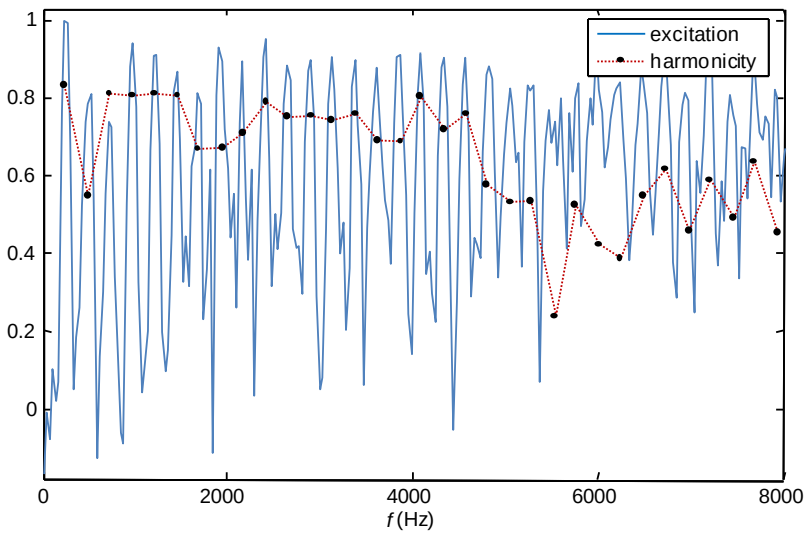


Figure 3.10 - Harmonicity of the excitation sub-bands superimposed on the normalized excitation.

3.4.2 Estimation of Harmonic Amplitudes

The harmonic amplitudes can be obtained from the tracking of the peak amplitudes of the DFT of speech at the frequency neighbourhoods around the integer multiples of the

fundamental frequency F_0. Alternatively, given an estimate of F_0, the following least square error estimation method can be used to obtain the harmonic amplitudes.

3.5 PITCH PROBABILITY DISTRIBUTION MODEL

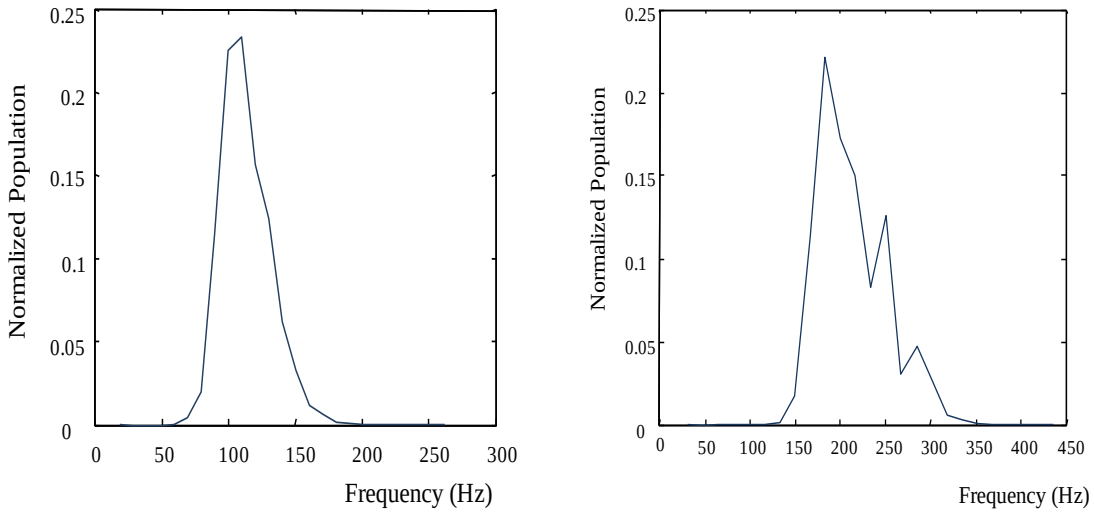
This section introduces the use of probability distribution models in pitch estimation.

3.5.1 Pitch Histograms

The probability density function (pdf) of pitch variations of a speaker may be obtained by fitting a curve to the histogram of the pitch extracted from the corresponding laryngograph records. Pitch pdf may be used as the prior function in a maximum a posteriori pitch estimation method.

In the following we plot the histograms for a number of individual male and female speakers and also for the group of male and female speakers. The variations in pitch among speakers imply that the pdf of a group of speakers has a larger variance and a broader shape than the pdf of an individual speaker. The pdf of combined group of male and female speakers shows two distinct modes corresponding to male and female genders. It is evident, as is well known, that pitch is a strong indicator of gender.

For male speakers the average pdf mode (peak) is at 120 Hz and most of the pdf is within the range of 70-200 Hz. For female speakers the average pdf mode (peak) is at around 185 and most of the pdf is within the range of 145-350 as shown in Figure 3.11 and Figure 3.12.



(a) (b)

Figure 3.11 - The normalised histogram of the pitch (a) of 11 male speakers: mean =112.7 Hz, standard deviation=18. Hz, and (b) of 10 female speakers, mean =213.3 Hz, standard deviation =37.3 Hz.

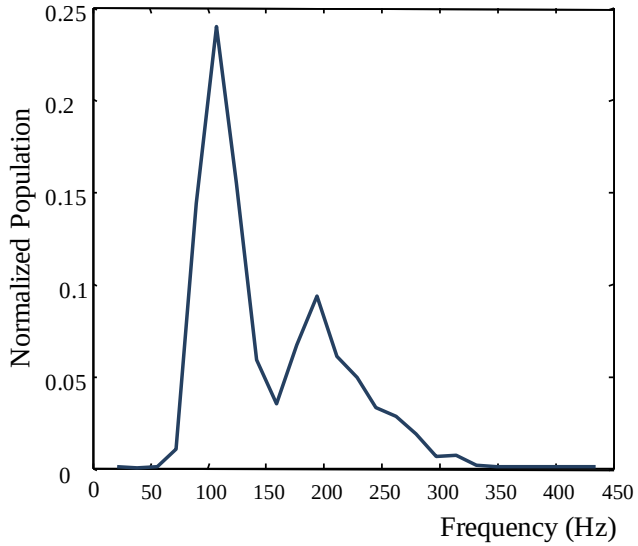


Figure 3.12 - The histogram of the pitch of all 21 speakers: mean= 151. Hz, standard deviation = 55.8 Hz.

Note two distinct modes one at 112 Hz corresponding to male speakers and the other at 213 Hz corresponding to female speakers.

3.5.2 Bayesian Formulation of Pitch Estimation Model

Given a segment of raw speech data $x = [x(0), x(1), \ldots x(N-1)]$, the posterior probability of pitch can be described, using the Bayes rule, as

$$f(F_0|x,a) = \frac{f(x,a|F_0)f(F_0)}{f(x,a)} \tag{3.19}$$

where $f(F_0|x,a)$ is the posterior probability of the unknown pitch value given the speech signal x and spectral envelope a, $f(x,a|F_0)$ is the likelihood of $\boldsymbol{x}$ and $\boldsymbol{a}$ assuming a value of F_0, $f(F_0)$ is the prior probability of F_0 and $f(x,a)$ here becomes a normalising factor.

The Bayes` estimate $\hat{F}_0$ is defined as

$$\text{Bayesian}(\hat{F}_0) = \min_{\hat{F}_0}\left(\int_{-\infty}^{\infty} C(\hat{F}_0, F_0) f(F_0|x) dF_0\right) \tag{3.20}$$

where $C(\hat{F}_0, F_0)$ is the cost of estimating an actual value of F_0 as $\widehat{F}_0, F_0$. The Bayesian function can be expressed, by expanding the posterior probability in terms of the likelihood and the prior, as

$$\text{Bayesian}(\hat{F}_0) = \min_{\hat{F}_0}\left(\int_{-\infty}^{\infty} C(\hat{F}_0, F_0) f(\boldsymbol{x}|F_0) f(F_0) dF_0\right) \tag{3.21}$$

The elements of the Bayes estimation are as follows;

1) **The Bayes cost function** $C(\hat{F}_0, F_0)$ associates a cost with estimation error $\hat{F}_0 - F_0$. Typical cost functions are mean squared error (MSE) and absolute value of error. The objective of Bayesian estimation is to minimise the cost of error.

2) **The likelihood function** $f(\boldsymbol{x}|F_0)$ provides the likelihood that the signal x is generated by a proposed value of pitch F_0.

3) **The prior function** f(F_0) provides a probability description of the distributions of the pitch value F_0 independent of any particular observation x. The prior distribution is therefore obtained beforehand (perhaps from a large database) and it describes the available knowledge of the distribution of the pitch of a person.

When $C(\hat{F}_0, F_0) = (\hat{F}_0 - F_0)^2$ the Bayesian *Minimum Mean Squared Error* (MMSE) solution is obtained.

3.5.2.1 Maximum a Posterior (MAP) Pitch Estimation

MAP is a special case of Bayesian function when the cost is $C(\hat{F}_0, F_0) = 1 - \delta(\hat{F}_0, F_0)$ then the Maximum a positerior method is obtained.

$$\text{MAP}(\hat{F}_0) = \hat{F}_0^{MAP} = \max_{\hat{F}_0}(f(F_0|x)) = \max_{\hat{F}_0}(f(x|F_0)f(F_0)) \tag{3.22}$$

The maximisation of the MAP function can be performed on the log probability as

$$\hat{F}_0^{MAP} = \max_{\hat{F}_0}\left(log(f(x|F_0)) + log(f(F_0))\right) \tag{3.23}$$

3.5.2.2 Maximum Likelihood (ML) Pitch Estimation

ML is a special case of Bayesian function when the cost is $C(\hat{F}_0, F_0) = 1 - \delta(\hat{F}_0, F_0)$ and when the prior function is uniform (i.e. when all the pitch value is equally likely or without any preference data)

$$\text{ML}(\hat{F}_0) = \hat{F}_0^{ML} = \max_{\hat{F}_0}(f(F_0|x)) = \max_{\hat{F}_0}(f(x|F_0)) \tag{3.24}$$

The maximisation of the ML function can be performed on the log probability as

$$\hat{F}_0^{ML} = \max_{\hat{F}_0} \left(log\left(f(x|F_0) \right) \right) \tag{3.25}$$

General Formulation of ML for Pitch Estimation

The most common approach to ML method is to calculate the likelihood of a residual signal obtained as the difference between the speech signal and an estimate of the periodic component of speech. Given a speech segment spectral vector $X = [X(0), \dots X(N-1)]$ and an estimate of the spectral envelop vector A, a harmonic plus noise model of speech may be written as

$$X = A\left(E_h(\hat{F}_0) + E_n \right) \tag{3.26}$$

Where $E_h(\hat{F}_0)$ and E_n are the harmonic and noise parts of the excitation signal.

The ML estimate is obtained via maximisation of the likelihood function

$$\mathrm{ML}\left(\hat{F}_0\right) = \hat{F}_0^{ML} = \max_{\hat{F}_0} \left(f\left(X \middle| F_0, \hat{A} \right) \right) \tag{3.27}$$

Using the harmonic plus noise model, the ML likelihood probability can be written as

$$\hat{F}_0^{ML} = \max_{\hat{F}_0} \left(f\left(X - \hat{A}\hat{E}_h(\hat{F}_0) \right) \right) = \max_{\hat{F}_0} \left(f\left(\hat{E}_n \right) \right) \tag{3.28}$$

where the likelihood of $\hat{F}_0$ is measured as the likelihood of the residual $E_n = X - AE_h\left(\hat{F}_0\right)$, i.e. the difference between the original speech segment $\boldsymbol{X}$ and the harmonic component $AE_h\left(\hat{F}_0\right)$ synthesised assuming that the fundamental frequency has a value of $\hat{F}_0$.

For the most commonly assumed form of the distribution of the residual, i.e. the Gaussian distribution, we have

$$f\left(E_n = X - AE_h(\hat{F}_0)\right) = \frac{1}{\sqrt{2\pi\sigma_{E_n}^2}} \exp\left(-0.5\,\frac{\left(X - AE_h(\hat{F}_0)\right)^2}{\sigma_{E_n}^2}\right) \qquad (3.29)$$

Different methods in the literature differ mainly in the way that the periodic component is modelled, synthesised and subtracted to yield the non-periodic residual.

3.5.3 Least Squared Error (LSE) Pitch Estimation

Least squared error may be considered as a special case of Bayesian estimation. When the prior function has a uniform distribution the MAP estimate reduces to an ML estimate. Furthermore when the likelihood function $f(\boldsymbol{x}|F_0)$ is Guassian then the ML estimate is the same as the *least squared error* (LSE) estimate due to the fact that maximisation of a Gaussian function is equivalent to minimisation of the exponent, e.g. $(\boldsymbol{x} - \hat{\boldsymbol{x}}(F_0))^2/\boldsymbol{\sigma}^2$, of the Gaussian function.

A note on LSE, MMSE and Bayesian MMSE Estimation

LSE and MMSE are commonly used as alternative descriptions of the estimation methods that minimise the average of the squared error of a target (original value or true value) of a signal and a synthesised/estimated version based on the values of some parametric function. For example in LSE or MMSE pitch estimation the mean squared error, due to a proposed value of pitch $\hat{F}_0$, may be obtained from the difference between the harmonics synthesised from $\hat{F}_0$ and the actual value of the signal.

In Bayesian MMSE estimation, the cost function is the squared error function $C(\hat{F}_0, F_0) = (\hat{F}_0 - F_0)^2$. However each value of the squared error, due to a proposed value of pitch $\hat{F}_0$, is weighted by the posterior probability of the true value of the unknown parameter $f(F_0|\boldsymbol{x})$. Hence, Bayesian MMSE is a probabilistically weighted MMSE method.

4

STATISTICAL MODELING AND SMOOTHING OF PITCH TRAJECTORIES

In this chapter a finite-state statistical model of the time-variations of the trajectory of the fundamental frequency of speech (pitch intonation) is presented. Based on this model three different and complementary post-processing methods for removing impulsive errors, step change errors and smoothing of the random fluctuations from the pitch trajectories are described.

4.1 INTRODUCTION

Examinations of many examples of the time-variations of the actual pitch trajectories, obtained from the laryngograph signals, shown in Figure 4.1, demonstrates that the pitch intonation is a low frequency slowly varying process.

As evident from Figure 4.1, the low frequency character of the intonation curve can be seen by observations of rise-connect-fall (RCF) pitch intonation **[**104**]** cycles, where each RCF usually spans a duration, of the order of several 100 ms, of a voiced utterance.

The observations that the pitch intonation trajectories are relatively smooth and slowly time-varying curves gives rise to the following general deductions that may be made regarding the smoothness and the continuity of pitch trajectories:

1) Continuity within each pitch utterance. Within each pitch utterance unit that is within a continuous voiced speech segment, the pitch trajectory is a highly correlated relatively slow time-varying, curve process, during which sudden changes such as step changes, impulsive changes, or short duration pulses in the value of pitch are not normally observed. Hence, within a voiced utterance it may be useful to employ filters or cost penalties that detect and remove sudden bursts of impulsive or step like changes in the pitch trajectory.
2) Change across two consecutive pitch utterances. Across two consecutive pitch utterance units, in many cases continuity of pitch trajectory is observed, however, there can be a step change in the pitch value from the end of one pitch utterance unit to the beginning of the next pitch utterance unit. Hence, at the start of a voiced segment one needs to allow for the possibility of a step change in the pitch value, relative to the value of the pitch at the end of the previous voiced segment.

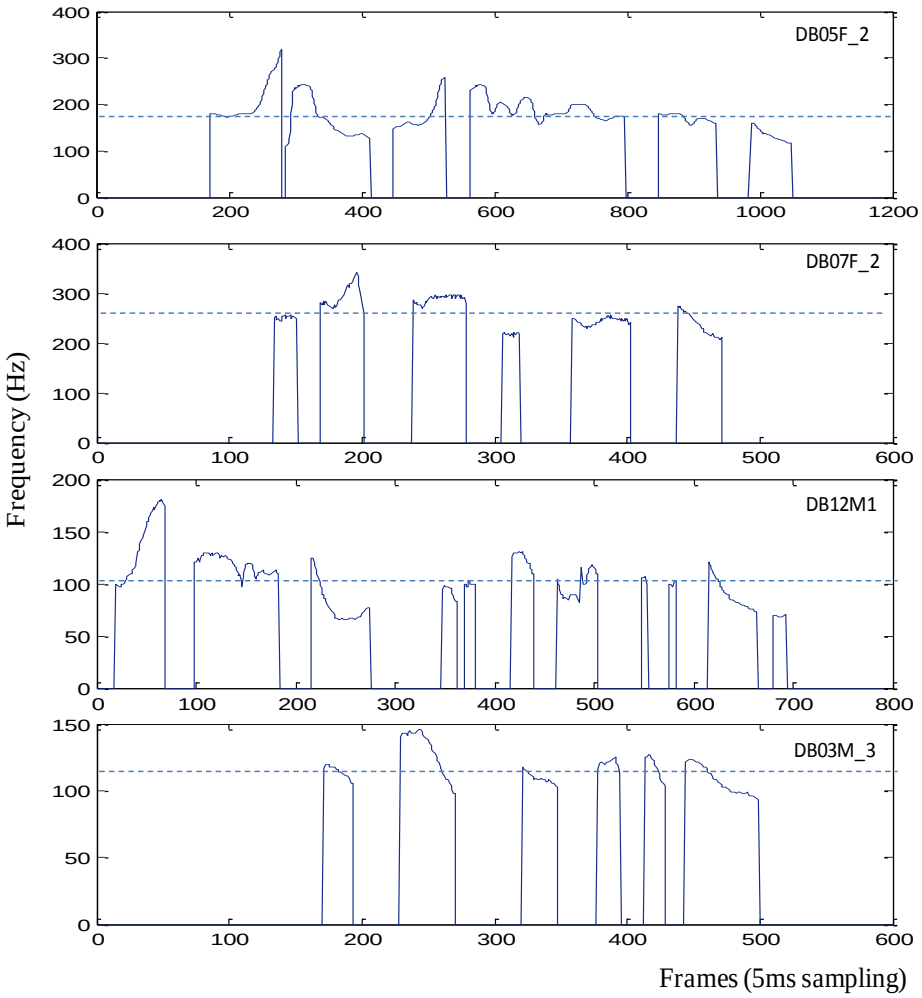


Figure 4.1 - Illustration of the time-variations of the actual pitch trajectories obtained from the laryngograph signals, and the dotted lines represent the mean pitch values.

3) Natural random variation in stressed or pathological voice. It should be noted that within each pitch utterance unit the pitch trajectory may oscillate, for example this happens when voice trembles during highly emotional voice expression, for example, as a signal expressing distress or extreme present fear.

Based on the above observation it is clear that a finite-state characterization of the pitch trajectory is an informative model for derivation of signal processing models, methods and algorithms that would limit errors in pitch estimation. For the purpose of pitch smoothing

and for limiting large erroneous impulsive and step changes in pitch, this thesis considers a finite-state model that is essentially composed of two states; a voiced state and an unvoiced state.

For the voiced state there are two types of transitions which impact the pitch trajectory differently:

1) Within-voiced-utterance state transition, in this state speech is already within a voiced state and the successive pitch values derived from the successive frames can be constrained to conform to variation within a smooth trajectory that excludes step or impulsive type changes.
2) Across-voiced-utterance state transitions which happens usually when there is a gap between two consecutive voiced segments e.g. when there is a voiced-unvoiced-voiced sequence or at the beginning of a new voiced utterance. This state-transition signals the beginning of a new voiced segment and a step change in the new pitch value relative to its previous value at the end of the last voiced segment is a possibility that needs to be allowed for.

Based on the above observations, in this chapter, first in section 4.2 we consider a finite-state model of pitch trajectory. Next, in the following sections, three different pitch post-processing methods are investigated for maintain smoothness and continuity in the pitch trajectory within each voiced utterance unit, these are:

1) Removal of impulsive change in pitch.
2) Removal of step changes in pitch.
3) Smoothing of the pitch trajectory.

4.2 PITCH TRAJECTORY MODELS

4.2.1 Finite-State Model

As explained the pitch signals are highly correlated processes that can be modeled by a slowly time-varying Markovian-Gaussian process. Figure 4.2 (a) shows a two state voiced-unvoiced model of pitch trajectory. Within the unvoiced state the speech signal is not periodic and hence it does not have a fundamental frequency as the vocal folds do not open and close in a periodic manner. Within the voiced state speech signal is periodic with a time-varying fundamental frequency.

The self-loop transition within the voiced state signals the continuation of the smooth evolution of the current pitch trajectory with no expected occurrence of step/impulse type changes in the pitch value. The transition, from the unvoiced state into the voiced state, signals the beginning of a new voiced state and should allow for an initial value of a smooth pitch trajectory that may be a step change different from the final pitch value of the previous utterance.

In other words, a step change discontinuity in the pitch may be observed across the gap in voiced-unvoiced-voiced speech segment but not within a voiced speech segment. An impulse type discontinuity is pitch is not expected in normal speech. The smooth variation of the pitch trajectory within each voiced state may be modeled by the well-known rise-fall or rise-connect-fall, RCF models. A set of relatively simple Markovian models of the pitch variations, within each pitch utterance, is a rise-fall and rise-connect-fall model as shown in Figure 4.2 (b - d). Note Figure 4.2 (b) is a two-state Markov model of the type of pitch utterance units that may be characterized and modelled by a rise and a fall.

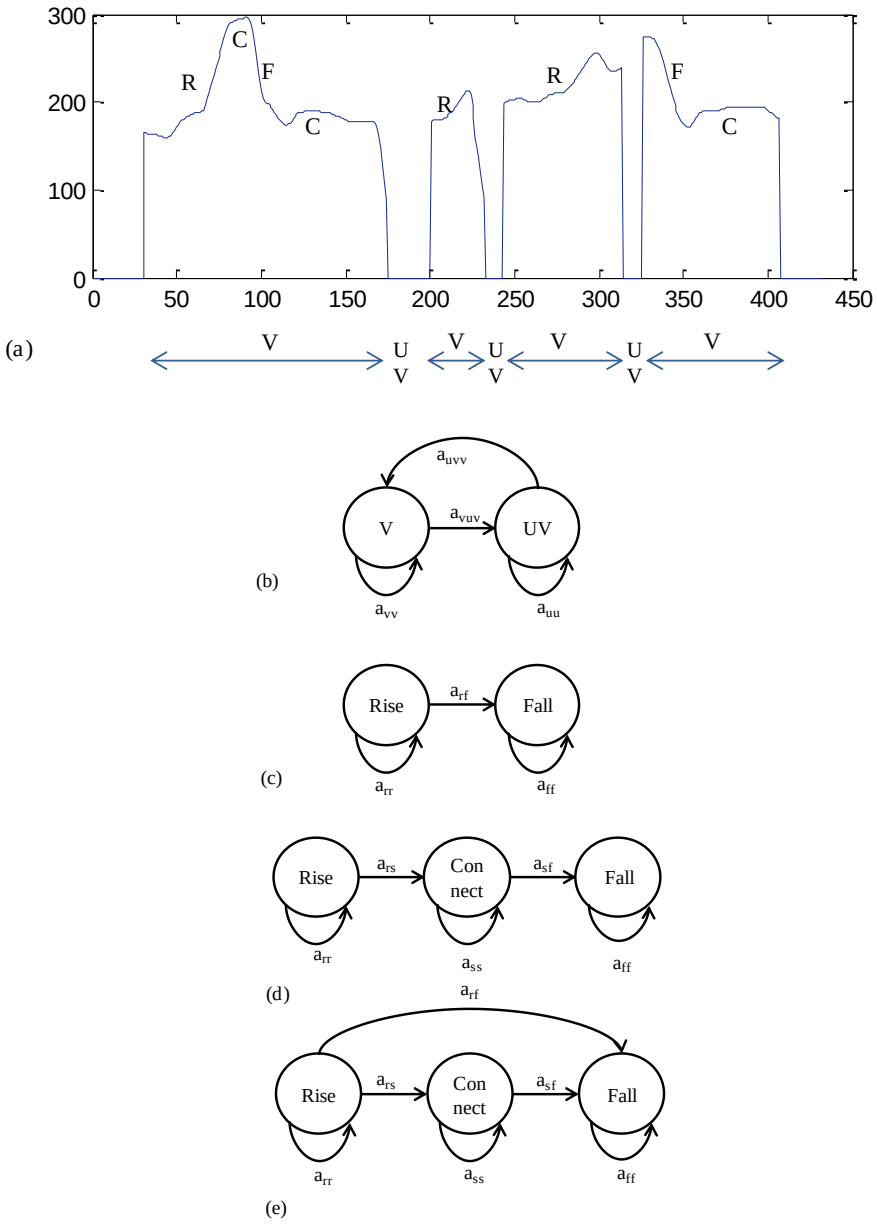


Figure 4.2 - Illustration (a) a pitch curve composed of a series of voiced-unvoiced, rise-connect- fall events and a set of Markovian model: (b)-(c) a two-state of rise and fall model, (d) a three-state of rise-connect-fall model, and (e) a combination of rise-fall and rise-connect-fall models that includes a skip-state transition.

Figure 4.2 (c) is a three state model that models the rise, connect and fall states of pitch utterance units. Figure 4.2 (d) is a Markov model combination of rise-fall and rise-connect-fall models that includes a skip state transition which allows the two models (4.2 b and c) to be integrated within one structure. Note the sum of state self-loop and exit transition is unity.

Within each voiced state the variation of the pitch trajectory is a slowly-time-varying process that may be modelled by a low order linear prediction model with a Gaussian input as explained in the next section.

4.2.2 Linear Prediction Model of Pitch

The smooth trajectory of pitch, within the voiced state of a finite-state model, can itself be modelled by a low order linear prediction model. For modelling the slow variations of pitch trajectory typically a linear prediction model of order 2-3 should be sufficient. Assume the sequence of pitch estimates within each utterance is denoted as $[\hat{F}_0(m)]$. For a pitch utterance unit, the prediction of the pitch value at frame m, $\bar{F}_0(m)$, given the previous P pitch estimated values, $\hat{F}_0(m-1)\cdots\hat{F}_0(n-P)$, may be modeled by a linear prediction function as

$$\bar{F}_0(m) = \sum_{k=1}^{P} a_k \hat{F}_0(m-k) \tag{4.1}$$

where $a_k{}'$s are the coefficients of the linear prediction model. The pitch prediction error, $\tilde{F}_0(m)$, is given as the difference between the estimated value and the predicted value as

$$\tilde{F}_0(m) = \hat{F}_0(m) - \bar{F}_0(m) = \hat{F}_0(m) - \sum_{k=1}^{P} a_k \hat{F}_0(m-k) \tag{4.2}$$

Note the 'true' pitch values, obtained from a laryngograph, are only available for evaluation purposes during the system research and development phase. During the system operation phase all we have is the pitch estimate sequence, $[\hat{F}_0(m)]$, for which smoothing and noise reduction functions may be developed.

Assuming that the pitch prediction error is a zero-mean random process with a Gaussian distribution, the probability of the estimate for the frame m, $\hat{F}_0(m)$, given the pitch estimates for the previous P frames, $\hat{F}_0(m-1) \cdots \hat{F}_0(m-P)$, may be expressed as

$$p\left(\hat{F}_0(m) \middle| \hat{F}_0(m-1) \cdots \hat{F}_0(m-P)\right) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{\left(\hat{F}_0(m) - \bar{F}_0(m)\right)^2}{2\sigma^2}\right) \tag{4.3}$$

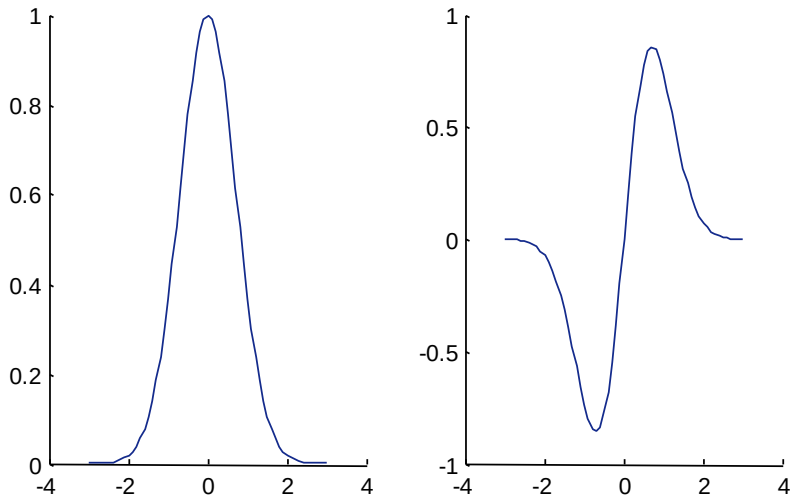
where σ is the variance of pitch prediction error.

The value of the variance of pitch prediction error can be used to define a limiting function that will limit sudden jumps to half pitch or double pitch values.

A linear prediction model incorporating a limiting function on the prediction error may be expressed as

$$\bar{F}_0(n,m) = \sum_{k=1}^{p} a_k \hat{F}_0(n, m-k) + g\big(e(m)\big) \tag{4.4}$$

where $g(e(m))$ is the limiting or influence function. An example of influence function is shown in Figure 4.3(b). Note the influence function in Figure 4.3 (b) is the derivative of the Gaussian function shown in Figure 4.3 (a).



(a) Gaussian pulse

(b) Influence pulse, derivative of Gaussian pulse

Figure 4.3 - Illustration of (a) the Gaussian pulse and (b) the influence pulse.

4.3 DETECTION AND REMOVAL OF IMPULSIVE AND PULSE NOISE FROM PITCH TRAJECTORY

This section is concerned with the modelling of the pattern of occurrence and removal of impulsive noise and short duration noise pulse in pitch trajectory [24].

4.3.1 Definition of an Impulse

An actual impulsive noise may be just one sample long or it may be a short duration pulse spanning several samples. The theoretical impulse function shown in Figure 4.4 is defined as a pulse of unit area with an infinitesimal time width as

$$\delta(t) = \lim_{\Delta \to 0} p(t) = \begin{cases} 1/\Delta, & |t| \leq \Delta/2 \\ 0, & |t| > \Delta/2 \end{cases} \quad (4.5)$$

The Fourier transform of the impulse function is obtained as

$$\Delta(f) = \int_{-\infty}^{\infty} \delta(t) e^{-j2\pi f t} dt = e^{0} = 1 \qquad (4.6)$$

where f is the frequency variable. Real impulsive noise is in fact short duration pulses with a finite duration and finite amplitude. They may also exhibit oscillatory characteristics which could be in part the impulse response of the system through which they propagate. Figure 4.5 shows the impulsive noise sequence as the output of an idealized impulse sequence and an impulse shaping filter [24].

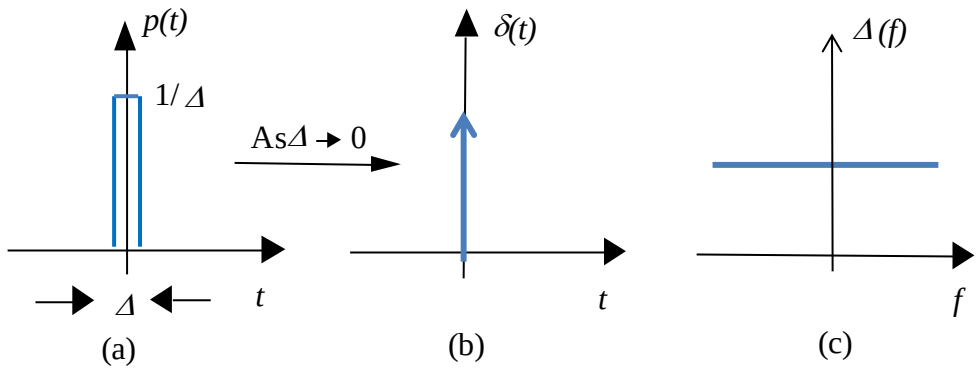


Figure 4.4 - (a) A unit-area pulse, (b) The pulse becomes an impulse as its duration $\Delta \to 0$, (c) The spectrum of the impulse function.

4.3.2 Probability Models of Impulsive Noise

An impulsive noise sequence $n_i(m)$ may be modeled as the output of a filter excited by an amplitude-modulated sparse random binary sequence as

$$n_i(m) = \sum_{k=0}^{P-1} h(k) n(m-k) b(m-k) \qquad (4.7)$$

where $b(m)$ is a binary-valued random sequence model of the time of occurrence of impulsive noise, $n(m)$ is a continuous-valued random process model of impulse

amplitude, and $h(k)$ is the impulse response of a filter that models the duration and shape of each impulse as illustrated in Figure 4.5. In the following three statistical processes for modeling an impulsive noise process are considered [4].

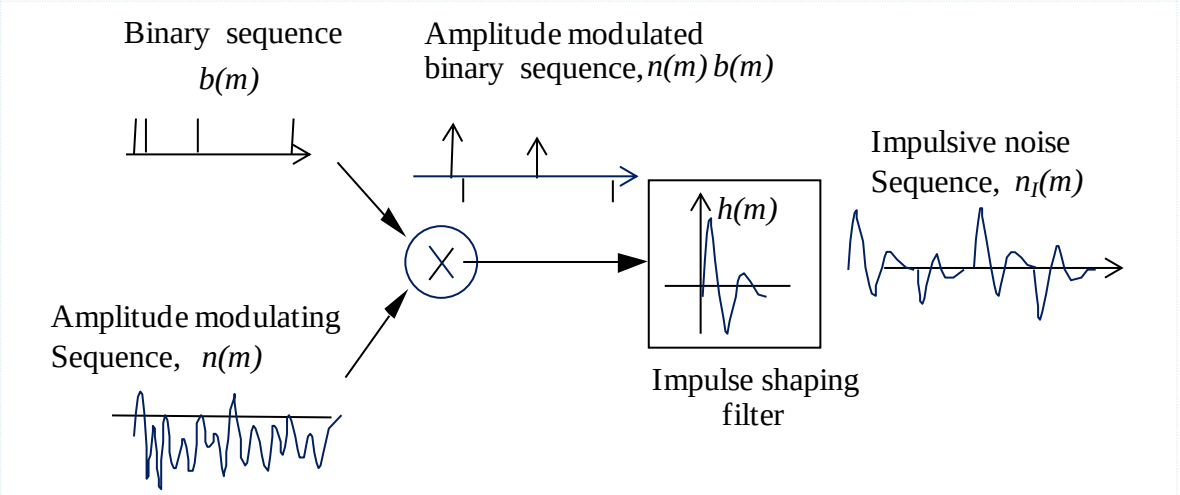


Figure 4.5 - Illustration of an impulsive noise model as the output of a filter excited by an amplitude-modulated binary sequence.

4.3.2.1 Bernoulli–Gaussian Model of Impulsive Noise

In a Bernoulli-Gaussian model of an impulsive noise process $n_i(m)$, the random occurrence of the impulses is modelled by a binary Bernoulli process $b(m)$ and the amplitude of the impulses is modelled by a Gaussian process as

$$f_N^{BG} = (1-\alpha)\delta\big(b(m)\big) + \alpha\delta\big(1-b(m)\big)f_N\big(n_i(m)\big) \tag{4.8}$$

where $\delta(\cdot)$ is the Kronecker delta function. The probability mass function of a Bernoulli process is given by

$$P_B\big(b(m)\big) = \begin{cases} \alpha & for\ b(m)=1, \\ 1-\alpha & for\ b(m)=0. \end{cases} \tag{4.9}$$

A Bernoulli process has a mean value μ_b of α and a variance of $\alpha(1-\alpha)$ [24].

4.3.2.2 Poisson–Gaussian Model of Impulsive Noise

In a Poisson model, the probability of occurrence or absence of an impulse in a short time interval Δt is given by

$$Prob(one\ \ impulse\ in\ a\ small\ time\ interval\ \Delta t) = \lambda \Delta t \tag{4.10}$$

$$Prob\ (zero\ impulse\ in\ a\ small\ time\ interval\ \ \Delta t)\ = 1 - \lambda \Delta t \tag{4.11}$$

The probability of k impulsive noise in a time interval of T is

$$p(k,T) = \frac{(\lambda T)^k}{k!} e^{-\lambda T} \tag{4.12}$$

In a Poisson–Gaussian model, the probability density function (pdf) of an impulsive noise $n_i(m)$ in a small time interval of $\lambda \Delta t$ is given by

$$f_N^{PG}(n_i\,(m)) = (1 - \lambda \Delta t)\delta\big(n_i(m)\big) + 1 - \lambda \Delta t f_N\big(n_i(m)\big) \tag{4.13}$$

where $f_N\big(n_i(m)\big)$ is the Gaussian pdf of Equation (4.8). From Equation (4.13) the mean and variance of the number of impulses in a time interval of T are given by

$$Expected\ (mean) number\ of\ impulse\ \ in\ T\ seconds = \lambda T \tag{4.14}$$

$$Variance\ of\ number\ occurances\ of\ impulse\ \ in\ T\ seconds = \lambda T \tag{4.15}$$

4.3.2.3 Hidden Markov Model of Impulsive and Burst Noise

Hidden Markov Models (HMMs) are defined by two sets of parameters, the Markovian state transition probabilities $\{a_{ij}\}$ and the state observation probabilities $\{b_{ik}\}$. The state transition probability models the pattern of occurrences of the impulses whereas the state observation probability models the amplitude of the impulses. A popular model for state observation probability is a Gaussian mixture model (GMM) [105].

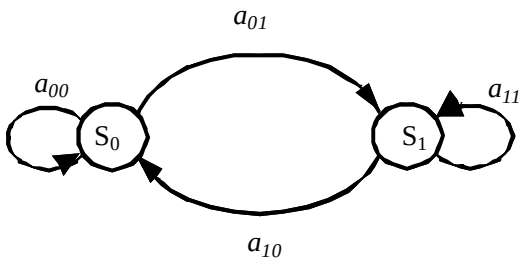


Figure 4.6 - A binary-state model of an impulse noise generator.

The state transition probabilities can affect a variety of different statistical patterns of the intervals of occurrences and the durations of impulsive noise. For example, as shown in Figure 4.6, the self-loop transition probability of S_0, a_{00}, can be used to control the duration of the impulse-absent whereas the self-loop transition probability of S_1, a_{11}, can be used to control the individual or burst nature of impulses emitted in state S_1.

4.3.3 Impulsive Noise Detection and Removal Using Linear Prediction Models

The impulsive noise removal system shown in Figure 4.7 consists of two subsystems: a detector and an interpolator. The detector locates the position of each noise pulse, and the interpolator replaces the distorted samples using the samples on both sides of the impulsive noise. Both the detector and the interpolator share the linear prediction analysis system.

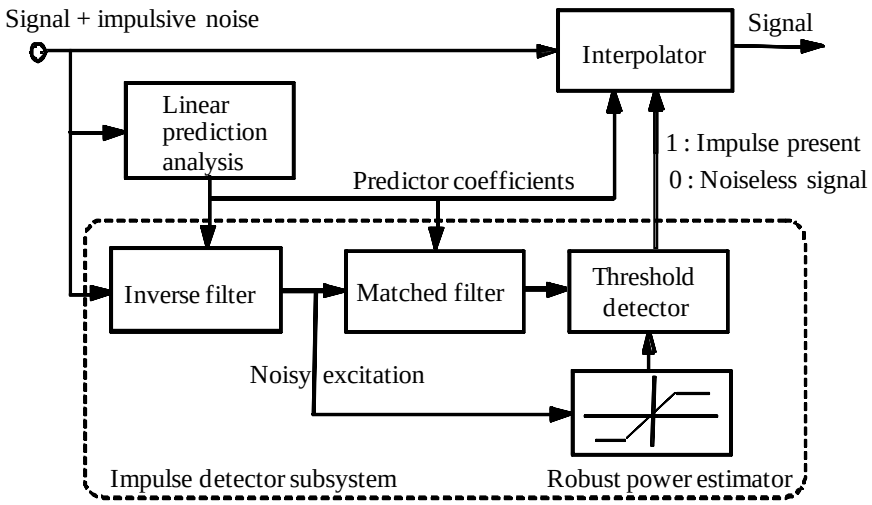


Figure 4.7 - Configuration of an impulsive noise removal system incorporating a detector and interpolator subsystems [taken from [4]].

A simple method for detection of impulsive noise is to employ an amplitude threshold, and classify those samples with amplitudes above the threshold, as noise, however, this method fails when the noise amplitude falls below the signal.

Detection can be improved by utilizing the characteristic differences between the impulsive noise and the signal. An impulsive noise, or a short-duration pulse, introduces uncharacteristic discontinuity in a correlated signal. The discontinuity becomes more detectable when the signal is differentiated. The differentiation (or, for digital signals, the differencing) operation is equivalent to decorrelation or spectral whitening which may be achieved by inverse filtering via transforming the noisy signal $y(m)$ to the excitation signal of a linear predictor which has the following effects:

(i) The scale of the signal amplitude is reduced to almost that of the original excitation signal, whereas the scale of the noise amplitude remains unchanged or increases.

(ii) The signal is decorrelated, whereas the impulsive noise is smeared and transformed to a scaled version of the impulse response of the inverse filter.

Both effects improve noise delectability.

4.3.4 Median Filters for Removal of Impulsive Noise

The classical approach to removal of impulsive noise is the median filter [24]. The median of a set of samples $\{F_0(m)\}$ is a member of the set $F_{0_{med}}(m)$ such that; half the population of the set is larger than $F_{0_{med}}(m)$ and the other half is smaller than $F_{0_{med}}(m)$. Hence, the median of a set of samples is obtained by sorting the samples in the ascending or descending order, and then selecting the mid-value.

$$F_{0_{med}}(m) = median(\{F_0(m)\}) = midpoint(sort(\{F_0(m)\})) \qquad (4.16)$$

In median filtering, a window of predetermined length slides sequentially over the signal, and the mid-sample within the window is replaced by the median of all the samples that are inside the window, as illustrated in Figure 4.8.

The output $\hat{F}_0(m)$ of a median filter with input $F_0(m)$ and a median window of length $2K+1$ samples are given by

$$\hat{F}_0(m) = F_{0_{med}}(m) = median[F_0(m-K), \cdots, F_0(m), \cdots, F_0(m+K)] \qquad (4.17)$$

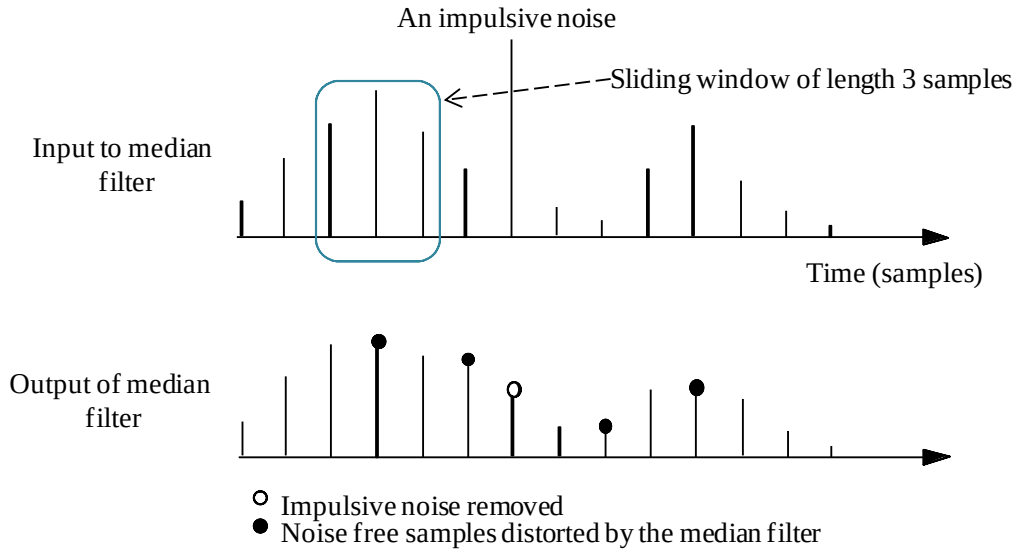


Figure 4.8 - Input and output of a median filter. Note that in addition to suppressing the impulsive outlier, the filter also distorts some genuine signal components when it swaps the sample value with the median.

4.4 REMOVAL OF STEP CHANGE DISCONTINUITY IN PITCH TRAJECTORY

Within each voiced pitch utterance unit, the pitch trajectory follows a smooth rise-sustain-fall curve or rise-connect-fall curve [104]. Hence, sudden step changes in the pitch trajectory, within an utterance unit, are unlikely to occur naturally and are most probably due to a large sustained error such as half pitch step change or double pitch step change estimation errors. Therefore, it is beneficial to apply a penalty to the overall pitch estimation cost function to penalize for step changes in pitch within a voiced utterance unit.

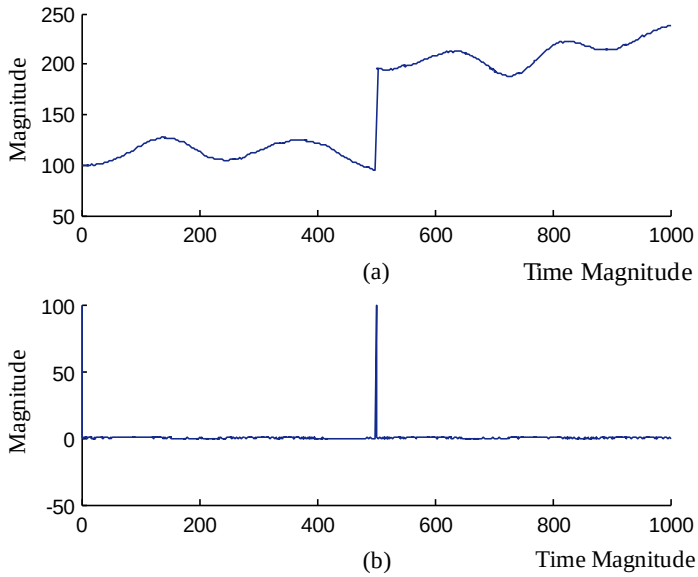


Figure 4.9 - Illustration of the step change values. (a) a signal with the distinct of step change at sample 500, (b) the step change detector.

A step change in discrete-time domain, by definition, usually implies that the overall trend in the neighbourhoods before and after the step change are similar, other than a step change in the mean value. The derivative of a step change would be a relatively large impulse and hence by monitoring of the first order differences in the pitch sequence estimate, it is possible to detect an uncharacteristically large change in pitch differences at the point where the step change happens. The derivative of the pitch function can be obtained by a first order FIR filter with coefficients

$$b = [1, -1] \tag{4.18}$$

The equation for this difference equation is

$$\Delta F_0(m) = F_0(m) - F_0(m-1) \tag{4.19}$$

Note this filter is a first order difference filter whose output is the difference between the current input sample and the immediately previous input sample.

Figure 4.9 illustrates this point, it shows a slowly varying curve with a step change at the sample number 500, the first order derivative of this curve shows a distinctive impulse at the point where the step changes happens.

The step change can be removed by processing the output of the difference filter with the following influence function (IF) as shown in Figure 4.10 (b)

$$y = xe^{-0.5x^2/\sigma^2} \tag{4.20}$$

The IF of Equation 4.20 is the derivative of the Gaussian function [106]. This function linearly passes the input to output for values of the input that fall within the variance or standard deviation but attenuates or may even block large outlying samples that are outside the variance or standard deviation. As shown in Figure 4.10 the variable σ^2can be increased/decreased to increase/decrease the linear region in which the signal is passed unaffected. Figure 4.10 and 4.11 show the impact of variation of σ^2 on the processed signal.

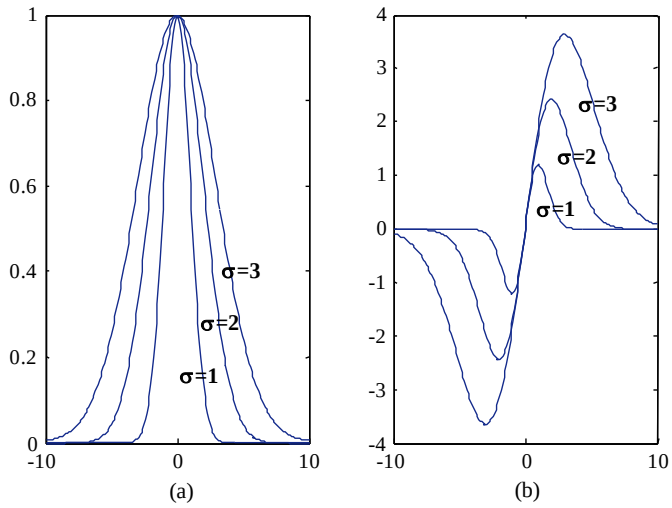


Figure 4.10 - Illustration of the variation of the shape of (a) Gaussian pulse and (b) its derivative, the influence function (IF), with three different values of the variance of σ^2.

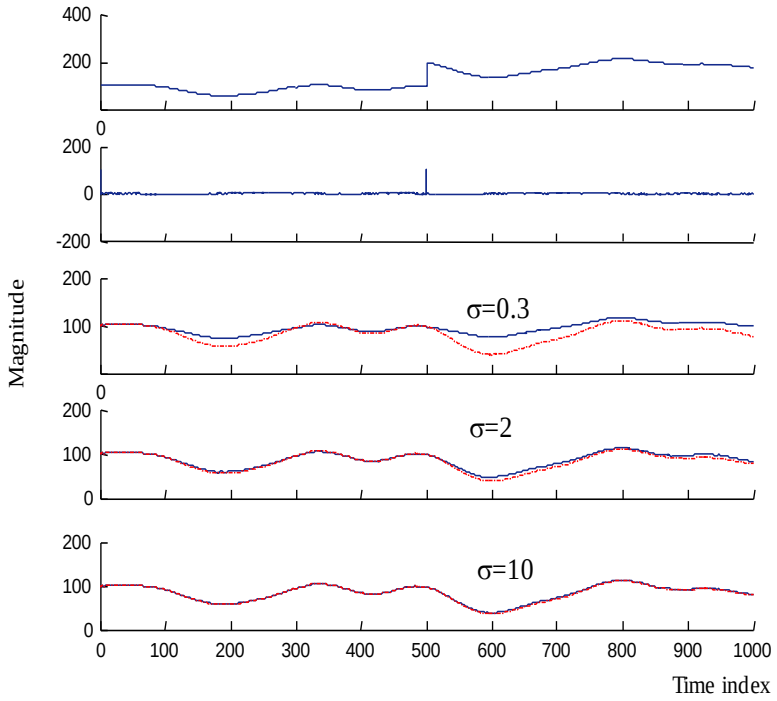


Figure 4.11 - Illustrate of the response of the variation of σ^2 of the speech signals with the dotted lines represent the actual curves.

4.5 SMOOTHING OF THE PITCH TRAJECTORIES

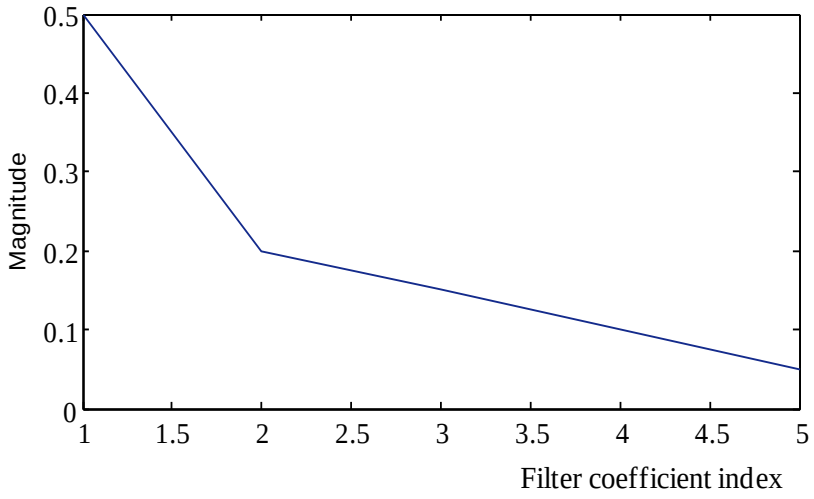
Moving Average Filter

Random fluctuations in pitch estimation can be reduced by a simple, relatively low-order, low-pass moving average (MA) filter [24],[107]**.** A moving average filter also known as a finite impulse response (FIR) filter is defined as

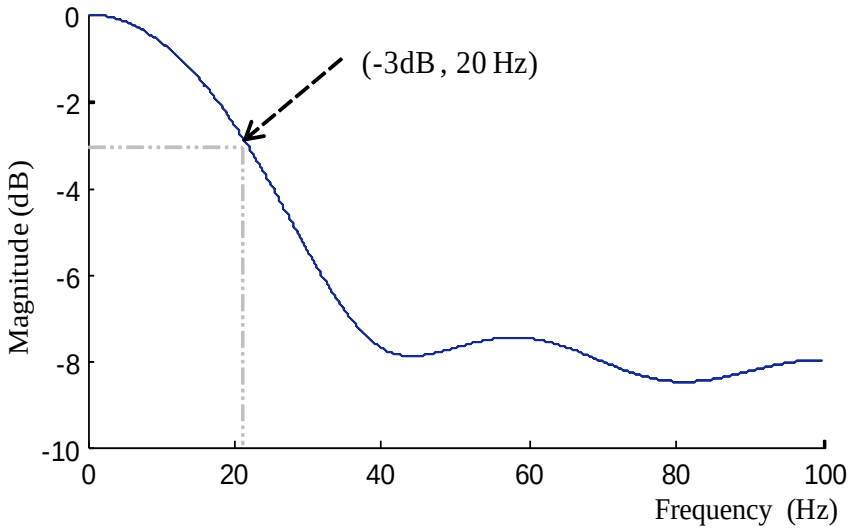
$$y(m) = b_0 x(m) + b_1 x(m-1) + \cdots + b_P x(m-p) \qquad (4.21)$$

Where $b = [b_0,\ b_1,\ \cdots b_P]$ are the coefficients of a filter of order P. In z-transform domain the MA filter can be expressed as

$$Y(z) = b_0 X(z) + z^{-1} b_1 X(z) + \cdots + z^{-P} b_P X(z) \qquad (4.22)$$



(a)



(b)

Figure 4.12 - Illustration of (a) the impulse response and (b) the frequency response of the moving average filters with coefficients vectors $b = [0.5, 0.2, 0.15, 0.1, 0.05]$.

A large number of choices are available for design of a such a filter such as the window design technique that employs the inverse Fourier transform of an idealized lowpass filter, or the Gaussian low-pass filter [106]. In this work a 5th order, one-sided, MA filter is used with the following coefficients vector, b,

$$b = [0.5, 0.2, 0.15, 0.1, 0.05] \tag{4.23}$$

Note the MA filter coefficients are chosen heuristically to give progressively less weights to the past samples as shown by the filter impulse response (note for MA filter impulse response is the coefficient set). The more distance a sample, the less weight it gets in the MA process.

Figure 4.12 (a-b) shows the impulse frequency response of the MA filter. The filter coefficients were chosen empirically.

4.6 CONCLUSION

The smoothness and the continuity of the time-variation of pitch trajectories may be optimized by considering the variation models, the removal of impulsive noise, and the removal of the step change within utterances and between utterances of the speech signals. The variation of pitch trajectories may be modelled using the probability of the finite-state model, or the linear prediction models. Several challenges such as impulsive noise, step change and smoothing commonly are the steps or techniques to improve the pitch trajectories in order to reduce the error of pitch estimation in speech processing.

The observations of many extracted pitch examples shows that of the three forms of estimation error (i.e., step change, impulsive noise and random noise), step changes and impulsive noise are the most significant and constitute the majority of pitch error. Pitch is generally a smooth process and hence smoothing random variations is not as much of an issue as that of the removal of large errors.

5

IMPACT OF WINDOW LENGTH AND MOMENT ORDER ON PITCH ESTIMATION

This chapter explores the impacts of the choice of the similarity criterion and the speech window segment length on pitch estimation. The similarity criteria explored are the second order moments and a set of modified moments including the modified second and higher order moments. The main finding is that the length of the window is by far the most dominant factor that affects the pitch estimation error. Various moment-based similarity criteria offer similar results while a modified second order method offers some robustness. All methods perform substantially better than the benchmark YIN method.

To obtain the modified higher order moment, MHOMs, each speech frame is split into a positive-valued and a negative-valued signal. The magnitudes of the HOMs for the positive and the negative valued signals are obtained separately and combined. HOMs form a sharper peak around the true pitch value compared to the correlation function.

The choice of the window length has a major impact on pitch estimation. For each criterion the variation of pitch error is obtained as a function of the window length. Depending on the moment criteria used a window size of 33 ms to 80 ms is optimal. To avoid excessive delay in real-time applications a two-stage method is proposed whereby an initial, coarse but robust, estimate of the pitch, from a longer window length, spanning the current and past speech frames is followed by fine-tuning within the current frame. This strategy imposes no additional delays.

The impact of the choice of the window length and the choice of similarity criterion are evaluated on a database of 10 male and 8 female speakers over a range of SNRs and noise types. For calculation of pitch errors, the pitch references are obtained from manually-corrected estimates obtained from laryngograph signals. The results for the second to fifth order moments are compared with magnitude difference criteria and the YIN method. The overall conclusion is that the HOMs provide similar performance to second order method with second order modified HOM being more robust than other methods. However, for each method very substantial improvement in pitch accuracy is obtained by selection of optimal window duration.

5.1 INTRODUCTION

Accurate estimation and smooth trajectory of the pitch, $T_0(m)$, or fundamental frequency $F_0(m)$, of a speech signal is a challenging task, especially in low signal-to-noise ratio, SNR, environments and also in dynamic channel conditions. Conventional pitch estimation methods are mostly based on the correlation (i.e. 2nd order moment) criterion, as the similarity measure for detection and estimation of the periodicity of speech signals. An alternative similarity criterion often cited in the literature is the average magnitude difference function (AMDF). Whereas the correlation criterion utilises the average product of two samples $x(m)$ and $x(m-\tau)$ as a measure of the similarity of the samples spaced by τ seconds, the AMDF uses the average difference between the samples as the similarity measure. This chapter investigates the use of higher order moments (HOMs) as an alternative to conventional pitch estimation criteria and explores the impact of the length of the signal window on the variations of the pitch estimation error.

For implementation on real-time systems, a two-stage method is proposed where the advantage of a robust coarse estimate obtained from a larger window, spanning a number of stored past speech frames and the current speech frame, is combined with fine-tuning over a short current window (i.e. spanning only the current speech frame).

5.2 MODIFIED HIGHER ORDER MOMENTS METHODS (MHOMs) AS PITCH ESTIMATION CRITERIA

Whereas autocorrelation (2nd order) based pitch extraction methods, utilise the average similarity between two samples, e.g. $x(m)$ and $x(m-T)$, the higher order moment

methods (HOMs) exploit the average similarity between three or more samples; .e.g. $x(m)$, $x(m-T)$, $x(m-2T)$ and so on [108]- [109].

For pitch extraction, where the intent is to estimate the period *T*, the general expression for calculation of the K^{th} order moment can be defined as

$$m_K(T) = \frac{1}{N-(K-1)T} \sum_{m=0}^{N-(K-1)T} [x(m)x(m-T) \dots x(m-(K-1)T)] \qquad (5.1)$$

where $K = 2, 3$ .... The contributions in the literature to the application of HOM method to pitch extraction are limited to a number of conference papers: [59] introduced the use of the higher order statistics where they can extract useful information of voiced frames and can separate speech from noise; [58]**,** [110] improve the reliability of the pitch estimation for unknown, periodic non-sinusoidal signals [111]**.**

5.2.1 Modified Higher Order Moments

The modified higher order moments (MHOMs) criteria are applied to the sample analysis of a periodic speech signal and the results are compared with conventional pitch extraction criteria. Pitch estimation methods based on MHOMs require a longer length of averaging window as the similarity of each sample with the corresponding samples up to (*K*-1) periods away is computed.

The impact of window length on the accuracy of pitch estimation and the implications of the use of a large window for practical implementation of the proposed pitch estimation methods on delay-sensitive communication systems are discussed and an appropriate solution for real-time communication systems is presented.

Conventional period/pitch estimation uses the average of the product of two samples (i.e. correlation or second order moment), $x(m)x(m-T)$, spaced at a distance of T as a measure of similarity or periodicity at T. At the values of the period T where $x(m)$ and $x(m-T)$ are similar, reinforcement of the periodic values occurs and hence a peak of the product $x(m)x(m-T)$ is observed. Theoretically, this idea of reinforcement of similar periodic samples can be extended to employ the average of the product of K periodic samples (i.e. K^{th} order moment), $x(m)x(m-T)\ldots.x(m-(K-1)T)$, as the similarity measure for a proposed value of the period T. The expected advantage gained is a greater degree of reinforcement of the product of K similar samples that theoretically may result in a sharper similarity criterion and hence less pitch estimation error as the moment order K increases. In reality the moment order, K, cannot be set to a value beyond five or six due to the non-stationary character of speech and pitch signals and because as K increases an appropriately larger averaging widow is required.

The general form of the equation for the MHOMs may be expressed as

$$m_K(T) = \frac{1}{N-(K-1)T} \sum_{m=0}^{N-(K-1)T} fn[x(m)x(m-T)\ldots x(m-(K-1)T)] \qquad (5.2)$$

where $fn[.]$ is some general function that may assume different forms. For this thesis a novel and particular form of $fn[.]$ termed the modified HOMs is obtained by splitting, rectifying, a signal $x(m)$ into a positive-amplitude $x_+(m)$ part and a negative-amplitude $x_-(m)$ part defined as

$$x_+(m) = \begin{cases} x(m) & if \quad x(m) > 0 \\ 0 & otherwise \end{cases} \qquad (5.3)$$

$$x_{-}(m) = \begin{cases} x(m) & if \quad x(m) < 0 \\ 0 & otherwise \end{cases} \tag{5.4}$$

Clearly, we have

$$x(m) = x_{+}(m) + x_{-}(m) \tag{5.5}$$

Note that using the electrical engineering terminology, $x_{+}(m)$ and $x_{-}(m)$ are obtained from positive and negative half-wave rectifications of the signal $x(m)$.

The K^{th} order modified moment is defined as

$$m_K(T) = m_{K_+}(T) + m_{K_-}(T) = \frac{1}{N_1}\left\{\sum_{m=0}^{N_1}\left(\prod_{l=0}^{K-1} x_{+}(m-lT) + \left|\prod_{l=0}^{K-1} x_{-}(m-lT)\right|\right)\right\} \tag{5.6}$$

where $m_{K_+}(T) + m_{K_-}(T)$ are the MHOMs of positive and negative parts of the signal and $N_1 = N - (K-1)T$, and $|.|$ is the absolute value operator.

For example the equation for the third order (K=3) moment is defined as the sum of

$$m_3(T) = \frac{1}{N-2T}\left\{\sum_{m=0}^{N-2T} x_{+}(m)x_{+}(m-T)x_{+}(m-2T) + \left|\sum_{m=0}^{N-2T} x_{-}(m)x_{-}(m-T)x_{-}(m-2T)\right|\right\} \tag{5.7}$$

The K^{th} order moments of a periodic signal are periodic. Hence, for estimation of the period, an energy maximizing function of the moments is defined as

$$E(T) = \frac{1}{N_T}\sum_{k=1}^{N_T} m_K\,(kT) \qquad T_{min} < T < T_{\max} \qquad (5.8)$$

where $N_T = fix(N/T)$ is the maximum number of periods that can be fitted in the function $E(\cdot)$. The estimate of the period T_0 is obtained as

$$T_0 = arg\ \max_T E\,(T) \qquad T_{min} < T < T_{\max} \qquad (5.9)$$

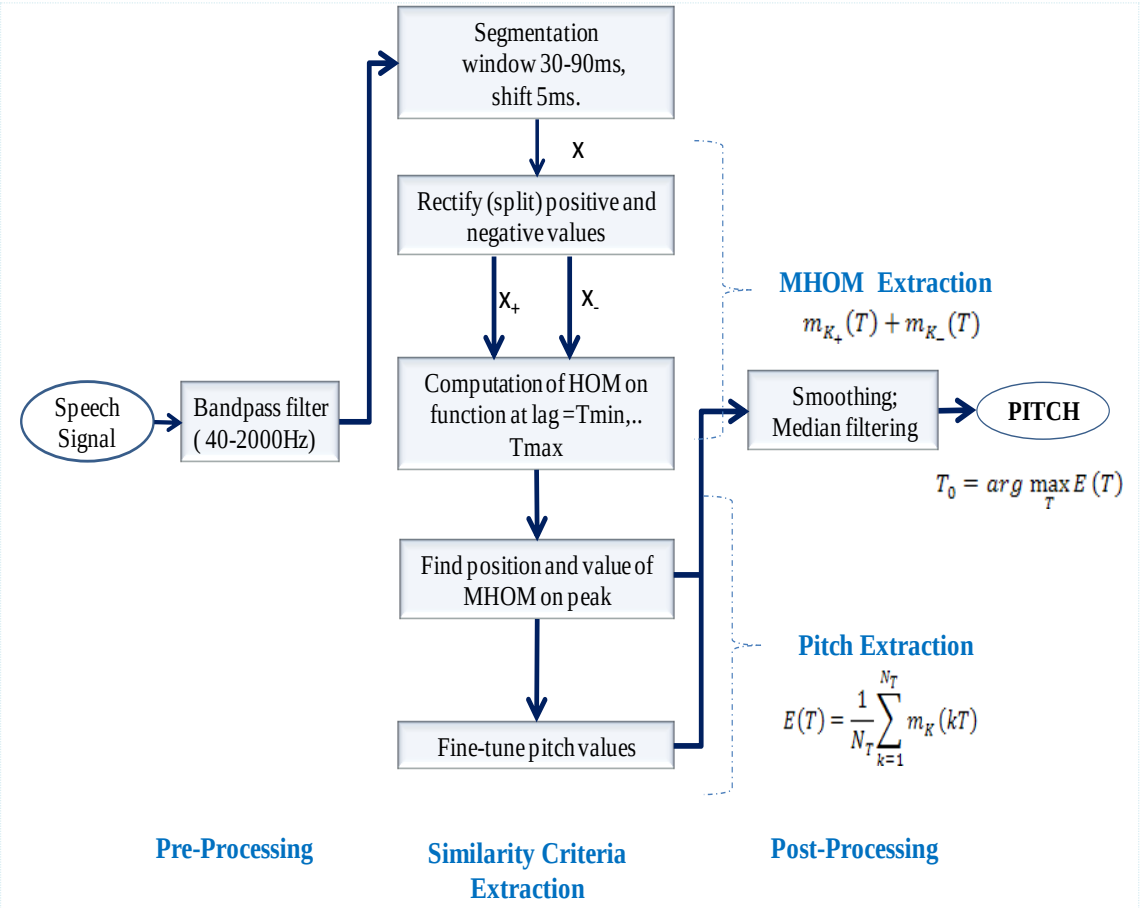


Figure 5.1 - Illustration of the general block diagram of a modified higher order moment (MHOM) pitch estimation system.

Figure 5.1 shows the general block diagram of the proposal modified higher order moment method of pitch estimation in time domain.

5.2.2 An Analysis of Modified Higher Order Moments

Figure 5.2, (top panel) shows a Hanning windowed segment of a periodic voiced speech together with a comparative illustration of the shapes of the similarity curve for various pitch extraction criteria namely the ACF, the AMDF and the third order, fourth order and fifth order MHOMs. The speech window is 400 samples long (i.e. 50ms at a sampling rate of 8 kHz). Note that for display purposes the AMDF curve has been inverted and zero-floored.

From Figure 5.2, it is evident that the modified HOMs forms a sharper peak around the true period value (T) and its integer multiples (kT) compared to the 2^{nd} order moment ACF and the AMDF functions. Furthermore, the peak of the third order moment at the correct period value T has a relatively higher value than the peak at $2T$ when compared to similar points of the correlation function and this relative difference is even more pronounced for the fourth order and fifth order moments.

Therefore, using MHOMs as period estimation criteria appear to promise more accurate period/pitch estimates with less large (e.g. double or half pitch) errors. A more in-depth experimental evaluation that validates this expectation follows in Section 5.6.

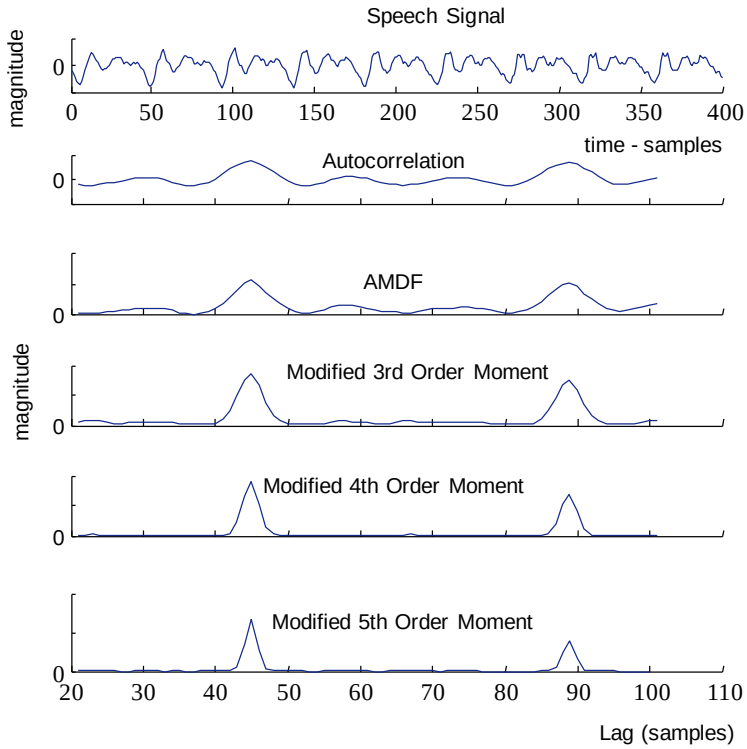


Figure 5.2 - Comparative illustration of the relative sharpness of ACF, AMDF, modified third order, modified fourth order and modified fifth order moment methods.

We postulate that the reason for the more concentrated and sharper shapes of the higher order moments is due to a greater degree of relative reinforcements that is obtained when the average of the products of three or more similar periodic samples (i.e. $x(m)x(m-T)\,x(m-2T)$ ...) are obtained.

5.3 THE EFFECT OF SPEECH WINDOW LENGTH

The choice of speech window length has a significant impact on pitch estimation. This is particularly the case for pitch estimation using MHOMs. It is easy to see that the higher the order K of the moments, the greater would be the $(K-1)T$ distance between the multiplied samples $x(m)$ and $x(m-(K-1)T)$ and hence by necessity a longer

averaging window is required for reasonably accurate and low-variance estimates of the higher order moments.

Four main considerations that influence the choice of the speech window length are:

1) The time-variations of the speech signals. Classical signal processing theory – such as Fourier transform, moment analysis, linear prediction models, etc. – assume that the signal processes are non-stationary; for this reason the window length should be short enough such that within the window the signal parameters, such as the pitch, remain approximately stationary [17], [112] - [113]. In most speech processing systems, speech is assumed stationary for about 20ms. However, note that there are differences in the degree of stationarity of different parameters of speech. For example, pitch is extra-segmental parameters that generally vary at a slower rate depending on the speaker and the style of speech. Extra segmental parameters are parameters that span relatively large segments composed of several frames of quasi-stationary speech.
2) The maximum allowable delay in voice communication. Due to the inevitable delay incurred in signal propagation through communication networks and channels, the processing delay, including the window length, should be kept at a minimum possible value. The maximum allowable delay in communication systems is about 300 ms and mobile communication systems strive to keep speech frame (window) delay within 20 ms [9].
3) The variance of the estimate of a moment decreases inversely with the number of samples used in the averaging process.
4) The accuracy of pitch estimation and the variance of pitch estimation error. For accurate estimation of stationary signal processes, the averaging window length

should be as large as possible; however this conflicts with the constraints on stationarity and delay explained in (1) and (2).

From the estimation theory, specifically the Cramer-Rao lower bound, the variance of estimation error decreases with the increasing observation length [24]. Hence, as expected the choice of the speech window (or frame) length has a substantial influence on the variance of the pitch error. Generally, pitch estimates improve with the increasing speech window length within a voiced segment of speech. However, if the window length is too large, the pitch estimate will not accurately follow the smooth variation of the pitch utterance curves and will give rise to a coarse estimate of the pitch curve that has a step-wise shape.

We suggest that a combination of a coarse but robust pitch estimate obtained from a longer speech window length, spanning the current and several past speech frames, followed by fine tuning over the current speech frame can be used to have a good advantage without imposing any additional delay and without compromising the delay constraint of the voice communication systems, hence, the following two-stage approach is utilized:

1) Initially a larger window length of N_L samples, based on a concatenation of the current speech frame of length N_S samples with $N_L - N_s$ samples extracted from the adjacent immediately preceding (and possibly overlapping) speech frames, is used to obtain a robust *gross* estimate of the pitch value F_0.

2) The pitch estimation process is repeated over a shorter window length, composed of the N_S samples of the current speech frame, to obtain a locally optimised *fine*

estimate of the pitch constrained around the current values of F_0 in the range of $F_0 \pm aF_0$, where $0 < a < 1$. Typically $a = 0.1$.

The Practical Implication of Using a Longer Window in Real-Time Applications

In this section it is briefly explained that the two-stage multiple window method proposed in this work can be implemented in real-time for current communication systems without imposing the prohibitive cost and limitation of an additional delay. To achieve this for every speech frame $\boldsymbol{x}(m)$ of length *N*-samples, $x(m) = [x(m) \dots x(m - N - 1)]$, a longer speech window of length $N(P + 1)$ speech samples may be obtained via concatenation of the current speech frame and *P* stored previous frames as $[x(m) \dots x(m - P)]$ as shown in Figure 5.3. In this way at any time there are two concurrent speech windows; the relatively short current speech frame and the longer asymmetric speech window spanning the current and past frames; these two windows can be used to implement the two-stage pitch estimation method proposed here. The asymmetric window has more weight on the current speech frame than previous frames. Note the main additional costs are the extra memory required to store *P* past frames and the additional processing time for calculation of MHOMs and extraction of pitch from a longer window.

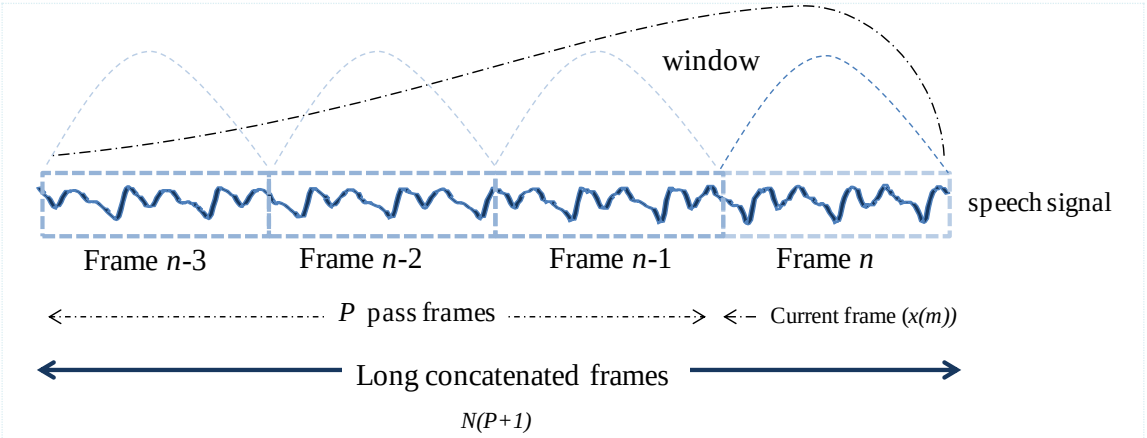


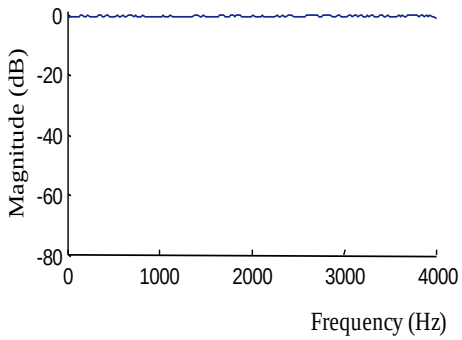
Figure 5.3- Illustration of the concatenation of frames to form larger segments for pitch estimation.

5.4 DATABASE OF SPEECH AND REFERENCE PITCH SIGNALS FOR EVALUATION OF RESULTS

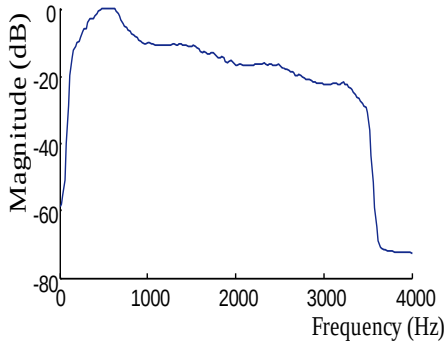
This section describes the databases of speech, reference pitch and noise, employed for evaluation of pitch extraction results. The method of extraction of the reference pitch values from the laryngeal signals is described.

5.4.2 Additive Noise Types

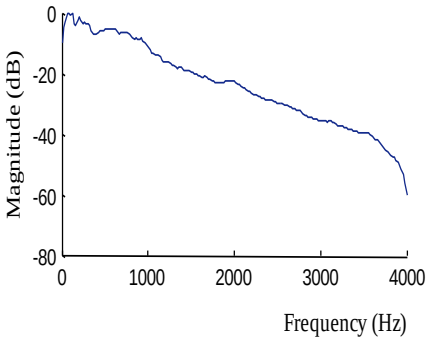
In order to investigate the robustness of the proposed criteria, four types of additive noise namely; Gaussian white noise, car noise, train noise, and babble noise are used [114]. The power spectra of the noise are shown in Figure 5.7. Gaussian white noise was generated using the MATLAB routine. Car noise and train noise were recorded by researchers at Brunel University, whereas the babble noise was obtained from publicly available data [115].



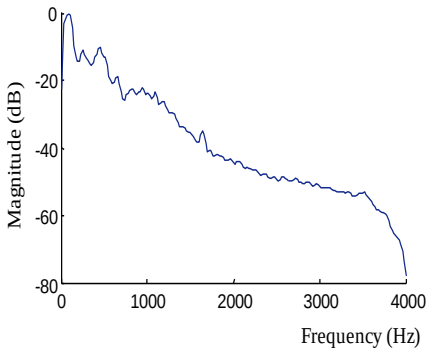
(a) Power spectra of white noise



(b) Power spectra of babble noise



(c) Power spectra of train noise



(d) Power spectra of car noise at motorway environment.

Figure 5.4 - The power spectra of the four types of noises used for evaluations.

The range of the variations of noise power from low frequency to 3.5 kHz is 0 dB (i.e constant power) for white noise, 30 dB for babble noise, 40 dB for train noise and 50 dB for car noise. Hence the noise with the most spread energy across frequency is the white noise followed by babble noise, then train noise and lastly car noise.

Note that car noise, babble, noise and train noise, in that order, have the largest coincidence of spectral energy with those of the fundamental frequency and the first and second harmonics of speech. Noting that since the fundamental frequency and the first two harmonics are usually crucial for accurate pitch estimation, it is expected that at a given overall signal to noise ratio, comparatively, car noise degrades the pitch estimation accuracy most followed by babble noise, train noise and then white noise.

All noise signals like speech, were lowpass filtered and band limited to 4 kHz and resampled at 8 kHz sampling rate for evaluation.

5.4.1 Speech Signals and Laryngograph Signals Databases

The raw databases used for the evaluation of the experimental results are publically available simultaneous recordings of speech and laryngograph signals [116] - [120].

(1) Speech Signals Databases

The speech databases used in this work, for the extraction of reference pitch values and the evaluation of pitch extraction methods, contain gender-balanced and phonetically-balanced (i.e. having approximately the same statistics of occurrence of phonemes as those that of a very large database) utterances from 10 male speakers with a total of 304 utterances and 8 female speakers with a total of 155 utterances with total durations of 1048 and 597 seconds respectively. All speech signals were band-limited to 4 kHz and resampled at 8 kHz which is the standard bandwidth and sampling rate used for mobile phones as shown in Table 5.1.

Table 5.1 - The databases of the speech and laryngeal signals.

Gender	Number of speakers	Number of utterances	Mean Duration of utterance (sec)	Duration (sec)	Bandwidth, Sampling Rate
Female	8	155	3.85	597	4 kHz, 8 kHz
Male	10	304	3.45	1048.2	4 kHz, 8 kHz

(2) Extraction of Reference Pitch from Laryngograph Signals

Laryngograph or electroglottograph (EGG) signals are records of the vocal folds phonatory vibrations during voice production. The laryngograph signals has been used as the "ground truth" or reference pitch values in our evaluation. The laryngograph signals provide a relatively clean recording of the glottal vibrations from which the speech signal period and its inverse, the pitch, can be extracted with a high degree of accuracy [121].

The EGG signals measure the variations of the contact area of the vocal folds. To measure the vibrations of vocal folds contact area, an EGG records the pattern of variations in the transverse electrical impedance of the larynx and nearby tissues by means of a small A/C electrical current in the mega Hertz region applied by electrodes on the surface of the neck. This electrical impedance will vary slightly with the area of contact between the moist vocal folds during that part of the glottal vibratory cycle in which the folds are in contact.

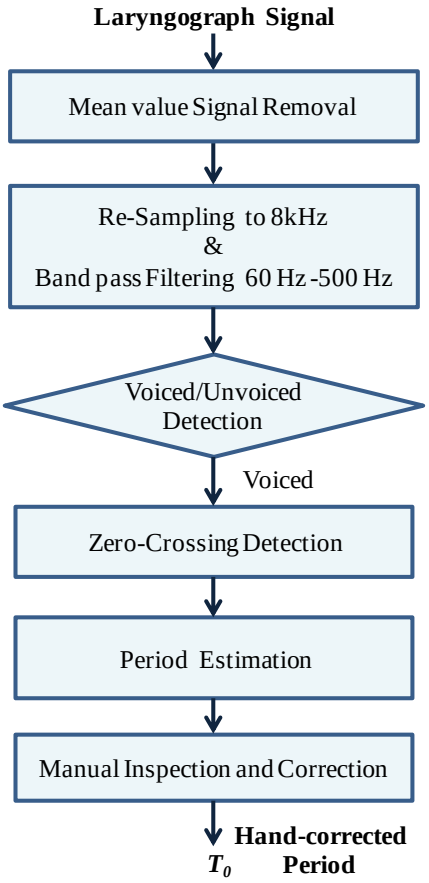


Figure 5.5 - The flowchart for period estimation (pitch marking) from laryngograph signal using zero-crossing method.

However, because the percentage variation in the neck impedance caused by vocal fold contact can be extremely small and varies considerably between subjects, no absolute measure of contact area is obtained, only the pattern of variation for a given subject [122].

(3) Pitch Marking of the Laryngograph Signals

Figure 5.5 illustrate the pitch marking process of laryngograph signals. Pitch marking refers to a process of, marking regularly spaced points, corresponding to period or

frequency of repetition (pitch), in time or frequency domain representation of the signal. The time variations of the period/pitch are derived from the pitch mark points.

A zero-crossing method was used for the initial marking and extraction of the period information from the laryngograph signal [123]. First the mean of the laryngograph signal is removed to obtain a zero-mean signal. A zero crossing is defined as when the magnitude of the zero-mean signal changes sign from a positive value to a negative value or vice-versa. Note that each period of a zero-mean signal is composed of two zero crossings, one from positive to negative and the other from negative to positive as shown in Figure 5.6 and Figure 5.7.

The period is calculated as the total number of samples, between the zero-crossing at the beginning of a period and the zero crossing at the end, multiplied by the sampling period.

All the extracted period data were visually inspected and hand-corrected where necessary. The pitch values extracted from the voiced activity of the laryngograph signals for the reference (or 'ground truth') values for various pitch extraction methods evaluated in this section.

The estimates of pitch from the laryngographs signals and from speech signals are performed at 5 ms intervals, which correspond to a window slide of 40 samples at 8 kHz sampling rate. The choice of 5 ms sampling interval is a standard that is also employed as the pitch update rate for mobile phone voice coding [9]. Note 5 ms sampling corresponds to a sampling rate of $F_S = 200$Hz and a maximum frequency of $\frac{F_S}{2} = 100$ Hz. This is sufficient for capturing the variation of pitch trajectory i.e. intonation.

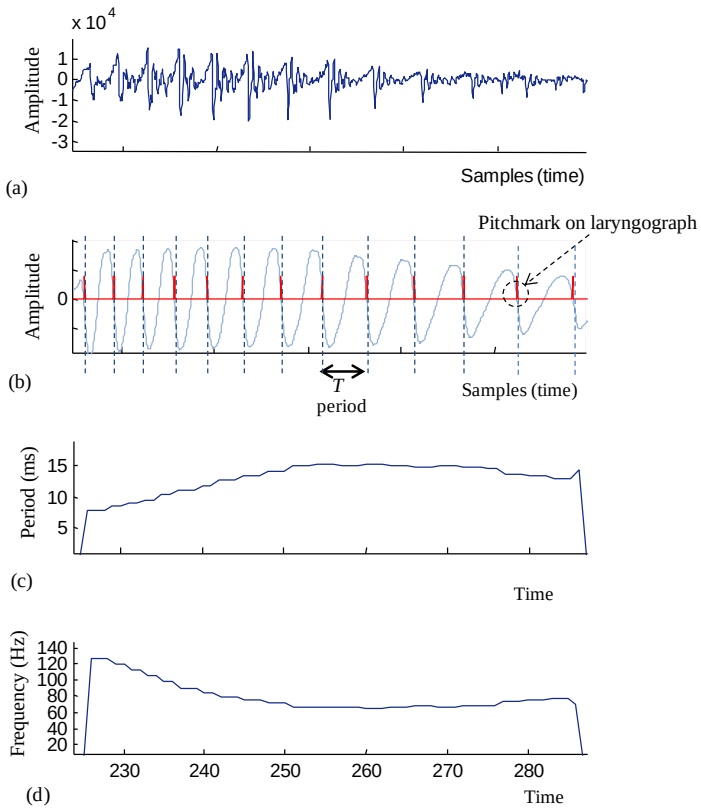


Figure 5.6 - (a) The speech signal, (b) the pitch marking on laryngograph signal, (c) the down-sampled period, and (d) the pitch or fundamental frequency of the male speaker.

Figure 5.6 and 5.7 show two examples of speech and the corresponding laryngograph signal, pitch marks, period curve and pitch curve for a male speaker and a female speaker. Given the visual inspection and hand correction process, the inspected/corrected reference pitch are highly accurate.

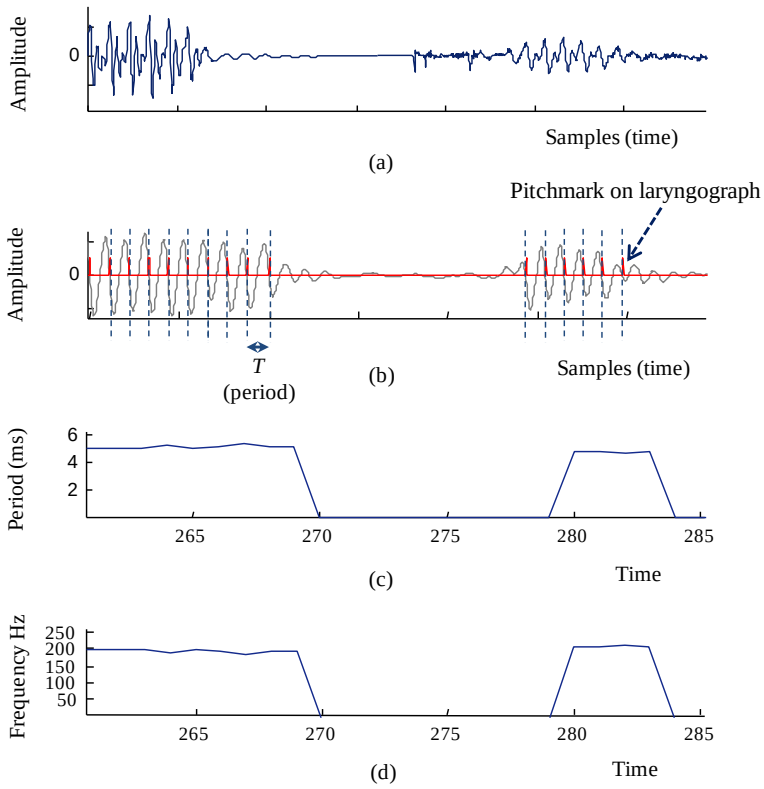


Figure 5.7 - (a) The speech signal, (b) the pitch marking on laryngograph signal, (c) the down-sampled period, and (d) the pitch or fundamental frequency of the female speaker.

5.5 EXPERIMENTAL EVALUATION AND DISCUSSION

In this section a set of experiments are performed to evaluate and compare the performance of the proposed pitch extraction methods with several well-established methods namely the autocorrelation function, ACF [17], [22] – [23], [112]- [113], [124]- [125]; the AMDF method [38], [55] – [57], [126] - [132], and the benchmark YIN method [34] and HOMs (3^{rd}, 4^{th} and 5^{th}) methods [58]-[59], [110]- [111]. The effects of different types of noise on the performance of pitch extraction methods are examined. The experiments are conducted on several databases with different noise types and in a range of signal to noise ratios. The details of the experimental setup are described next.

5.5.1 Pitch Error Analysis Method

For error measure, the average percentage absolute value of pitch estimation error is defined as

$$E = mean\left(\frac{\left|\hat{F}_0(m) - F_0(m)\right|}{F_0(m)}\right) \times 100 \quad \% \tag{5.10}$$

where $F_0(m)$ and $\hat{F}_0(m)$ are the true (aka 'ground-truth') value (obtained from manually-corrected laryngographs) and the estimate of pitch, for the m^{th} speech frame, respectively. Note that the pitch error is calculated over voiced frames only. The voicing information is reliably obtained from the laryngeal signal and visually inspected and manually corrected.

The choice of the average % of absolute pitch error, as opposed to other error measures such as mean square error etc, conforms to the normal practice employed in other works [17], [126], [133]- [134].

For analysis of pitch accuracy six categories of mean percentage absolute value of pitch error are considered:

1) E_{Total}; the average percentage of overall absolute value of pitch errors for all values of pitch error large and small, Equation (5.10).
2) E_{Fine} the average percentage of small absolute value of pitch errors that are less than or equal to 20%, called fine percentage error (FPE),

$$E_{Fine} = mean\left(\frac{\left|\hat{F}_0(m) - F_0(m)\right|}{F_0(m)} \leq 0.2\right) \times 100 \quad \% \tag{5.11}$$

3) E_{Gross} the average percentage of large absolute value of pitch errors that are greater than 20%, called gross percentage error (GPE)

$$E_{Gross} = mean\left(\frac{\left|\hat{F}_0(m) - F_0(m)\right|}{F_0(m)} > 0.2\right) \times 100 \quad \% \tag{5.12}$$

4) *Var*; variance of pitch errors for all values of pitch error large and small.

5) The percentage of population of errors, P_{Fine}, for small value of pitch error, $P_{Fine} < 20\%$.

6) The percentage of population of errors, P_{Gross}, with large value of pitch errors $P_{Gross} \geq 20\%$.

Note that is $E_{Total} = E_{Gross} + E_{Fine}$.

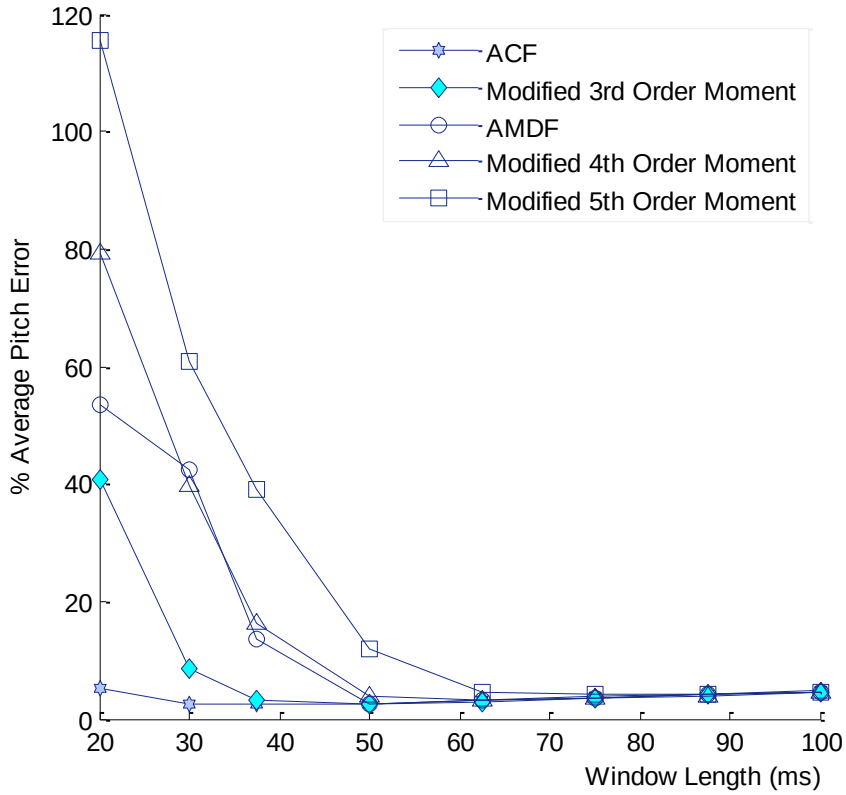
The motivation for splitting the pitch estimation errors into gross and fine errors is to provide an indication of robustness of the pitch estimator. Generally a robust system will yield less outlier and large errors including double and half pitch errors.

The choice of a threshold value of 20%, dividing the boundary between small and large errors, is arbitrary, however, this division is also used by other researchers [34], [71] to assess the tendency of a pitch estimation method to produce gross errors including half and double pitch estimates.

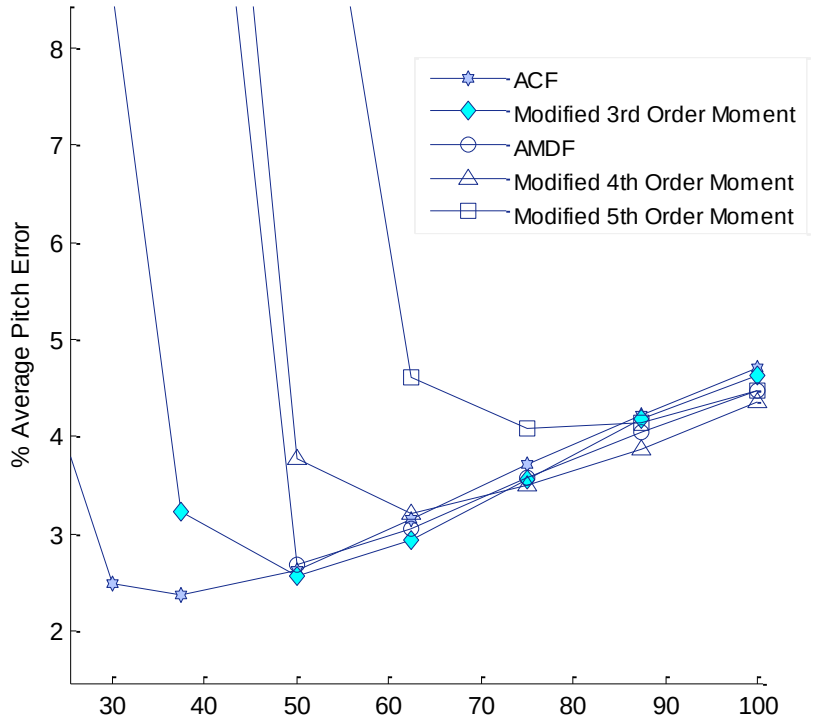
For pitch error evaluation purposes, the voicing detection is based on the laryngograph signals whose high signal to noise ratio recording process allows very accurate voiced/unvoiced detection. The voicing detections are visually inspected and manually corrected where necessary.

5.5.2 Analysis of the Effect of Varying Window Length on Pitch Estimation Error

An experiment was conducted to determine the variations of the pitch estimation error with the variation of speech window length and hence to select an optimal value of speech window for each of the MHOM and the second order methods [135].



(a)



(b)

Figure 5.8 - The mean of (%) pitch error versus speech window lengths for clean speech signals (30dB SNR): (a) widow length varying from 20 ms to 100 ms windows length and (b) widow length zoomed in 30 ms to 100 ms windows length.

Figure 5.8 Illustrates the variations of pitch estimation error, for different methods, with a range of window lengths of: 20ms (160 samples at a sampling rate of 8 KHz), 30ms (240 samples), 37.5 ms (300 samples), 50ms (400 samples), 62.5 ms (500 samples) ,75ms (600 samples), 87.5ms (700 samples) and 100 ms (800 samples).

The overall trend of variations of the pitch estimation error with the increasing window length is an initially steep reduction in the pitch error rate which levels off and slightly increases beyond the minimum error point. Since the statistical (moments) theory assumes that within the observation window the pitch signal is stationary, as the window length increases there is a point (around 30 - 40 ms for ACF) where the pitch error starts to

increase due to the considerable time variations of the actual pitch within the large time window.

Figure 5.8, suggests that the pitch estimation method based on the modified third order moment criteria provides the least error value for a window length of 50 ms and slightly more error for longer window lengths. The fourth and fifth order methods performs best for a window length of greater than 60 ms and the pitch error levels off at a minimum error value around a window length of 87 ms.

A further interesting point is the behaviour of the correlation-based pitch estimation method compared to the benchmark YIN method as shown in Figure 5.9. As the window length increase the curve of pitch error of the correlation method crosses that of YIN method and the correlation method significantly outperforms YIN. This underscores the importance of the influence of the window length as a dominant factor in pitch estimation. Increasing window length alone can outperform the combined effect of all the optimizations steps employed in the YIN method.

When estimating periods it is useful to place limits on maximum and minimum values of the period. For minimum and maximum values of period, T_{min}, T_{max} and pitch F_{min}, F_{max}, the following values were chosen as tabulated in Table 5.2.

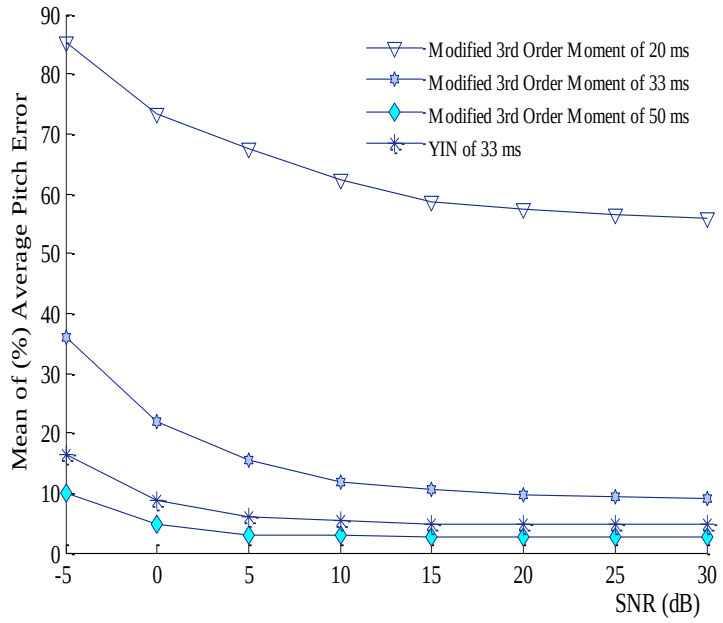
Table 5.2 - The limitation of period and fundamental frequency for the evaluation

Period	$T_{min} =$ 2.5 ms	$T_{max} =$ 25
Fundamental Frequency	$F_{max} = 1/T_{min} =$ 400 Hz	$F_{min} = 1/T_{max} =$ 40 Hz

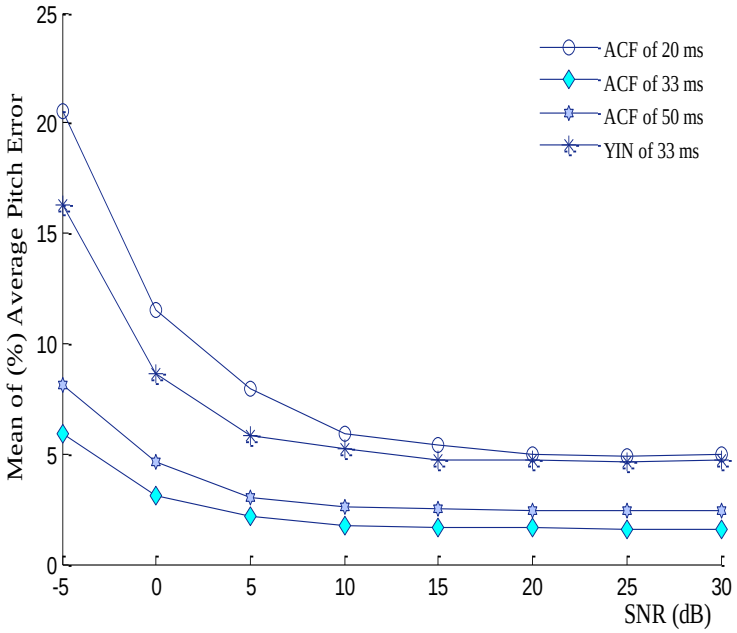
These limits can be varied. Based on the results, as explained in section 5.3 we employed a two-stage pitch estimation strategy whereby in the first stage a large window length of 50 ms (400 samples at 8 kHz sampling rate) is employed for the pitch estimation method based on the third order modified moment and a window length of 87.5 ms (700 samples) is employed for the pitch extraction methods based on the fourth and fifth order modified moments. Note that for a delay sensitive communication system these larger speech windows may include a number of stored past frames so that the system will not incur additional delays. The relatively large window length employed at the first stage of pitch estimation provides a coarse but robust initial estimate. In the second stage of the pitch estimation process, the coarse estimate, obtained from the first stage, is fine-tuned in the locality of the current short frame of 160 (20 ms) sample.

5.5.3 Analysis of Performance of Pitch Extraction Methods in Noisy Environments

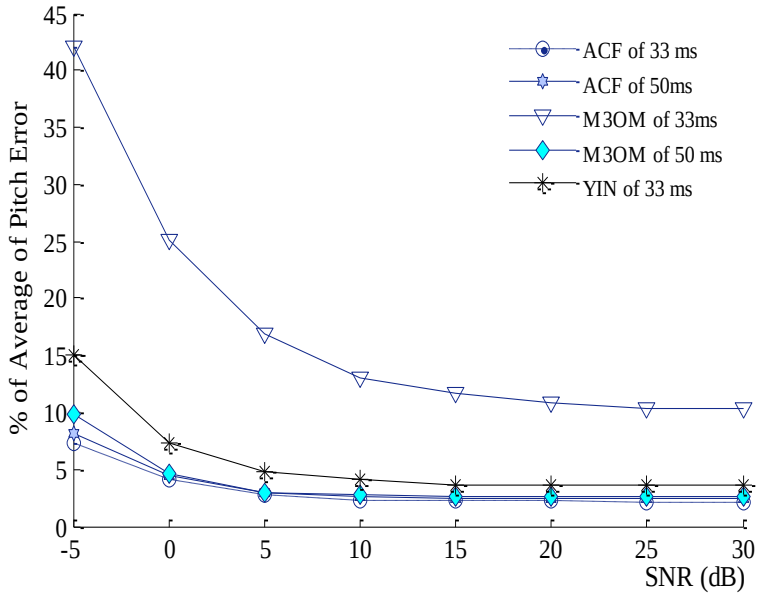
The pitch estimation methods were evaluated in a range of signal to noise ratios from 30dB down to -5dB with a SNR step size of 5dB. The noisy speech samples were obtained by adding several common types of noise; Gaussian white noise, car noise, train noise, and babble noise to clean speech signal **[**136**]**.


Modified 3rd Order Moment of 20 ms
Modified 3rd Order Moment of 33 ms
Modified 3rd Order Moment of 50 ms
YIN of 33 ms
Mean of (%) Average Pitch Error
SNR (dB)
90
80
70
60
50
40
30
20
10
0
-5
0
5
10
15
20
25
30

(a)


ACF of 20 ms
ACF of 33 ms
ACF of 50 ms
YIN of 33 ms
Mean of (%) Average Pitch Error
SNR (dB)
25
20
15
10
5
0
-5
0
5
10
15
20
25
30

(b)



(c)

Figure 5.9 - The % overall pitch error for a different windows length (20 ms, 33 ms, and 50 ms) as a function of SNR of Gaussian white noise; (a) Modified third order moment method , and (b) autocorrelation function method with YIN method, and (c) the ACF, modified third order moment, and YIN with the window length of 33 ms and 50 ms.

For these experiments the performance of two pitch extraction method based on autocorrelation and modified third order criteria, were evaluated for windows of 20 ms, 33 ms and 50 ms with speech contaminated with white Gaussian noise in a range of SNRs from 30 dB down to -5 dB. The results plotted in Figure 5.9 (a-c) shows a definitive pattern of improvement in the pitch estimation accuracy with the increasing window length at all SNRs.

Again, the most important finding from Figure 5.9 (a-c) is that increasing the signal widow length, for correlation or modified third order moment methods, can alone outperform all the various optimisation steps employed in YIN; this \underscores the importance of the length of the window as a dominant factor.

Figure 5.10 to Figure 5.13 provide a comparative analysis of the percentage of the overall pitch estimation error for the proposed modified HOM methods and the conventional pitch extraction (i.e. ACF and AMDF) and the YIN methods. The evaluations are performed for a range of signal-to noise ratios, in the range of -5 dB to 30 dB and the four aforementioned types of noise. Note that car, train and babble noise, due to a greater concentration of their power at low frequencies, where the fundamental frequency of speech resides, result in significantly larger pitch errors compared to white noise. For white noise, the second, the third and fourth order moment methods result in almost 50% less error compared with the bench mark YIN method as shown in Figure 5.10. For other noise types (i.e. car noise, train noise and babble noise) also the second to fourth order moments consistently yield less pitch errors than the YIN method. However, the fifth order method does not perform as well as the second to fourth order moments; we postulate this is due to a limit reached, in that for the fifth order moment the distance of four times the period between the samples, $x(m)x(m-T)\ x(m-2T)x(m-3T)x(m-4T)$ is too large and significant changes in pitch and speech signal occurs for distance of $4T$ and beyond.

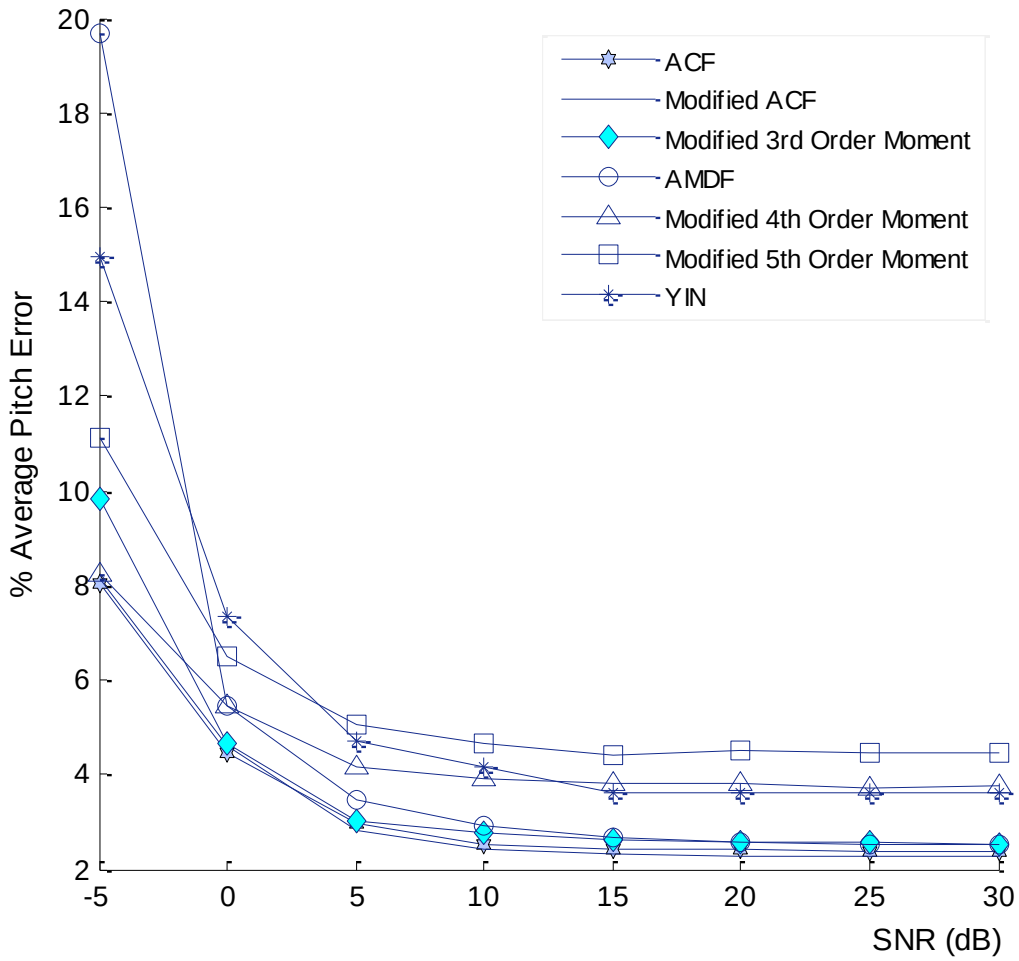


Figure 5.10 - The % overall pitch error of Gaussian white noise as a function of SNR

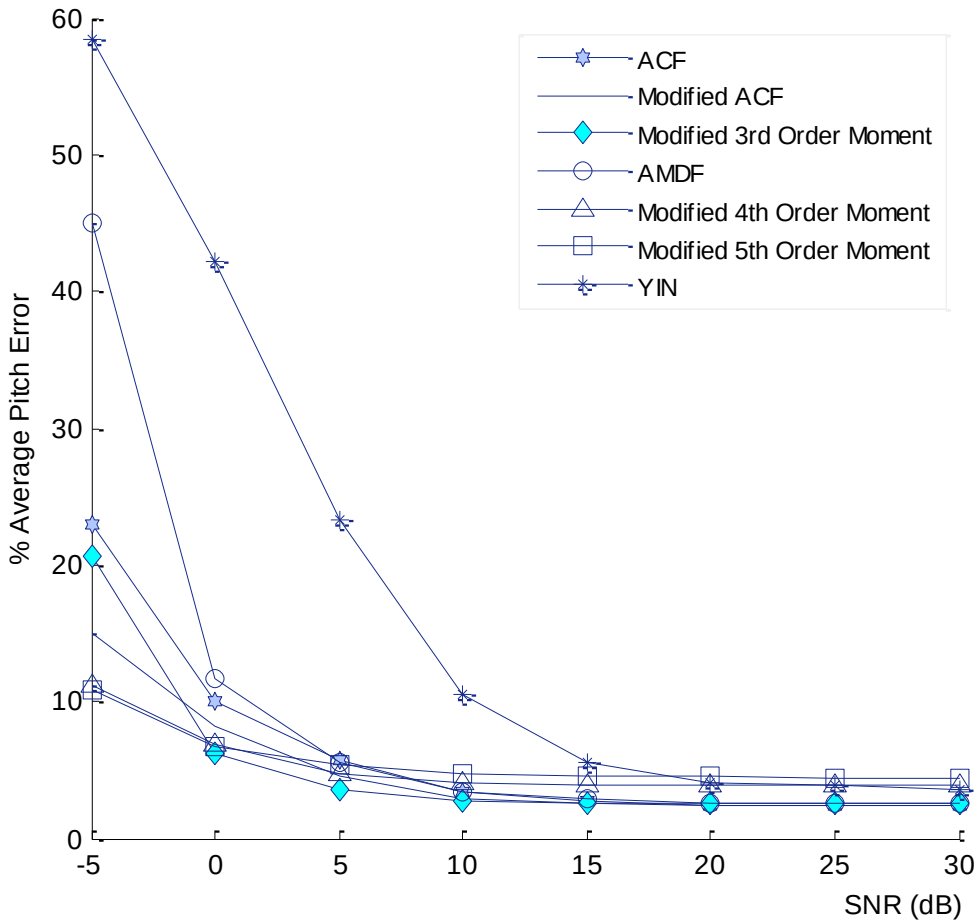


Figure 5.11 - The mean of overall (%) of pitch error of car noise as a function of SNR

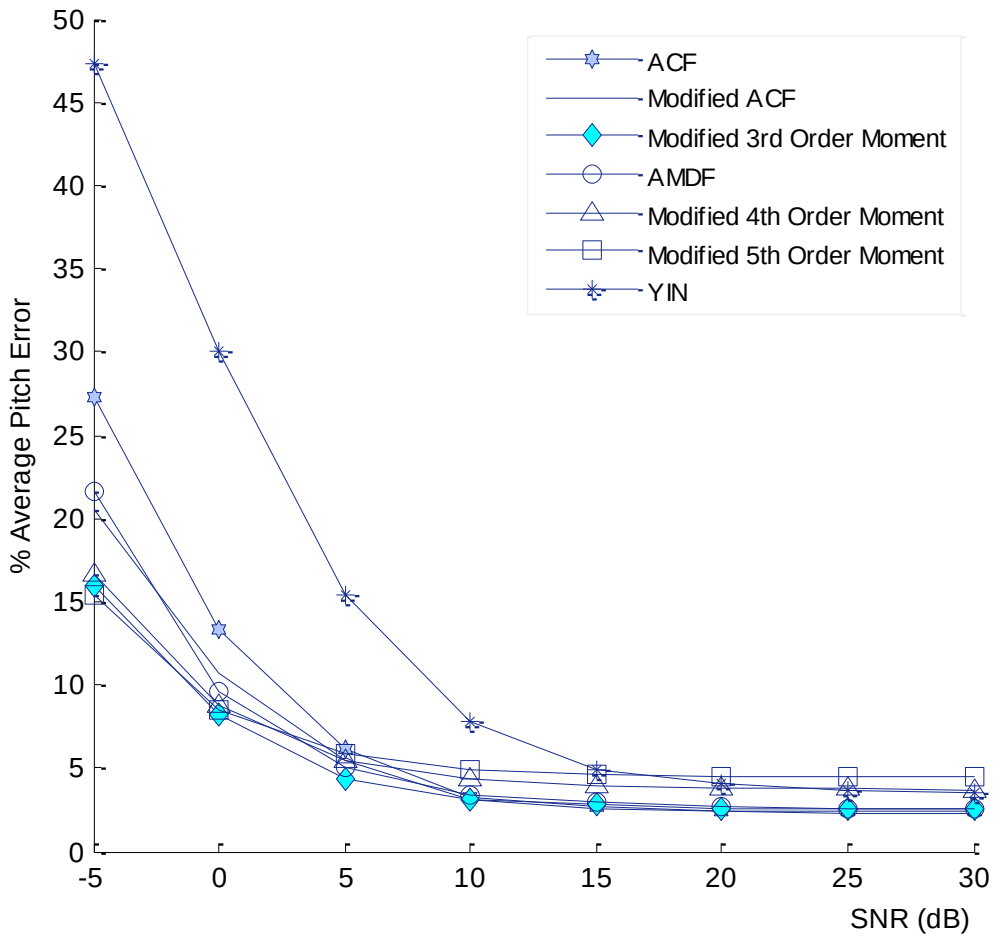


Figure 5.12 - The mean of overall (%) of pitch error in train noise as a function of SNR

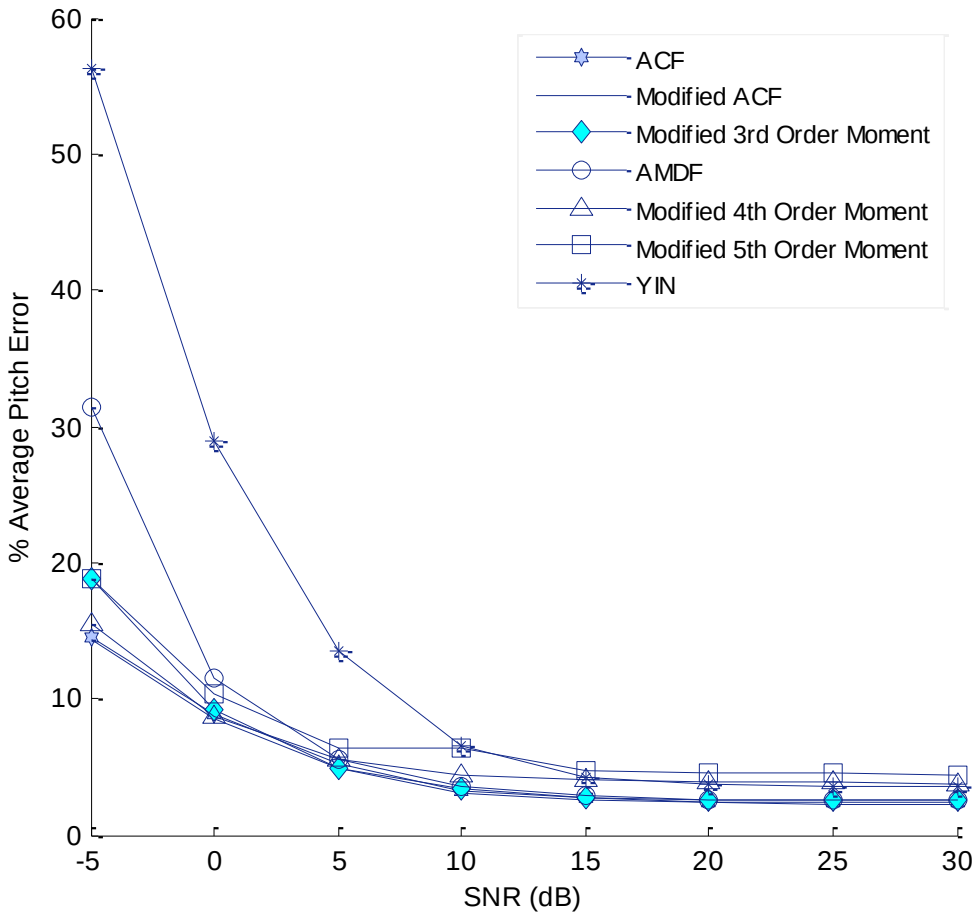


Figure 5.13 - The mean of overall (%) pitch errors of babble noise as a function of SNR.

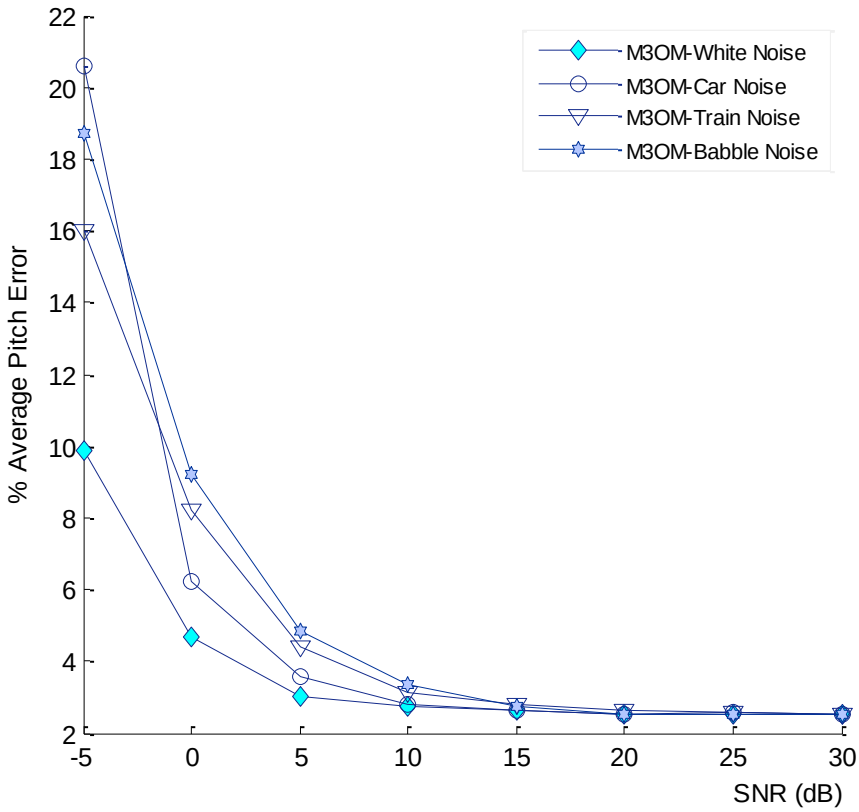


Figure 5.14 - Comparative pitch estimation error of the modified third order moment for four types of noise the function of SNR.

Figure 5.14 illustrate the comparison of pitch estimation error for the modified third order moment method with four different types of noise.

5.5.4 Analysis of the Variance of Pitch Errors

Figure 5.15 shows the plots and the values of the variances of the overall pitch errors for different criteria for the range of SNR from -5dB to 30dB for white noise. The proposed methods based on third and fourth order moments methods display less variance and compete well with the conventional ACF method and the YIN pitch extraction method.

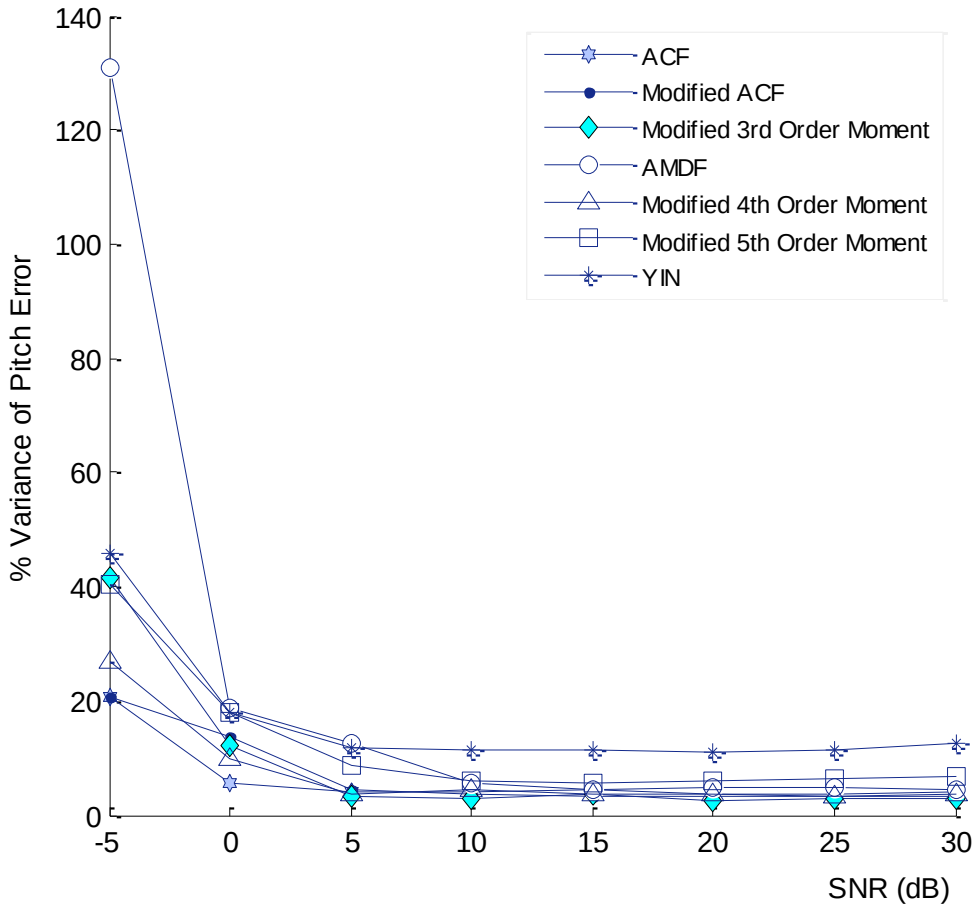


Figure 5.15 - Illustration of the variance of the pitch error of Gaussian white noise.

5.5.5 Analysis of the Weighted Average Fine and Gross Pitch Errors

For more detailed pitch error analysis, the percentage pitch errors are divided into two categories; the population weighted fine pitch errors defined as

$$E_{fine}^{\text{weighted}} = P_{fine}E_{fine} \qquad 5.13$$

Where P_{fine} is the percentage population of fine errors and $E_{fine}^{\text{weighted}}$ designate the weighted percentage pitch errors less than or equal to 20% (FPE) as shown in Figure 5.16.

The population weighted gross pitch errors defined as

$$E_{Gross}^{\text{weighted}} = P_{Gross} E_{Gross} \qquad 5.14$$

Where P_{Gross} is the percentage population of gross errors and $E_{Gross}^{\text{weighted}}$ designate the weighted percentage pitch error greater than 20% (GPE) as shown in Figure 5.17. The choice of the threshold of 20% as a dividing line between small and large pitch errors is arbitrary but this value of threshold is also used by other researchers [34], [124].

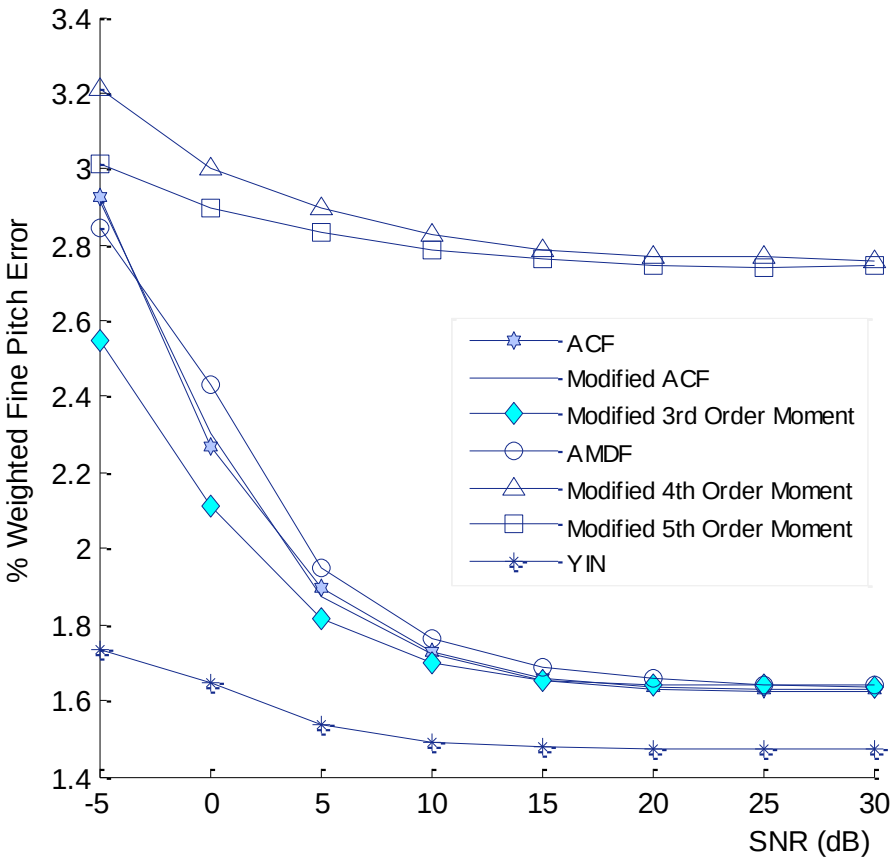


Figure 5.16 - The % weighted fine pitch error of Gaussian white noise as a function of SNR.

The evaluation results plotted in Figure 5.16 and Figure 5.17 display the weighted average pitch error for fine and gross pitch error respectively. It is evident that the combination of an optimal widow length and MHOMs criteria achieve a distinct improvement in terms of the accuracy of relative to the benchmark YIN for both fine and gross pitch evaluations which resulted in smaller error values in a range of -5dB to 30dB SNR.

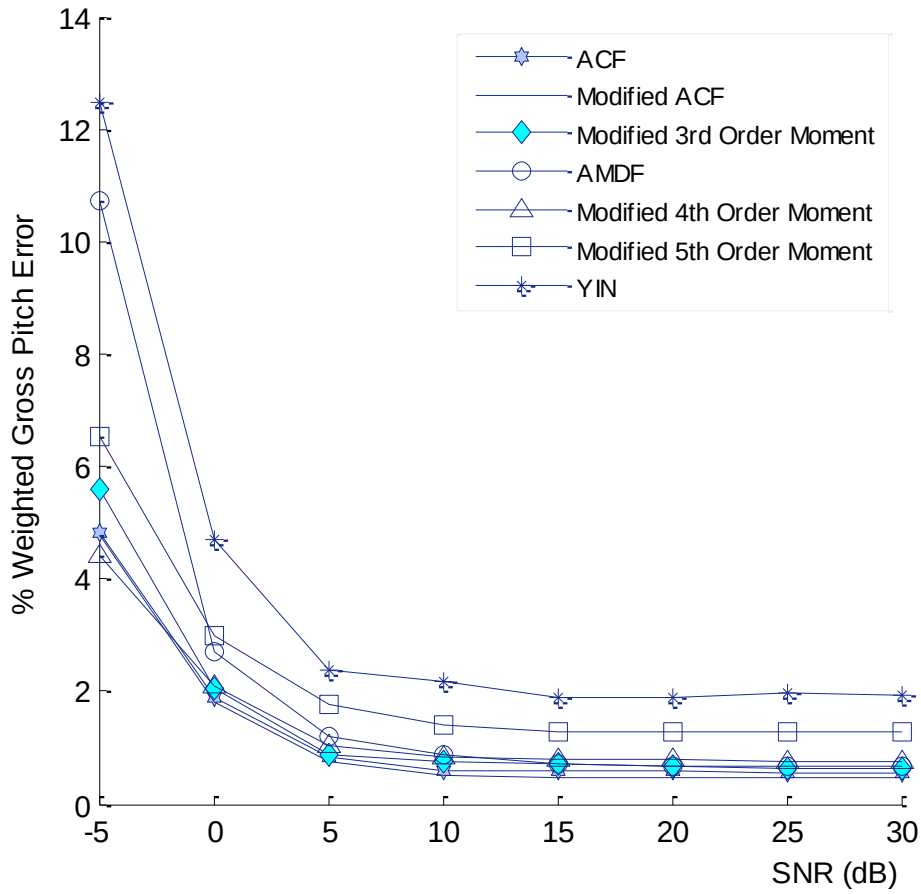


Figure 5.17 - The percentage weighted gross pitch error of speech in Gaussian white noise as a function of SNR.

5.5.6 Analysis of the Population of Fine and Gross Pitch Errors

Figure 5.18 and Figure 5.19 are the plot of the percentage population of gross and fine pitch errors, using various pitch extraction methods, respectively. In Figure 5.19, the pitch extraction based on ACF and MHOMs result in a significantly lower percentage of population of the gross pitch errors (that are greater than 20%) compared to the bench mark YIN methods. Conversely, as shown in Figure 5.18, the ACF and MHOMs methods yield a relatively higher percentage of population of pitch error that are less than 20%. The results indicate that while the pitch extraction methods based on the third and fourth order modified moments provide a competitive overall average pitch error, they are also robust in that they have a relatively lower proportion of large pitch errors; low overall error and fewer occurrence of large errors are two desirable features of a pitch extraction method offered by higher order moments. Furthermore, the correlation method performs well compared with YIN when the strategy of estimating the pitch from a longer window followed by a localized estimate is used. This result further underscores the importance of window length in pitch estimation in that longer windows in addition to providing lower overall errors also result in fewer large errors.

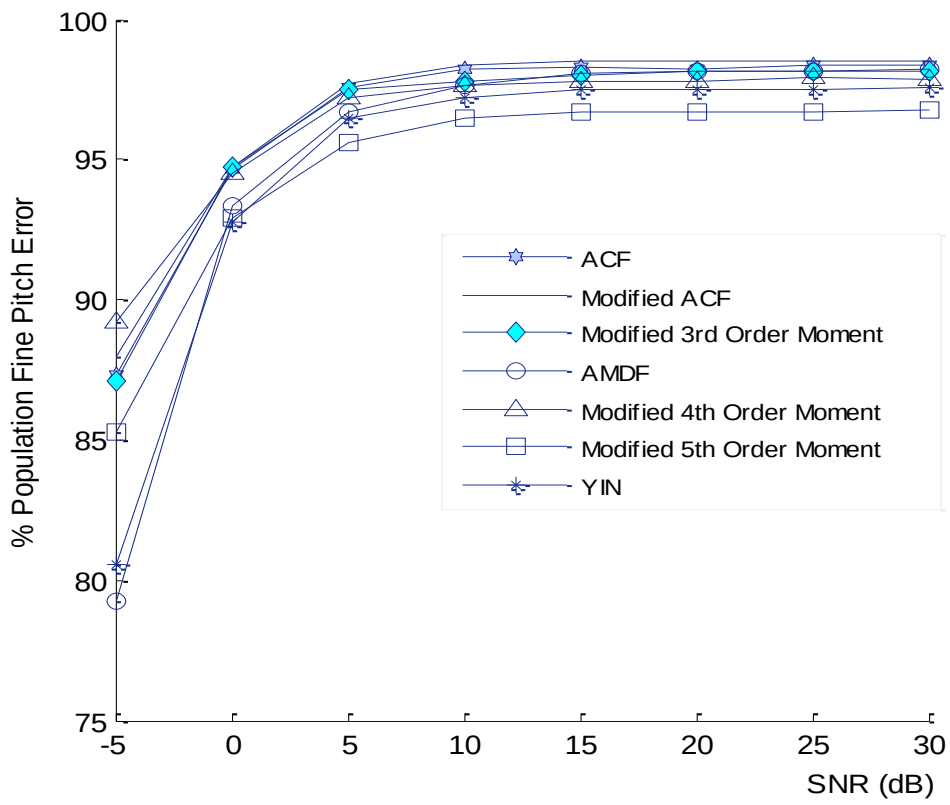


Figure 5.18 - The % population fine pitch error of speech in Gaussian white noise as a function of SNR.

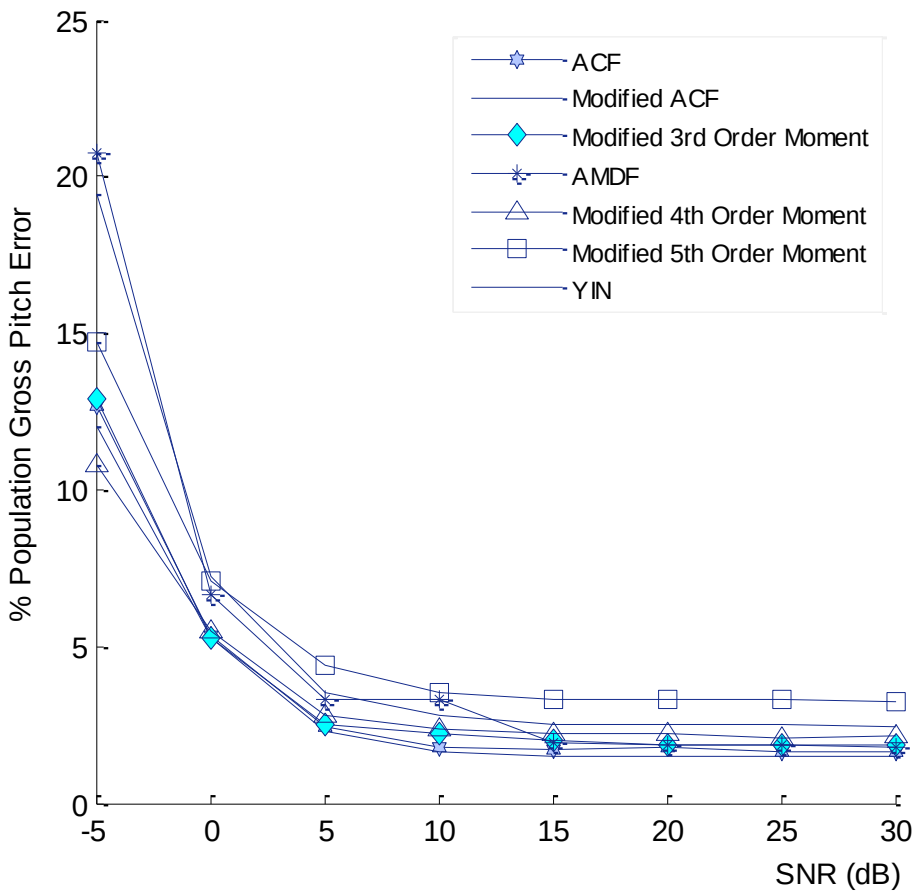


Figure 5.19 - The % population gross of pitch error of Gaussian white noise as function of SNR.

5.7 CONCLUSIONS

The main contributions of this chapter are (a) comparative evaluation of a set of similarity criterion for pitch estimation and (b) determination of the variation of pitch accuracy as a function of the window length.

The pitch extraction methods, based on the second order, modified second order, modified third order, fourth order and fifth order moments are evaluated with varying widow length and in four different types of noise and in a range of SNRs from -5 dB to 30 dB; the

results compete favourably with the conventional methods (i.e. autocorrelation with 20 ms window length and AMDF) and the benchmark YIN method.

A significant finding of this research work is the relatively dominant impact of the choice of window length on the pitch estimation error. Since MHOM methods involve the average of the products of samples, $x(m) \cdots x(m-(k-1)T)$, that are apart by two or more times the maximum allowable period, the reliable estimation of MHOMs requires appropriately larger windows. A set of experiments were conducted to determine the curves of the variations of pitch error versus speech window length for conventional and higher order pitch estimation methods. The results reveal that the pitch error decreases with the increasing speech window length, despite the impact of a non-stationary process, and that the choice of speech window length is the most influential factor effecting pitch estimation accuracy. For example pitch estimation based on correlation method using a window length of 50 -70 ms outperforms the bench mark YIN method of pitch extraction. The apparent downside of choosing a large window is an increase in delay. However, this is overcome by the proposed two-stage solution whereby an initial estimate of the pitch from a large window spanning the current and past speech frames is followed by fine-tuning around the current speech frame.

Of the higher order methods experimented with, the modified third order moment in particular and the fourth order moment methods perform well. Beyond the fourth order moment the relatively large length of speech window required, contains significant variation of pitch and this affects and limits the accuracy of pitch estimation. Significantly the third order and fourth order methods are robust in that in addition to yielding smaller pitch error also result in a small percentage of large pitch errors.

The potential advantages of higher order moment criteria may be cited as:

1) A higher level of reinforcement of similarity resulting from multiplication of more than two similar samples as shown in Figure 5.2.
2) Robust and improved performance at low SNRs (i.e. Figures 5.10 to Figure 5.13).
3) Competitive performance across SNRs (i.e. Figures 5.10).

The potential disadvantages of higher order moment criteria may be cited as:

1) Increase computational complexity;
2) Increased delay (the proposed two-stage method resolves this issue); and
3) Coarse estimate resulting from the averaging of pitch within a large window.

6

PITCH ESTIMATION VIA ANALYSIS-SYNTHESIS OF *N*-BEST CANDIDATES

The similarity criteria (e.g. moments or average magnitude difference function) used for the estimation of the period, or its inverse the fundamental frequency, yields multiple competing candidates at the extrema points giving rise to errors when the similarity at one of the maxima/minima other than the correct pitch is strongest. This chapter addresses the problem of determination of the best pitch value among a number of *N* proposed pitch candidates selected at the extrema of the similarity criterion used for pitch estimation. For each prospective pitch candidate, $F_{0_i}, i = 1, \dots, N$, the harmonic part of speech with the proposed fundamental F_{0_i} is synthesised in frequency domain as the product of an estimate of the spectral envelope of speech and an estimate of

the spectrum of the harmonics of the excitation. The synthesised speech spectrum is subtracted from the actual speech spectrum and the difference is used to yield a harmonic synthesis distortion measure such as the harmonicity distance, HD, signal-to-noise ratio, SNR, minimum mean squared error, MMSE etc. Furthermore, the harmonicity of speech at the proposed pitch and its harmonics is used as an additional component of the overall distortion score. The pitch candidate yielding the smallest synthesis distortion is selected as the most likely pitch value. The choice of the harmonic synthesis model and its parameters and the choice of the distortion measure are critical and these should be selected so as to maximise the mismatch between the harmonic part of speech signal reconstructed from an incorrect pitch candidate and the actual speech signal. For evaluation of errors, the pitch reference (aka 'ground truth') values are calculated from manually-corrected estimates of the periods obtained from laryngograph signals.

6.1 INTRODUCTION

Speech signals are composed of a combination of quasi-periodic and non-periodic signals. The term quasi-periodic implies that the signal is seemingly, but not strictly, periodic because the period varies over time. The pattern of time-variation of the pitch, known as intonation, conveys such information as pragmatics of speech, intent, style and accent.

Pitch extraction methods utilise the similarity of speech samples at time t, $x(t)$, with the speech samples a period of T seconds away; $x(t+T)$ or $x(t-T)$.

For example, correlation-based pitch extraction methods estimate the period as the value of T for which the average of the product of $x(t)x(t-T)$ over a frame of speech samples, known as the short-time correlation, attains a maximum value [25]. Magnitude-

difference-based pitch extraction methods estimate the period as the value of T for which the average magnitude difference $|x(t) - x(t-T)|$ over a frame of speech samples attains a minimum.

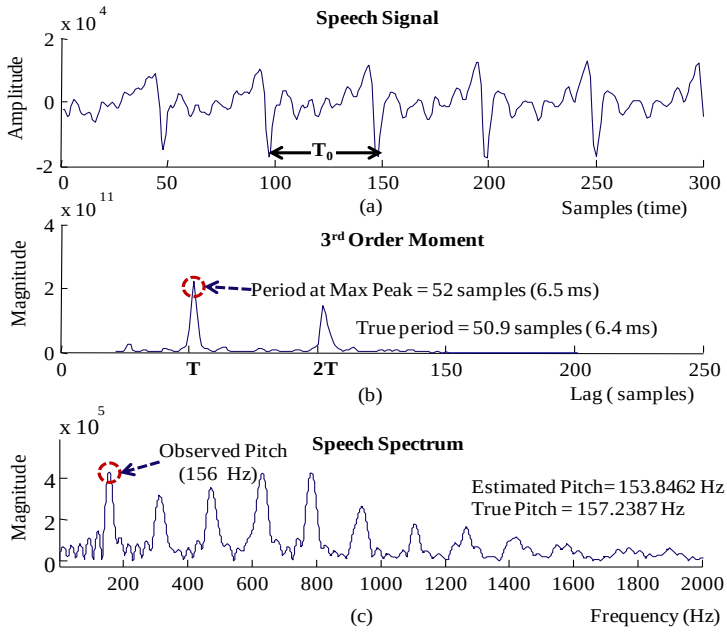


Figure 6.1 - Illustration of: (a) a periodic speech segment, (b) the peaks of the 3rd order moment as the candidates for period/pitch estimation, the max peak position 52 sample is close to the true period of 51 samples and (c) the frequency-domain representation of the harmonic structure of periodic speech.

A periodic signal with a period of T is also periodic at integer multiples of T i.e. $2T$, $3T$, ... etc. Hence similarity exterma points will occur also at integer multiples of Tas illustrated in Figure 6.1.

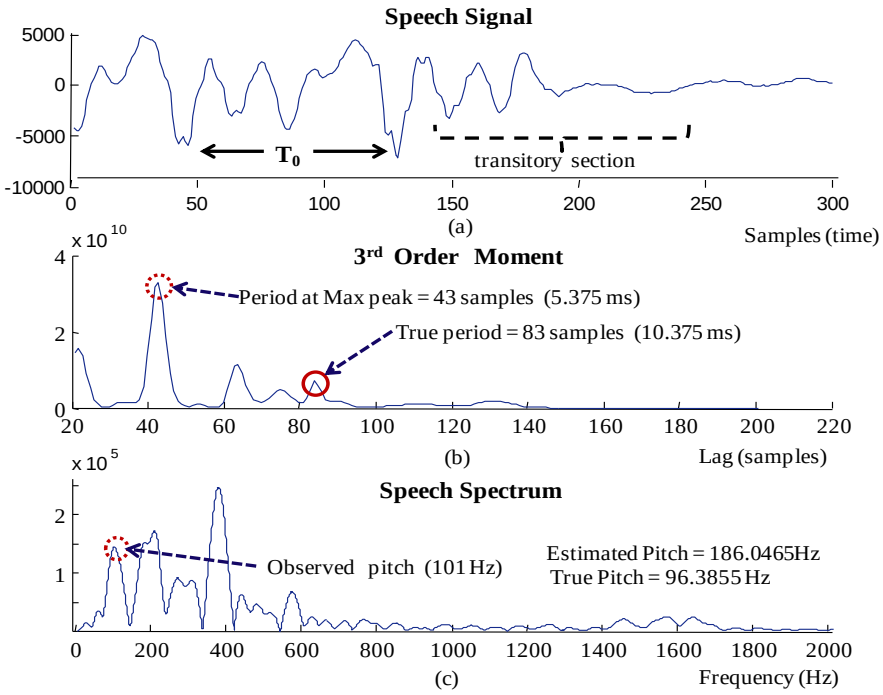


Figure 6. 2 - Illustration of: (a) a transient speech segment, (b) double pitch (i.e. half period) estimation from the max peak of the 3rd order moment, and (c) the frequency-domain representation of the harmonic structure of speech

In cases when the k^{th} harmonic of speech coincides with a strong resonance, and hence becomes the dominant harmonic, strong periodicity will be seen at a period of T/k, for example when the second harmonic is stronger than the first harmonic then a correspondingly stronger similarity will be seen at a lag of half period $T/2$ as shown in Figure 6.2.

Due to non-stationary and indeterminate nature of some speech segments sometimes a strong similarity extrema occurs at a period other than that expected. This is particularly the case for speech segments at the transitory sections as shown in Figure 6.2 (a), i.e. at the beginning or the end of an utterance or at the boundary of a transition between two utterances, when the signal period within a given segment window varies considerably from the beginning to the end [137] .

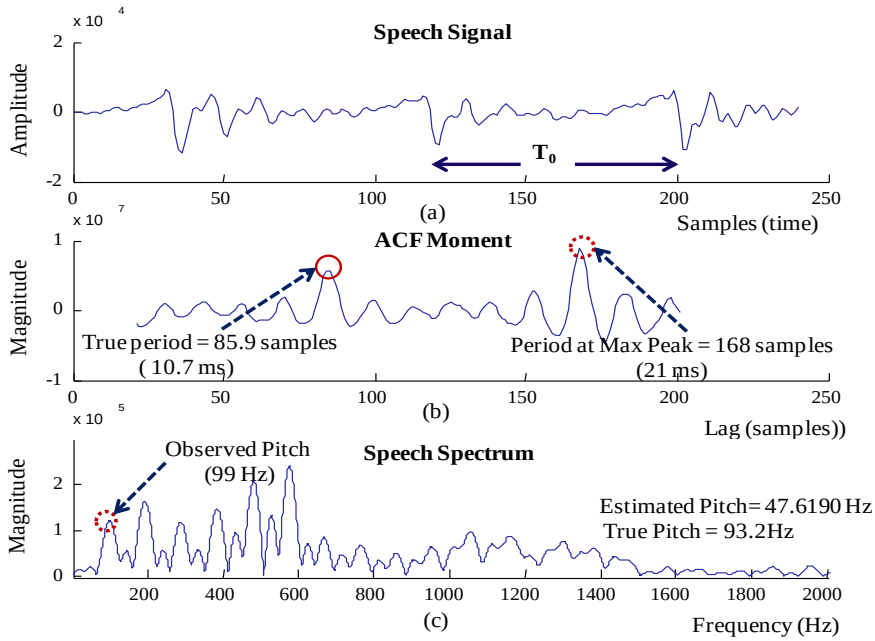


Figure 6.3 - Illustration of: (a) a periodic speech segment, (b) the correct pitch estimation from the ACF (2nd order moment), and (c) the frequency-domain representation of the harmonic structure of periodic speech.

Figure 6.3 shows an example where the maximum of the autocorrelation function method, ACF similarity criterion corresponds to a twice the true period which is equivalent to half the true pitch.

6.2 THE PROPOSED *N*-BEST CANDIDATES PITCH ESTIMATION METHOD

Figure 6.4 shows an outline of the proposed *N*-best pitch estimation method. A similarity criterion, such as the k^{th} order moment, $M(T)$, is calculated for the speech signal for a range of values of the period, T, spanning the minimum and the maximum expected values such as $T_{min} = 2.5$ ms and $T_{max} = 25\,ms$ corresponding to fundamental frequencies of $F_{max} = 400$ Hz and $F_{min} = 40$ Hz respectively. Alternatively, the

similarity criteria can operate on the frequency domain signal obtained from the Fourier transform.

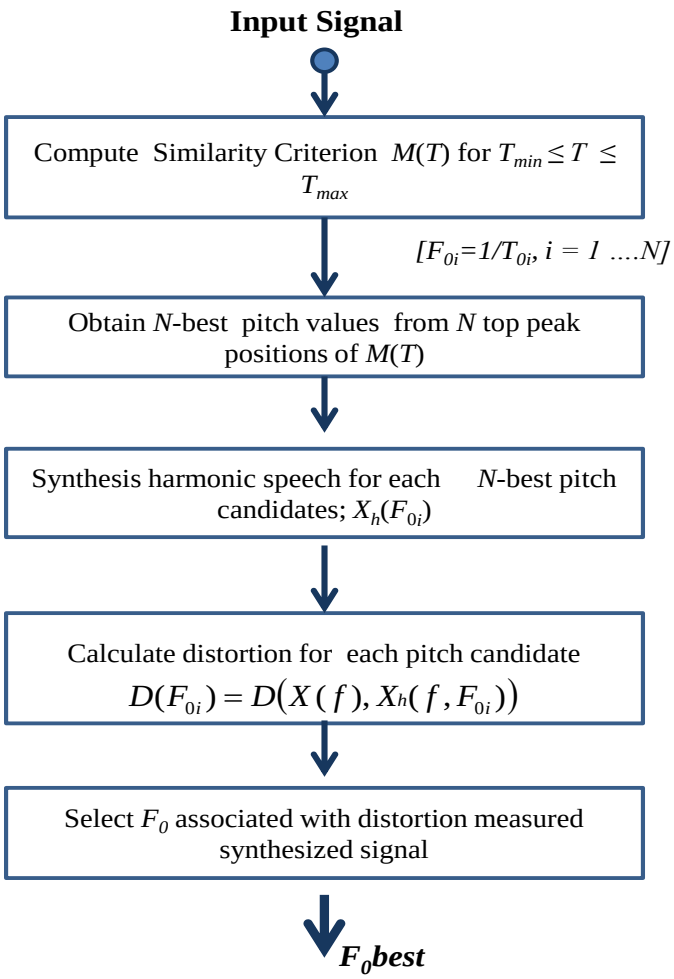


Figure 6. 4 - An outline of N-best pitch estimation method.

From the set of N candidate periods corresponding to the positions of the N top peaks of the similarity curve $[T_{0i}, i = 1 \cdots N]$ a set of N pitch candidate values are obtained as $[F_{0i} = 1/T_{0i}, i = 1 \cdots N]$.

For each pitch candidate $[F_{0i}, i = 1 \cdots N]$, the harmonic part of speech spectrum is synthesised to yield $[\hat{X}_h(F_{0i}), i = 1 \cdots N]$. A set of spectral distortion measures, such as weighted SNR or weighted MMSE, harmonicity etc. are accumulated as $D(F_{0i}) =$

$\sum_f D(X(f), \hat{X}_h(f, F_{0i}))$. The pitch candidate with the least distortion is selected as the best estimate of the pitch. The success of the N-best strategy depends on the methods used for synthesis of the harmonic part of speech and the choice of the distortion measure as explained next.

For each of the N-best pitch candidates, $F_{0i}, i = 1 \ldots N$, the harmonic plus noise model of speech is defined as

$$\hat{X}(f, F_{0i}) = \sum_{k=1}^{N_h} A(k) G_k(f - kF_{0i}) + V(f) \tag{6.1}$$

where F_{0i} is the fundamental frequency, $A(k)$ is the k^{th} harmonic spectral amplitude, $G_k(f)$ is k^{th} harmonic excitation spectral shape function, N_h is the number of harmonics and $V(f)$ is the non-harmonic part of speech [138].

For each of the N-best pitch candidates, F_{0_i}, i =1…, N, the harmonic part of speech is synthesised as

$$\hat{X}_{hi}(f, F_{0i}) = \sum_{k=1}^{N_h} A(k) G_k(f - kF_{0i}) \tag{6.2}$$

The distortion measure for the original and synthesised spectra, X and X_h, for the i^{th} pitch candidate is defined as

$$D\left(F_{0_i}, X_{hi}, X\right) = \sum_{f=0}^{N-1} d\left(X(f), \hat{X}_{h_i}(f, F_{0_i})\right) \tag{6.3}$$

Various forms of distortion measures evaluated are described in section 6.6.

The success of the N-best strategy depends on the efficiency of the model used for synthesis of the harmonic part of speech and on the choice of the distortion measure as explained next. The best pitch value among the N candidates is obtained as

$$F_0 = \min_{i=1\ldots N} D\left(F_{0_i}\right) \tag{6.4}$$

The implementation of the N-best methods requires the estimation of the following parameters:

1) Estimation of N-best pitch candidates, $[F_{0_i}, \ldots F_{0_N}]$, obtained from the top N extrema of a similarity criterion.
2) An estimate of the parameters of the shape of each harmonic excitation signal $G_k(f)$ to fit the actual speech harmonic shapes.
3) An estimate of the spectral envelope $A(f)$; this is a critical part of the method; the spectral envelop should be such that the error in harmonic synthesis is an increasing function of the pitch error **[**139] - [141].
4) A spectral distortion measure of the difference between the actual speech signal and the harmonic synthesised signal [142].
5) A selection method for finding the best candidate, this may be a straightforward selection of the pitch candidates that yields the best synthesised harmonics of speech or it may additionally employ the past history of the pitch estimates within nearest neighbour.

Figure 6.5 illustrates the N-best Pitch estimation method proposed in this chapter. The bandpass filter limits the signal to band of 40 Hz to 2000 Hz where the signal harmonics are expected to have the strongest harmonicity. The segment length depends on the type of the similarity criteria and is set between 30 ms to 62.5 ms as explained in section 5.3 (chapter 5). The segment shift 5 ms corresponding to 200 estimations of the pitch value per second. Each segment is windowed by the commonly used Hanning window.

After bandpass filtering, segmentation and windowing, the similarity criterion yields *N*-best competing candidates corresponding to the positions of the *N* top extrema points sorted in order of the decreasing magnitude of the extrema.

For each pitch candidate the harmonic synthesis module synthesises the harmonic part of the speech spectrum using a combination of the estimates of the spectral envelop of speech and the spectral details of the harmonic excitation. The estimation of spectral envelope and excitation details of speech are described in section 6.4.

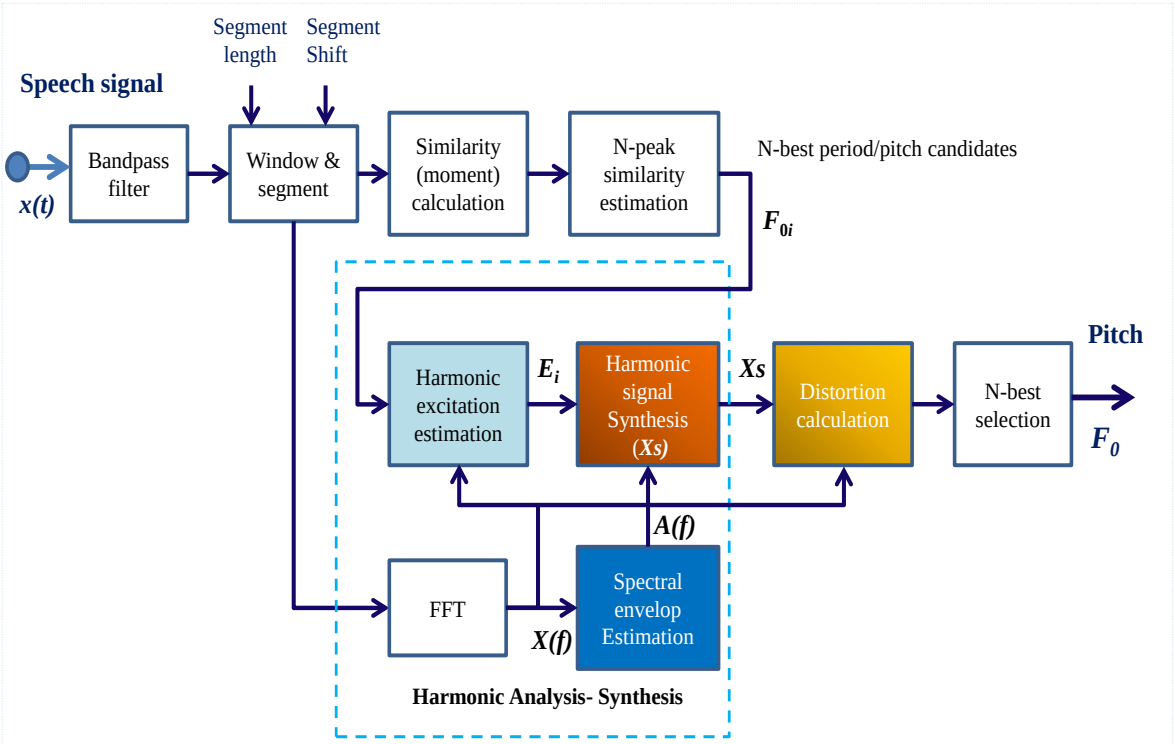


Figure 6.5 - Illustration of *N*-Best pitch estimate algorithms

6.3 HARMONIC EXCITATION MODEL ESTIMATION

Figure 6.6 shows several sub-processes of the harmonic signal model estimation; these are:

1) Harmonics' frequency adjustment,
2) Excitation pulse shape estimation, and
3) Selection of the number of harmonics for analysis-synthesis.

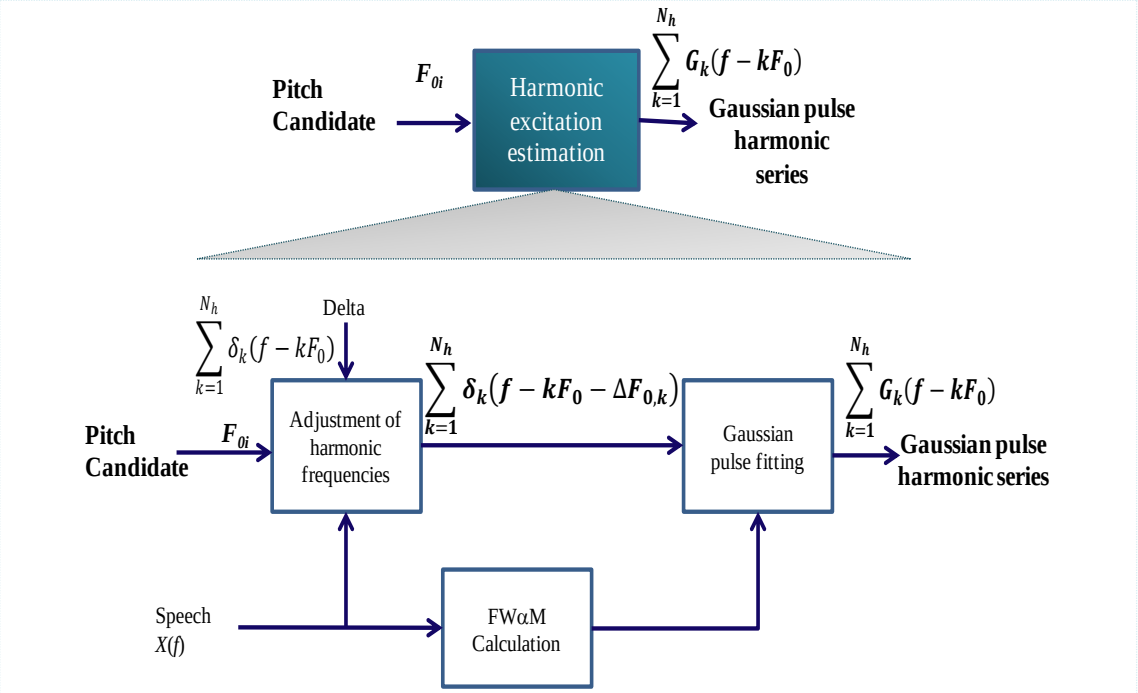


Figure 6.6 - The block diagram of harmonic excitation estimation model.

6.3.1 Harmonic Frequency Adjustment

The peaks of the harmonics of voiced speech do not necessarily occur at exact integer multiples of the fundamental frequency as may be the theoretical expectation. The reason for deviation of the position of the harmonics from the integer multiples kF_0, is the time-varying nature of the excitation signal and the configuration and resonances of the voice tube through which the excitation propagates. Hence a more accurate model of the position of the frequencies of the harmonics may be modelled as

$$E_h(f) = \sum_{k=1}^{N_h} \delta\left(f - kF_0 - \Delta_{kF_0}\right) \tag{6.5}$$

where Δ_{kF_0} is the deviation of the k^{th} harmonic from the nominal value of kF_0. The value of Δ_{kF_0} is found around the locality of kF_0 by a peak search in the region of $kF_0 - \delta elta$ to $kF_0 + \delta elta$ where $\delta elta$ is a user-set search region parameter. Δ_{kF_0} is obtained as

$$\Delta_{kF_0} = \underset{-\delta elta \le l \le \delta elta}{\operatorname{argmax}} X(kF_0 + l) \tag{6.6}$$

After adjustment of the harmonic frequencies as shown in Figure 6.6, a weighted estimate of the fundamental frequency is obtained as summation of the harmonic frequencies as

$$F_{0i} = \sum_{k=1}^{N_h} w(k) \frac{F_{0ik}}{k} \tag{6.7}$$

where $\frac{F_{0ik}}{k}$ is an estimate of the fundamental frequency obtained from the adjusted k^{th} harmonic frequency normalised by the harmonic number k as shown in Figure 6.7.

The weights $w(k)$ may reflect the strength of each harmonic component and they may be obtained from a number of different measures such from the normalised SNR obtained for each synthesised harmonic:

$$w(k) = \frac{SNR(k)}{\sum_l SNR(l)} \tag{6.8}$$

or from the estimate harmonicities at F_{0ik} as

$$w(k) = \frac{Harmonicity(k)}{\sum_l Harmonicity(l)} \tag{6.9}$$

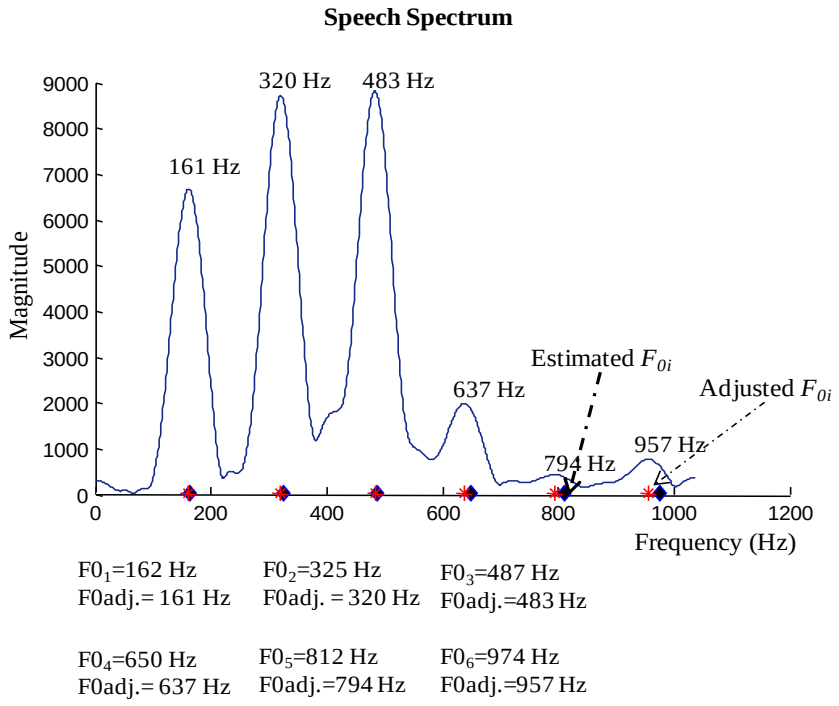


Figure 6.7 - The frequency adjustment of speech spectrum. The star (red) represents the adjusted F_0 and diamond (black) is the estimated F_0 before adjustment.

6.3.2 Harmonic Excitation Shape Estimation

The frequency spectrum of the harmonic part of the excitation for each segment of speech, $E_h(f)$, is modelled as a sequence of periodic Gaussian functions positioned at the frequencies corresponding to the fundamental F_0 and the harmonic frequencies kF_0

$$E_h(f) = \sum_{k=1}^{N_h} G_k(f - kF_0) \tag{6.10}$$

where $G_k(f - kF_0)$ is a unit-amplitude Gaussian function fitted to the k^{th} harmonic pulse centered at the frequency of kF_0.

The parameters of a Gaussian function are the mean value μ and the variance σ^2. For each harmonic signal the mean of the Gaussian function is set to the harmonic frequency; i.e. the mean or the centre of the Gaussian function for the k^{th} harmonic is $kF_0$and the Gaussian function is given by

$$G_k(f - kF_0) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-0.5\left(\frac{f-kF_0}{\sigma}\right)^2} \tag{6.11}$$

Removing the scaling factor $1/\sigma\sqrt{2\pi}$ yields a unit-amplitude Gaussian pulse, i.e. with a maximum value of 1, as required in the context of the signal synthesis in this work

$$G_k(f - kF_0) = exp^{-0.5\left(\frac{f-kF_0}{\sigma}\right)^2} \tag{6.12}$$

For each excitation harmonic of speech, the variance of the Gaussian function is calculated such that the width of the Gaussian function at a pair of points, at a predetermined fraction of the maximum value of the Gaussian pulse, fits the width of excitation.

For fitting the Gaussian function to the harmonic pulses of the speech signal, we select to fit the width of a Gaussian function to the width of the harmonic pulse, at such symmetric points about the mean, $(-f_\alpha, f_\alpha)$ where the Gaussian function magnitude is a fraction α of the maximum value, these points are known as full-width-α-maximum, F$W\alpha$M [143]. This process is illustrated in Figure 6.8. Hence for each harmonic pulse the width, F$W\alpha$M, between the two points on either side of the maximum peak at the harmonic, where the harmonic pulse magnitudes are a fraction α of the peak value, are measured. The variance is then calculated, as described next, such that F$W\alpha$M of the Gaussian function is the same as F$W\alpha$M of the speech excitation harmonic pulse.

Consider a Gaussian function $exp^{-0.5\left(\frac{f}{\sigma}\right)^2}$ with its mean centred at $f = 0$, for the following derivation we do not need the scaling constant $\frac{1}{\sigma\sqrt{2\pi}}$. Since the maximum value of the Gaussian pulse is unity, i. e. $\max\left(exp^{-0.5\left(\frac{f}{\sigma}\right)^2}\right) = 1$, the frequency f_α at which this function will have a magnitude equal to a fraction α of the maximum magnitude is given by

$$\alpha = exp^{-0.5\left(\frac{f_\alpha}{\sigma}\right)^2} \tag{6.13}$$

$$\ln \alpha = -0.5\left(\frac{f_\alpha}{\sigma}\right)^2 \tag{6.14}$$

$$f_\alpha = \pm\sigma\sqrt{2\ln(1/\alpha)} \tag{6.15}$$

The width between the points $(-f_\alpha, f_\alpha)$ at which the magnitude of the pulse is a fraction α of the maximum value is given by

$$FWaM(\alpha) = f_\alpha - (-f_\alpha) = 2f_\alpha \tag{6.16}$$

$$FWaM(\alpha) = 2\sigma\sqrt{2\ln(1/\alpha)} \tag{6.17}$$

For a value of a=0.5 we have the full width at half magnitude, FWHM,

$$FWHM = 2\sqrt{2\,ln(2)}\,\sigma = 2.3548\,\sigma \tag{6.18}$$

Alternatively we can use a fitting of the full width at other pointes such as at three quarter of the maximum or at a quarter maximum values.

For a value of a=0.75 we have the full width at three quarter of the maximum magnitude, i.e. a quarter below the maximum, FW3QM, as

$$FW3QM = 2\sqrt{2\ ln(4/3)}\ \sigma = 1.517\ \sigma \tag{6.19}$$

For a value of a=0.25 we have the full width at quarter of maximum magnitude, i.e. three quarter below the maximum, FWQM, as

$$FWQM = 2\sqrt{2\ ln(4)}\ \sigma = 3.33\ \sigma \tag{6.20}$$

Note the maximum value of the Gaussian pulse is at its centre with a value of 1 and this will be scaled during synthesis with the value of the envelope of the harmonic of speech at kF_0.

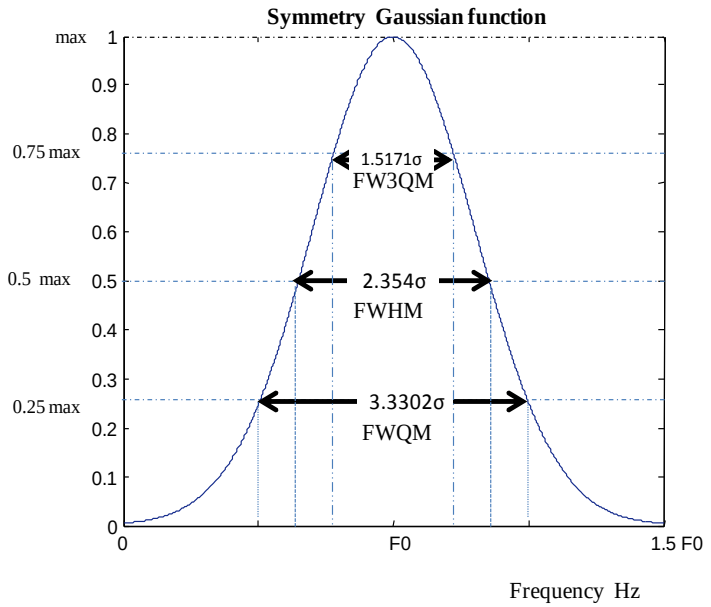


Figure 6.8 - Illustration of asymmetric Gaussian distribution model, with full width three quarter maximum (FW3QM), full width half maximum (FWHM), and full width quarter maximum (FWQM).

6.3.3 Asymmetric Gaussian Pulse Shape for Harmonic Excitation

Although theoretically, the shape of the excitation harmonics in frequency domain may be desired to be symmetric and regularly spaced at integer multiple of the fundamental frequency, kF_0, in practice, due to influence of the resonances and the anti-resonances of the vocal tract and the time-varying nature of speech, the shape of the harmonic pulse, the position of the peak points of the harmonics of speech and the frequency distances between the successive harmonics varies along the frequency.

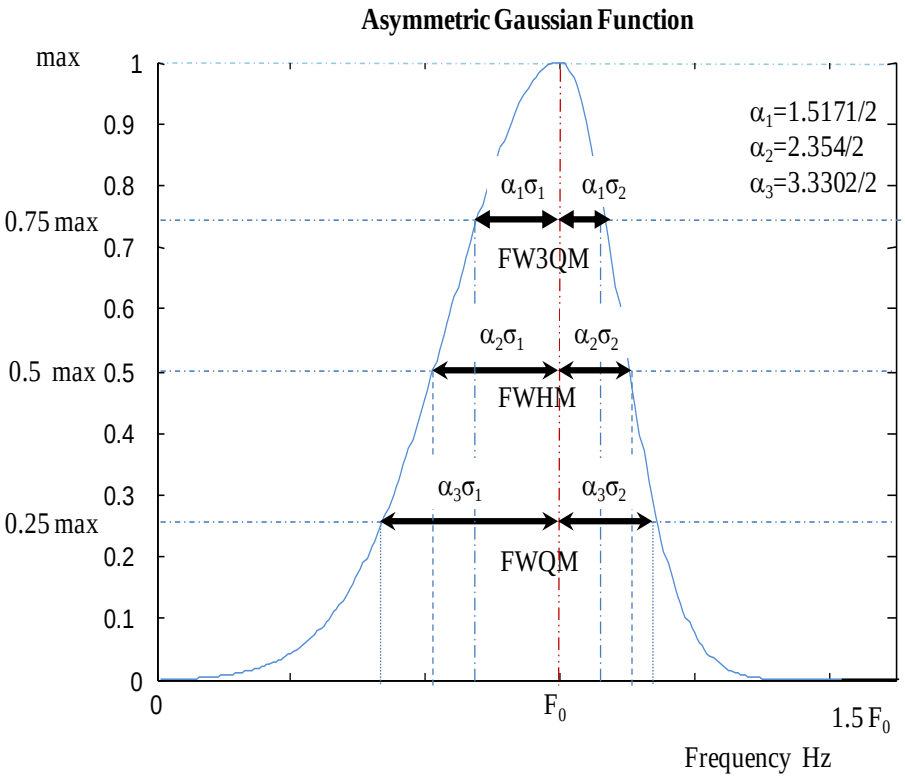


Figure 6.9 - Illustration of asymmetry Gaussian distribution model.

Hence, for example we are likely to have an asymmetric harmonic pulse shape such that the number of frequency samples measured from the mid points between two successive harmonics to the peak point (at harmonic frequency) is different on the two sides of the

harmonic. In addition the harmonic pulse shape itself may be skewed having two different slopes at different sides of the harmonic frequency. The asymmetry of the Gaussian pulse is modelled in order to optimise the shape of the synthesised speech spectrum as shown in Figure 6.9.

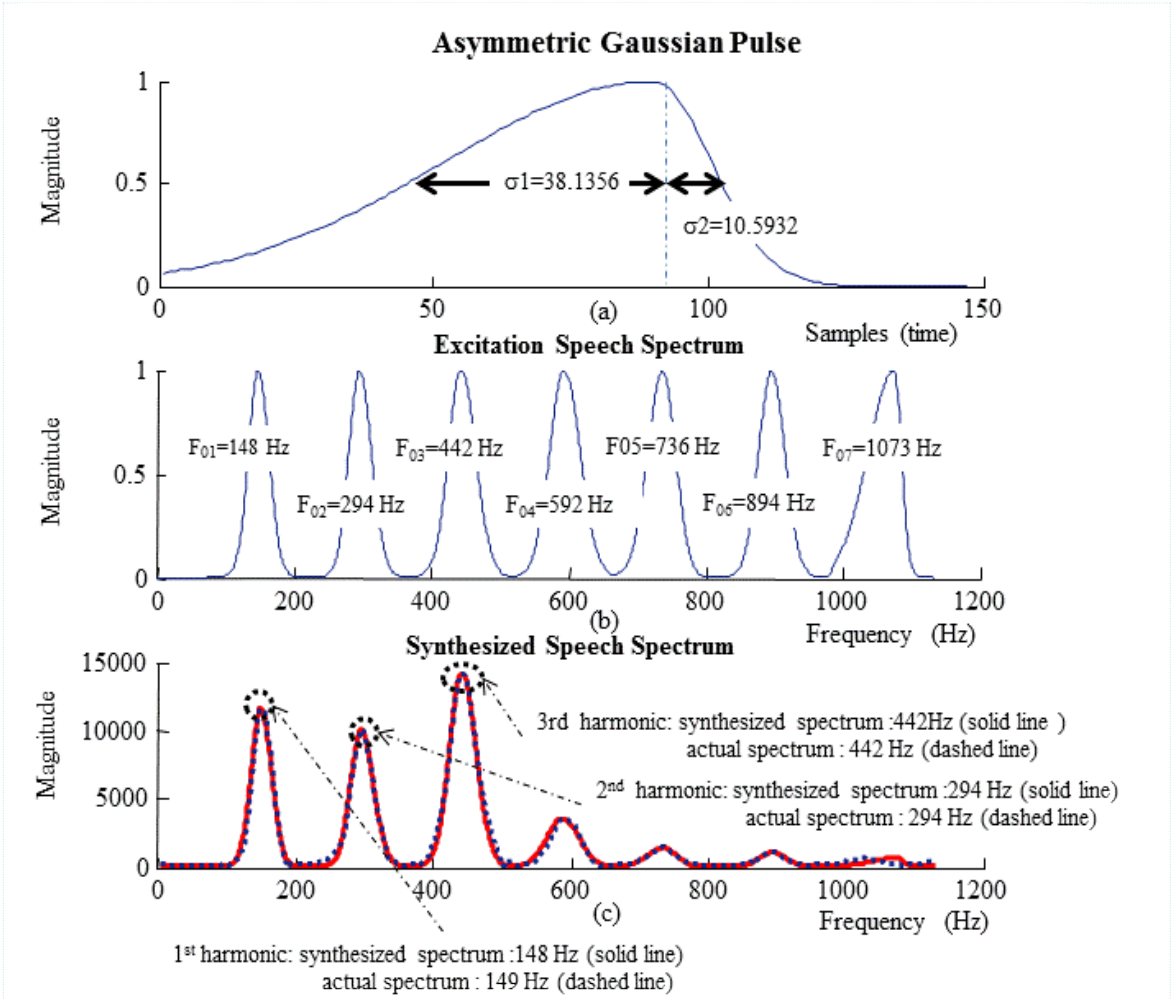


Figure 6.10 - Harmonic excitation of speech spectrum with (a) an asymmetric Gaussian pulse model, (b) their harmonic excitation with multiple integer speech spectrum, and (c) the synthesised speech spectrum (solid line) and the original speech spectrum (dashed line).

6.3.3.1 Generation of an Asymmetric Gaussian Pulse from Two Half-Gaussian Pulses

An asymmetric Gaussian pulse can be formed from the generation and concatenation of two half Gaussian pulses (the left half and the right half about the peak value) of different variances, σ_1, σ_2, as shown in Figure 6.10. To achieve this, two separate values of variance of the Gaussian pulses to fit the full widths at a fraction α of the maximum. FW□M, one for the harmonic curve to the left of harmonic peak value and one to the right of the curve are computed. Then, given the values of σ_1, σ_2, a software code generates and concatenates the two half Gaussian curves.

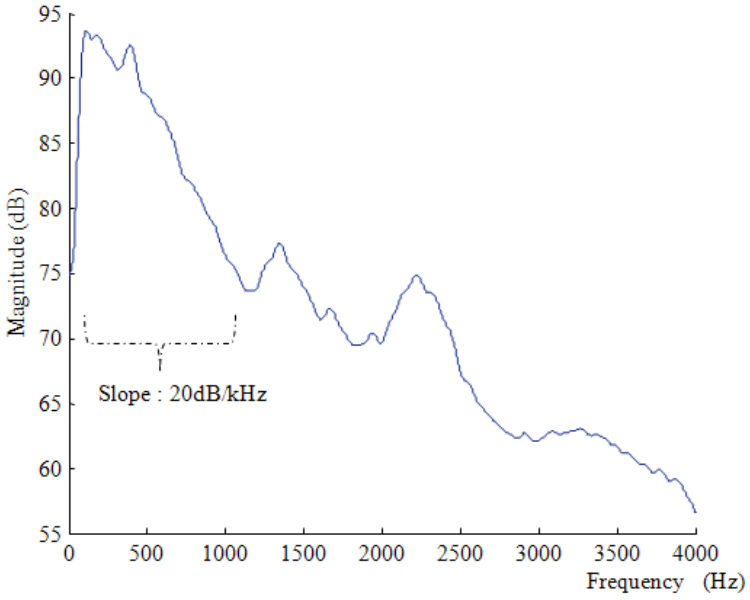
Figure 6.10 illustrates the generation, synthesis and fitting of an example of a speech segment with asymmetric excitation.

6.3.4 Selection of Number of Harmonics N_h

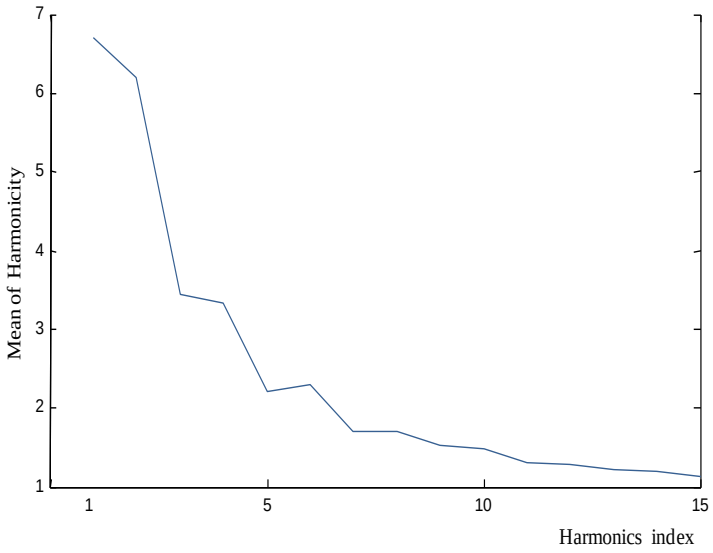
The choice of the number of harmonics used for estimation of the fundamental frequency will give impacts to the accuracy of pitch estimation.

In general for a bandwidth of $BW = F_s/2$ Hz, and a fundamental frequency of F_0 there is a maximum number of harmonics of $N_h = fix(BW/F_0)$. However, for most speakers only the first few harmonics are well defined with a high harmonicity and a shape that can be conveniently modelled by a Gaussian pulse. Experimentally, we have arrived at the choice of a 1000 Hz as the bandwidth that contains the most significant harmonics as shown in Figure 6.11. This choice is justified by the two experimental observations: (1) that the power spectrum of voiced speech drops in magnitude by more than 20 dB after 1000 Hz as shown in Figure 6.11 (a), and (2) the most well defined harmonics, i.e. those with the highest values of harmonicity, are the first 5 harmonics which on average reside in a frequency bandwidth of 1000 Hz, as shown in Figure 6.11 (b).

Hence, in the N-best pitch estimation method developed in this thesis the number of significant harmonics for analysis-synthesis, given a pitch candidate proposal F_{0i}, is calculated as $N_h = fix(1000/F_{0i})$.



(a)



(b)

Figure 6.11 - (a) showing the power spectral density of voiced speech signal; note that at 1 kHz the power is down by 20 dB, and (b) the harmonicity at the first 15 harmonics of the voiced speech of a male speaker.

6.4 SPECTRAL ENVELOPE ESTIMATION, $A(f)$

Spectral envelope estimation is one of the most crucial and critical parts of speech processing applications such as, speech coding, speech synthesis and speech recognition where the system performance is affected by the accuracy of the spectral envelope estimate [139] - [140].

In this application an estimate of the synthesized signal is obtained via multiplication of an estimate of the spectral envelop, $A(f)$, by an estimate of the excitation harmonic, for the proposed pitch F_{0i}, $E_h(f, F_{0i})$. For maximising the mismatch, between the magnitude spectrum of the actual speech signal and an incorrectly synthesised harmonic (for example a harmonic incorrectly synthesized with half pitch or double pitch estimates), the spectral envelop should ideally pass through the peaks of speech magnitude spectrum at the harmonic frequencies but it should not go through other points and in particular the envelope should not pass through the local peaks that reside at the spectral troughs in between the harmonics.

There are several established alternatives for the modelling of the spectral envelope or the vocal tract frequency response of speech, each method has merits and shortcomings, these methods are discussed in section 3.3 (Chapter 3) and in the following references [139], [144].

In particular LPC envelope and cesptrum envelop were experimentally explored as two alternative methods of spectral envelop estimation. However, both methods fail to conform to the requirement that the spectral envelop should go through the significant peaks at the harmonics and ideally should not miss any peaks or exhibit false peaks. LPC

and Cepstrum envelop do not have sufficient accuracy in tracing the significant spectral peaks and often exhibit sharp peaks where none exists in the actual signal.

The preferred choice for estimation of the spectral envelop in this work, justified in the followings, is the polynomial interpolation through the peaks of the spectrum at the harmonics.

Broadly, the spectral envelop via interpolation of a curve through the harmonic peaks, or the significant peaks, involves the following stages:

1) Identification and estimation of magnitude and position of the most significant spectral peaks; here a peak is defined as a turning point that stands above a preset number of neighbouring samples on both sides of the peak.
2) Pruning of the spectral peaks; this is an iterative process of identification and retaining of the most significant peaks; it uses two threshold parameters, these are peak prominence and minimum distance between successive harmonics.
3) A polynomial interpolation method, interpolates a spectral curve through the spectral peak points.

For the specific application of estimating a spectral curve that rests on the peaks at the harmonics, a number of constraints have been employed as:

1) Even distribution of the peaks across the bandwidth. In order to ensure that the spectral peaks are relatively evenly spread (as would be expected from a harmonic structure), the bandwidth is divided into N segments and in each segment M peaks are selected. Typically for the first 4 kHz (telephony) speech bandwidth, $N = 20$ and $M = 2$. This implies that each frequency segment has a bandwidth of 200 Hz

width ($BW/N = 4000/20$) and for each segment 2 peaks are estimated, hence for each 200 Hz band two peaks can be estimated.

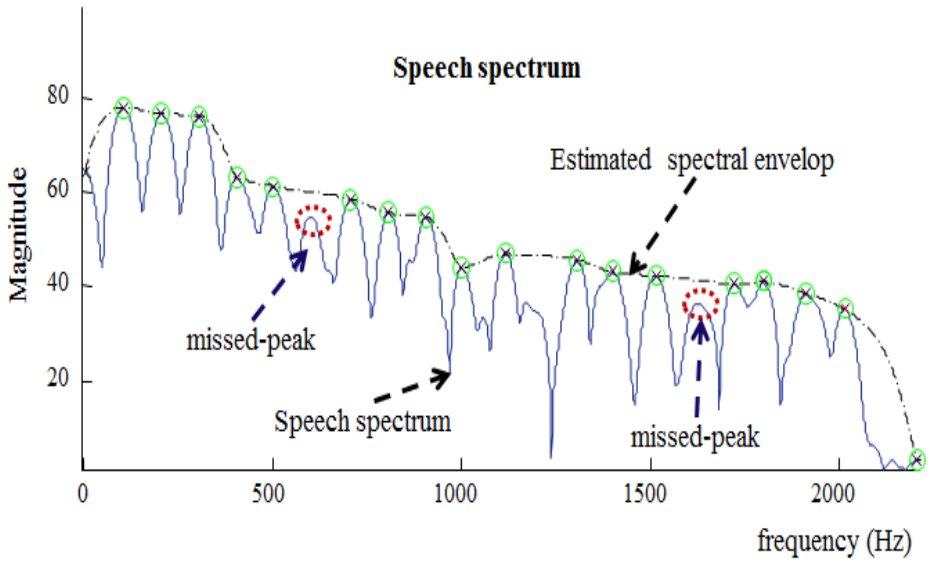
2) Maximum frequency difference between the positions of two consecutive peaks, the maximum frequency difference is set to 350 Hz which is in the upper range of the female pitch. The setting of this threshold may result in inclusion of some non-harmonic peaks data, above the threshold, in the estimation of the envelope.
3) Minimum frequency difference between the positions of two consecutive peaks, the minimum frequency difference is set to 80 Hz. The setting of this threshold helps to prevent inclusion of non-harmonic peaks data below the minimum frequency.
4) Maximum dB drop in magnitude between two successive peaks, this is set to value of 20 dB.

6.4.1 Harmonic Peak Identification: Optimising the Trade-off between the Miss-Rate and the False-Alarm Rate

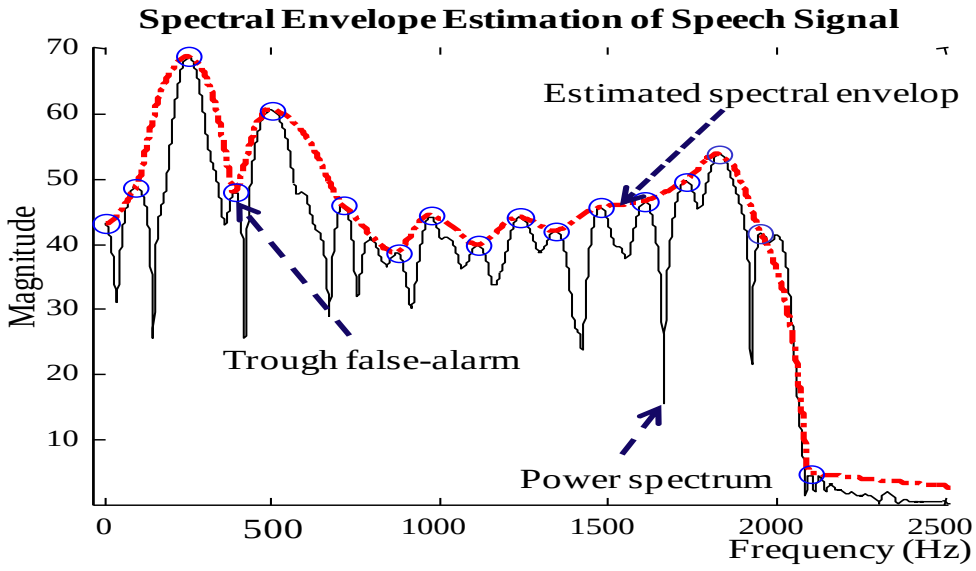
The classical trade-off in the communication, detection/estimation, theory between the miss-rate and the false-alarm rate can be applied to the objective of identifying the correct peaks at the harmonic frequencies while avoiding the misidentification of smaller peaks at non-harmonic frequencies (akin to false-alarms).

This can be guided by arguing that for the particular N-best pitch estimation problem the cost of missing some of the least prominent peaks at the harmonics (miss-rate) may be greater than the cost of misidentification of incorrect peaks that may for example reside at trough between the harmonics as shown in Figure 6.12 (a).

The trade-off between miss-rate and false-alarm rate can be set by the choice of thresholds used in the envelop estimation process as shown in Figure 6.12 (b) and explained further in the next section.



(a)



(b)

Figure 6.12 - The Spectral envelopes using polynomial interpolation (PCHIP), (a) the speech segment, (b) the spectral envelop with missed-rate estimation, and (c) the spectral envelop with trough false-alarm estimation (dashed line).

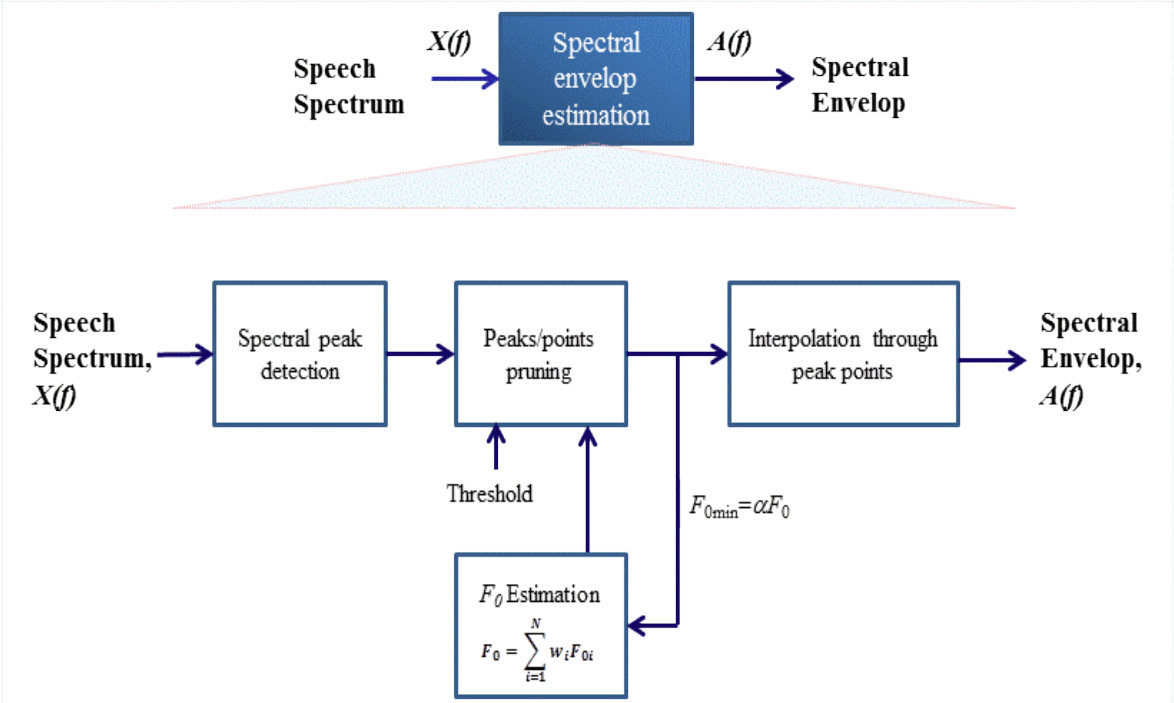


Figure 6.13 - The block diagram of spectral envelop estimation model.

6.4.2 Algorithm for Spectral Envelop Estimation

Figure 6.13 and 6.14 illustrate the block diagram and the flow chart for spectral envelop estimation algorithm.

Set the number of subbands, $M = 20$;

Set the number of peaks per subbands, $N = 2$;

This implies that *N* peaks will be considered for each frequency segment of width *BW/M*.

Initial Steps (1 & 2)

Step 1*:* Find N distinct spectral peaks in M subbands ($N \times M$ peaks) subject to the constraints that successive peaks should have a minimum frequency spacing $of\ F_{min}$, a maximum frequency spacing of F_{max} and a difference in amplitude of no more than $dB_{threshold}$.

Step 2: Obtain an initial estimate of the fundamental frequency F_0 from amplitude weighted combination of the frequency position of the harmonic peaks divided by the harmonic number $[F_{0k}/k\,,\ k = 1 \ldots N_h]$ (assumed to be the harmonics) of speech.

$$\hat{F}_0 = \sum_{k=1}^{N_h} W(F_{0k}) \frac{F_{0k}}{k} \tag{6.21}$$

where the weights are obtained as

$$W(F_{0k}) = X^{\propto}(F_{0k}) \Big/ \sum_{k=1}^{N_h} X^{\propto}(F_{0k}) \tag{6.22}$$

Where $X(f)$ is the frequency spectrum of the signal and $\propto$=1 or 2. Note the weights give more emphasis to estimates obtained from the more significant high amplitude harmonics.

Iterative Steps in a loop (3 & 4)

Switch to adaptive F_{min} and $F_{max,}$ iteration index t.

Step 3*:* Prune the peaks/peak positions using the constraints of

1) An adaptive minimum frequency spacing of $\beta\hat{F}_0$ between successive harmonics obtained from the current estimate of the pitch $\hat{F}_0$ and a maximum difference in peak level of $dB_{threshold}$. The value of β is experimentally selected in the range between $0.5 \leq \beta \leq 0.8$.
2) An adaptive maximum frequency spacing of $\gamma\hat{F}_0$ between successive harmonics obtained from the current estimate of the pitch $\hat{F}_0$ and a maximum difference in peak level of $dB_{threshold}$. The value of γ is experimentally selected in the range between $3 \leq \gamma \leq 4$.

3) A maximum amplitude difference threshold between successive peaks, peaks below $dB_{threshold}$ are deemed insignificant.

Step 4*:* Obtain an updated estimate of $\hat{F}_0$ from amplitude weighted combination of the frequency position of the pruned peaks (assumed to be the harmonics) of speech.

Convergence test: If the convergence ($Df = F_{0_t} - F_{0_{t-1}}$) criterion is not satisfied go to step3. Df may be set to a small value of 2 Hz. Where F_{0_t} is the pitch estimate at iteration t.

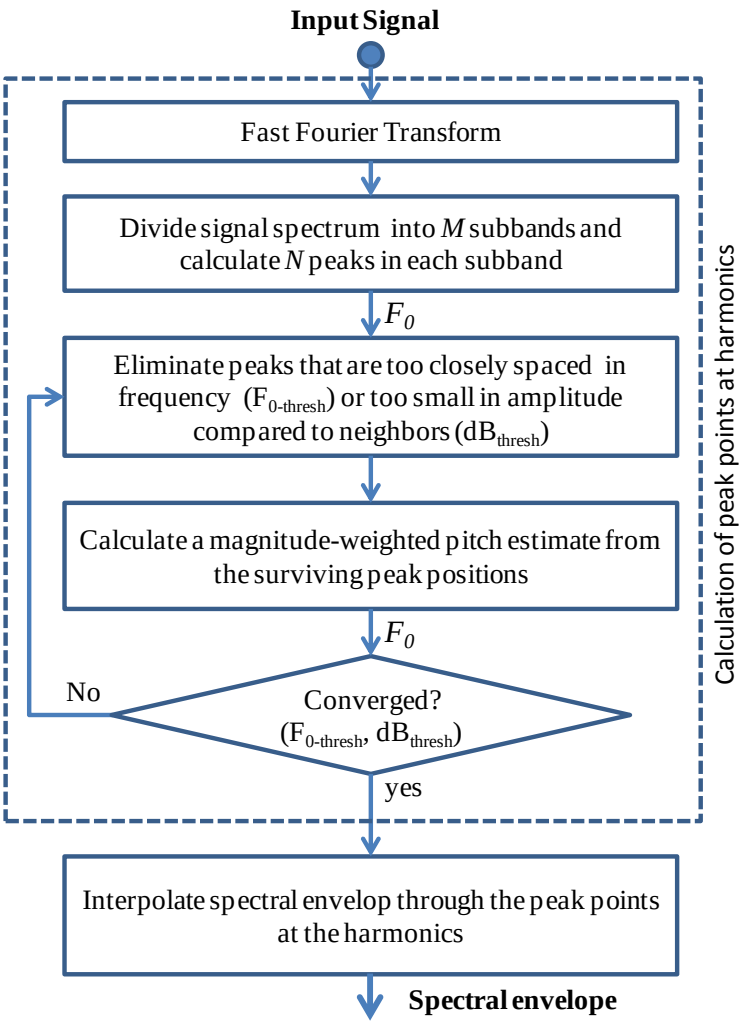


Figure 6.14 - Spectral envelope estimation algorithm

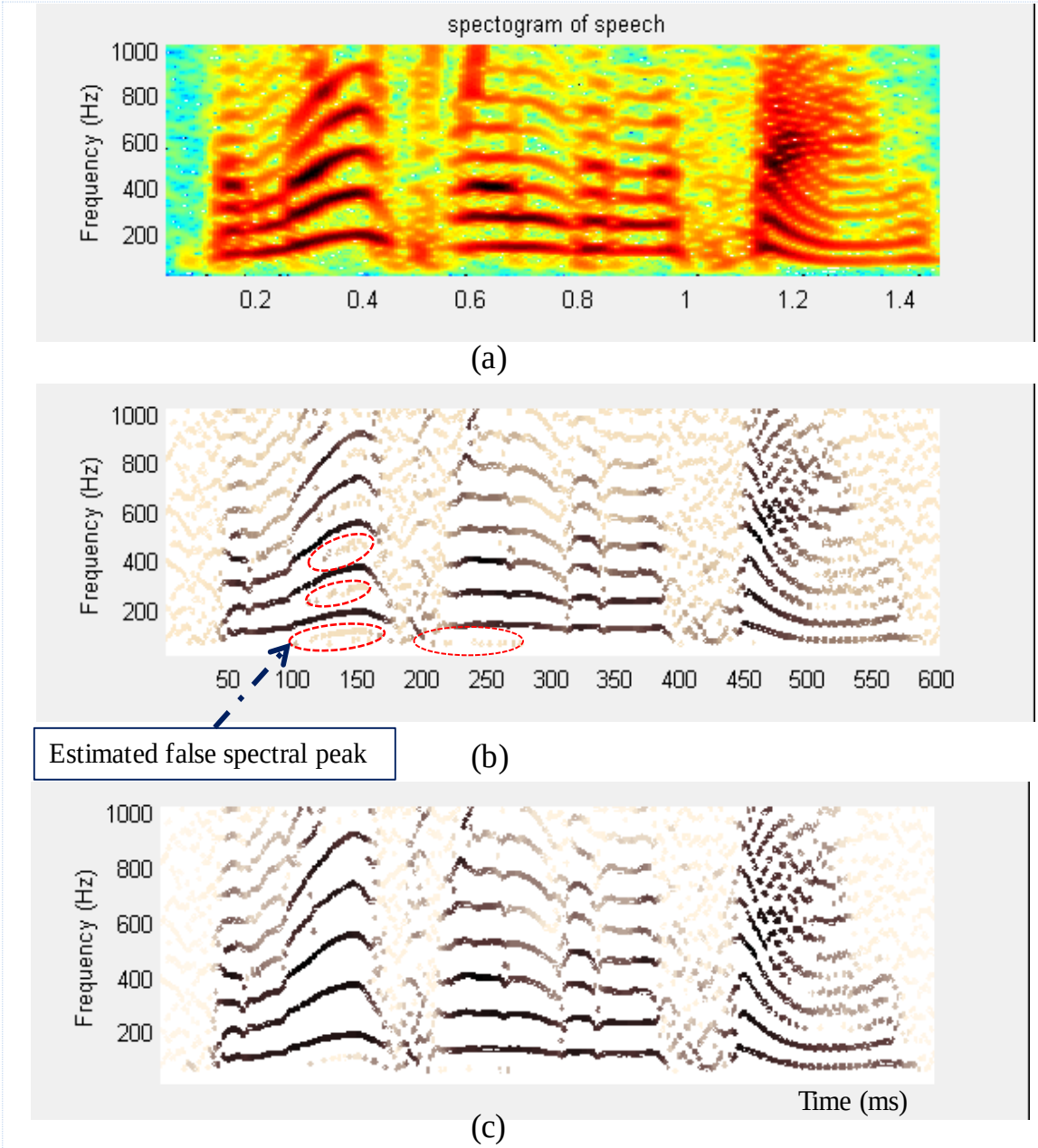


Figure 6.15 - Spectrogram of (a) the speech spectrum, (b) and (c) the pitch trajectory from spectral envelope using PCHIP interpolation pitch estimation method subject to the constraint of peak selection. The circles in (b) show areas with relatively higher proportion of false harmonic peak estimates which are improved by iterative post processing as shown in (c).

Figure 6.15 (a) shows the spectrogram of speech signal where the pitch trajectories are clearly visible. Figure 6.15 (b) are the initial estimate of the frequency-time trajectories of the spectral peaks that are used for estimation of the spectral envelop of speech, these estimates are obtained using the initial constraints on minimum frequency distance and maximum magnitude difference between the successive spectral peaks. As hoped, the spectral peaks of speech spectrum largely trace the pitch trajectories. However, there are instances of false peak detections some of which are shown encircled. Figure 6.15 (c) shows that some of the false spectral peaks are removed in the following iterations that impose constraints on the minimum and maximum frequency distance between the peaks and constraints on maximum magnitude difference between successive harmonics. A Piecewise Cubic Hermite Interpolation Polynomial (PCHIP) is used as the interpolation of the spectral envelop.

6.5 HARMONIC SIGNAL SYNTHESIS

For each proposed pitch candidate, $F_{0i}, i = 1 \cdots N$, the harmonic part of the speech signal is synthesised in the frequency domain as the product of the estimate of the spectral envelop $A(kF_{0_i})$ and the estimate of the harmonic excitation sequence $G_k(f - kF0_i)$ as expressed in equation (6.2)

$$\hat{X}_{hi}(f, F_{0_i}) = \sum_{k=1}^{N_h} A(kF_{0_i}) G_k(f - kF_{0_i}) \qquad i = 1 \cdots N \tag{6.23}$$

The N synthesised signal spectra, $\hat{X}_{hi}(f, F_{0_i})$ $i = 1 \cdots N$, are then compared with the original signal $X(f)$ to determine which of the N proposed candidates is best capable of generating the harmonic part of the signal.

Figure 6.16 shows an example of the synthesized harmonic spectrum of 2nd order moment (ACF) pitch method.

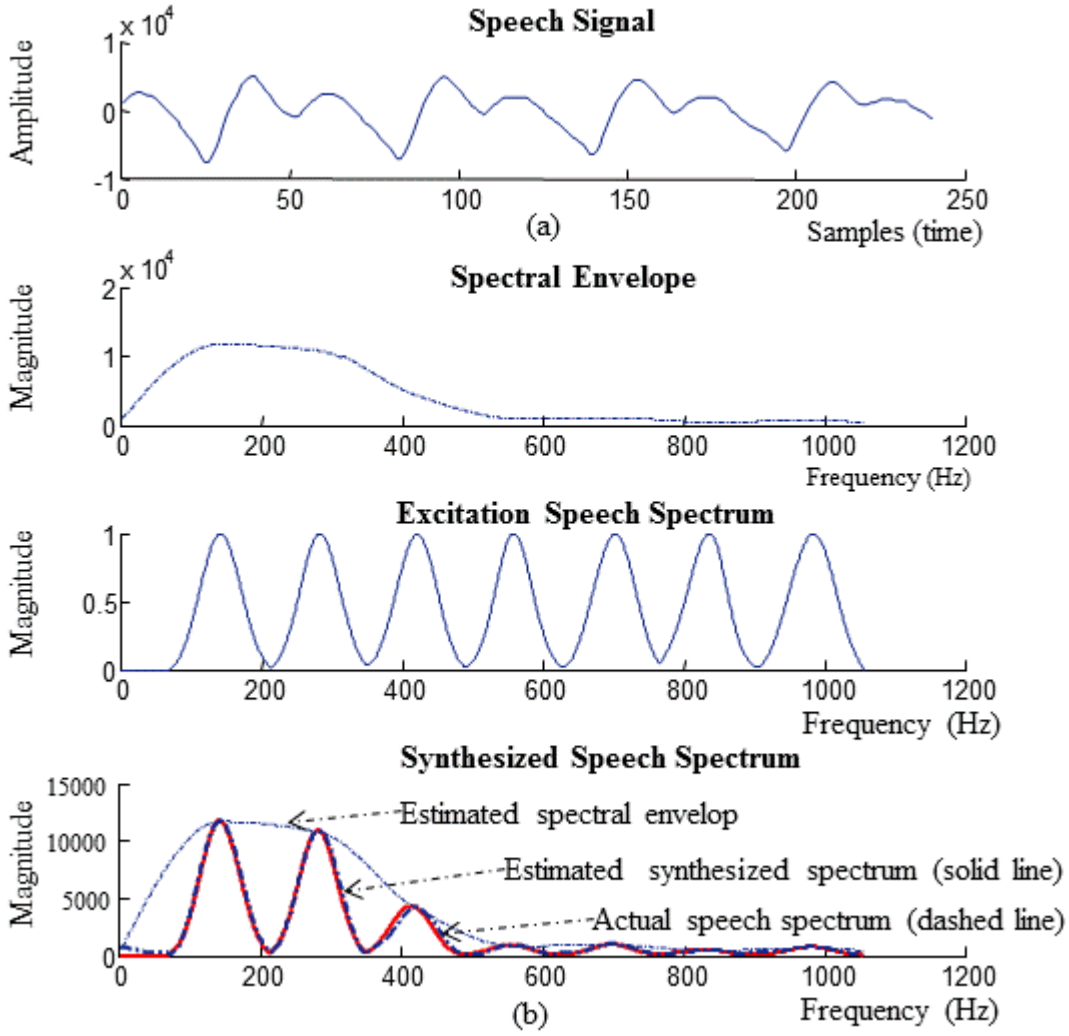


Figure 6.16 - Illustration of (a) the periodic speech signal in time, and (b) the spectral envelop fitted on the actual speech spectrum and the synthesised speech spectrum.

6.6 SPECTRAL DISTORTION MEASURES

For each pitch candidate F_{0i} a distortion function provides a numerical measure of the difference between the synthesized speech spectrum X_s and the actual speech spectrum X

$$D\left(F_{0_i}, X_s, \mathrm{X}\right) = \sum_{k=1}^{Nh} W\left(kF_{0_i}\right) d\left(kF_{0_i}, X_s, \mathrm{X}\right) \tag{6.24}$$

where $W(\cdot)$ are a set of weights and $d\left(kF_{0_i}, X_s, \mathrm{X}\right)$ is the part of the distortion incurred in the synthesis of the k^{th} harmonic of the proposed candidate

$$d(kF_{0i}, X_s, \mathrm{X}) = \sum_{l=-N1}^{N2} d(X_s(kF_{0i} + l), \mathrm{X}(kF_{0i} + l)) \tag{6.25}$$

where N_1 and N_2 denote the span of the k^{th} harmonic, in the following these are the mid-points between the successive harmonics.

The distortion measure is expected to have the following attributes

1) Reward the well fitted segments of the synthesised harmonics at the actual harmonic frequencies.
2) Penalise the poorly fitted segments for missing (double pitch) and false (half pitch) excitation pulses and for any other deviations of the synthesised harmonics from the actual harmonics.
3) Weight the good fits and the poor fits with a function that appropriately scales the distortion values.

The distortion measures explored in the following are based on the weighted harmonicity distance (WHD), weighted minimum mean squared error (WMMSE) and weighted signal to noise ratio (WSNR).

6.6.1 Harmonicity Distance

Harmonicity distance is used to measure the distortion of the harmonic structure of speech; a harmonicity contrast function may be defined in general terms as

$$H(n) = f(peak(n), trough(n)) \quad (6.26)$$

where $f(peak(n), trough(n))$ is a function of the peak at the n^{th} harmonic and the troughs on either side of the n^{th} harmonic. For example, one choice of the harmonicity measure is the ratio of the energy at a band of $(2N+1)$ frequency samples centred around the peak at the harmonic to the mean of the energy at a band of $(2N+1)$ frequency samples centred around the troughs on the two sides of the harmonic as

$$H(n) = \frac{2\sum_{l=-N}^{N} X^{\alpha}(F_0(n)+l)}{\sum_{l=-N}^{N} X^{\alpha}(midpoint\ (n-1,n)+l) + \sum_{l=-N}^{N} X^{\alpha}(midpoint\ (n,n+1)+l)} \quad (6.27)$$

Where $F_0(n)$ is the frequency of the peak spectrum at the n^{th} harmonic and the midpoints centred at troughs are defined as

$$midpoint\ (n-1,n) = 0.5\big(F_0(n-1) + F_0(n)\big) \quad (6.28)$$

$$midpoint\ (n,n+1) = 0.5\big(F_0(n) + F_0(n+1)\big) \quad (6.29)$$

where typically a = 1 or 2.

Alternatively one can use the following average of harmonicity on two sides

$$H(n) = \frac{0.5\sum_{l=-N}^{N} X^{\alpha}(F_0(n)+l)}{\sum_{l=-N}^{N} X^{\alpha}(midpoint\ (n-1,n)+l)} + \frac{0.5\sum_{l=-N}^{N} X^{\alpha}(F_0(n)+l)}{\sum_{l=-N}^{N} X^{\alpha}(midpoint\ (n,n+1)+l)} \quad (6.30)$$

Figures 6.17 show an example of voiced speech, their corresponding spectrum and their harmonicity.

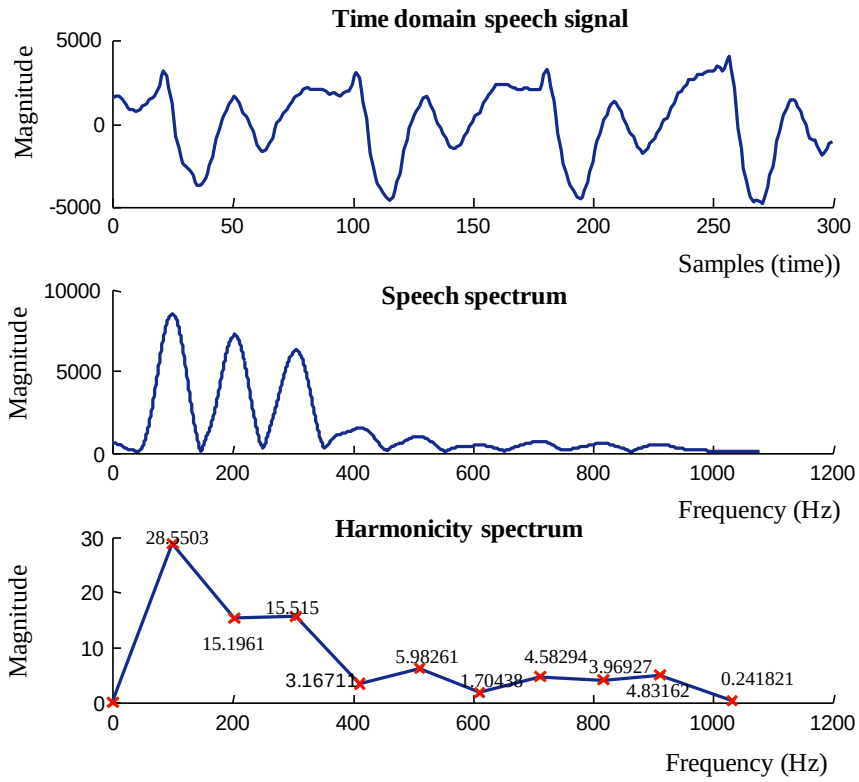


Figure 6.17 - The harmonicity at each peak spectrum of a segment of speech.

6.6.2 Minimum Mean Squared Error (MMSE) Distortion

The weighted MSE distortion measure is defined as the sum of the partial distortions at the harmonics as

$$MSE(F_{0i}, \mathrm{Xs}, \mathrm{X}) = \sum_{k=1}^{N_h} W(kF_{0i}) SE(kF_{0i}) \tag{6.31}$$

Where X and Xs are the actual and the synthesised spectra, N_h is the number of synthesised harmonics for a proposal F_{0i}.

The squared magnitude spectral error, SE, of the k^{th} proposed harmonic is given by

$$SE(kF_{0i}\,\mathrm{Xs},\mathrm{X}) = \sum_{l=mid(kF_{0i})}^{mid((k+1)F_{0i})} (X(l) - X_h(l))^2 \tag{6.32}$$

and the weights $W(kF_{0i})$, calculated as normalised average energy around each harmonic, are given by

$$W(kF_{0i}) = \frac{1}{mid\big((k+1)F_{0i}\big) - mid(kF_{0i})} \sum_{l=mid(kF_{0i})}^{mid((k+1)F_{0i})} |A(l)|^2 \tag{6.33}$$

where $A(l)$ is the spectral envelop. Alternatively, the spectral envelop weights can be calculated as the maximum of the original and the synthesised spectra as

$$W(kF_{0i}) = \max\big(X(kF_{0i}), X_s(kF_{0i})\big) \tag{6.34}$$

The best pitch candidate is chosen as

$$F_0 = \underset{F_{0i}}{argmin}\big(MSE(F_{0i},\mathrm{Xs},\mathrm{X})\big) \qquad i = 1,\dots,N \tag{6.35}$$

6.6.3 Weighted Signal to Noise Ratio Distortion

The weighted signal–to-noise ratio (WSNR) of the synthesized signal Xs relative to the original signal X is defined as the weighted sum of the segmental SNRs around the proposed harmonics as

$$WSNR(F_{0i},\mathrm{Xs},\mathrm{X}) = \sum_{k=1}^{Nh} W(kF_{0i}) SNR(kF_{0i},\mathrm{Xs},\mathrm{X}) \tag{6.36}$$

where the SNR for the k^{th} harmonics is given by

$$SNR(kF_{0i}, \mathrm{Xs}, \mathrm{X}) = 10log_{10}\left(\sum_{l=mid(kF_{0i})}^{mid((k+1)F_{0i})} |X(l)|^2\right) \Big/ \left(\sum_{l=mid(kF_{0i})}^{mid((k+1)F_{0i})} (X(l) - \mathrm{Xs}(l))^2\right) \tag{6.37}$$

Since $SNR(f)$ is a relative value of signal to noise ratio at frequency f, it is weighted by the estimate of the spectral envelope A as

$$SNR(kF_{0i}\ \mathrm{Xs}, \mathrm{X}) = W(kF_{0i})SNR(kF_{0i}\ \mathrm{Xs}, \mathrm{X}) \tag{6.38}$$

where

$$W(kF_{0i}) = \frac{1}{mid\big((k+1)F_{0i}\big) - mid(kF_{0i})} \sum_{l=mid(kF_{0i})}^{mid((k+1)F_{0i})} |A(l)|^2 \tag{6.39}$$

The consequence of weighting with the spectral envelop is that more rewards or penalties are given at high spectral envelop energy than at low spectral envelope energy as desired. The best pitch candidate is chosen as

$$F_0 = \underset{F_{0i}}{argmax}\big(WSNR(F_{0i}, \mathrm{Xs}, \mathrm{X})\big) \qquad i = 1, \ldots, N \tag{6.40}$$

Focused segmental SNR

The segmental SNR between synthesized harmonics and the original signal may be calculated at a focused region around the harmonics where the signal energy should be high and the expected contrast between a correctly synthesized and an incorrectly synthesized harmonic may be well pronounced. The modified SNR may be obtained as

$$SNR(kF_{0i}, \mathrm{Xs}, \mathrm{X}) = 10log_{10}\left(\sum_{l=kF_{0i}-delta)}^{kF_{0i}+delta} |X(l)|^2\right) \Big/ \left(\sum_{l=(kF_{0i}-delta)}^{kF_{0i}+delta} (X(l) - \mathrm{Xs}(l))^2\right) \tag{6.41}$$

where $delta = \alpha F_{0i}$, the choice of α is typically between $0.1 \geq \alpha \geq 0.25$.

6.6.4 *N*-Best Selection using Viterbi Network Process, F_0

The N-best pitch estimation is implemented using a selection method that finds the best value among *N* proposed candidates. Alternatively, for future work, the past history of *N* best proposed candidates pitch can be utilised within a Viterbi network.

6.6.5 *N*-Best Cost Functions

The Viterbi method finds the best route in a truncated distortion matrix computed as

$$E(F_{0_i}, t) = \sum_{f=0}^{Nh-1} d(X(f,t), X_{hi}(f,t))^2 \qquad i = 1, \ldots N \tag{6.42}$$

Where *t* is the speech frame number. Viterbi network is used to find the best sequence of states (*N*-Best pitch candidates). *N*-Best algorithm is a time-synchronous Viterbi-style pitch estimate procedure that is assured to find the *N* most likely pitch candidates that are within the pitch values prediction [145]**.**

6.7 EVALUATION AND PERFORMANCE ANALYSIS

This section provides an analysis of the distances from the true pitch, of the positions of the N top extrema points of the similarity functions, used for pitch extraction, in terms of the number of times each of the extrema points is closest to the actual pitch value. The similarity criterions considered are the autocorrelation method ACF (2nd order moment),

the average magnitude difference function AMDF and the higher order moment methods HOMs.

The N-best candidate pitch estimation method has been applied to the following moment criteria; the 2^{nd} to 5^{th} order moments and the average magnitude difference function (AMDF). The results were compared with N-best =1, and the YIN Method.

The speech sampling rate for the evaluation is 8000 Hz and the file speech databases format is 16 bits precision .wav. Prior to signal moment analysis, the speech signals are band-pass filtered to a range of 40-2000 Hz. The window lengths selected for different moments are as follows:

1) 2^{nd} order (ACF) and AMDF methods, 30 ms (240 samples);
2) 3^{rd} order moment, 37.5 ms (300 samples), and
3) 4^{th} order and 5^{th} order moments, 62.5 ms (500 samples)

The window overlap parameter is set to a value of 5 ms (40 samples) for all methods, this equates to 200 updates per second of the pitch estimate.

The similarity criteria, using various moments, were calculated in the range $T_{min} = 2.5\ ms$ (20 samples) and $T_{max} = 25\ ms$ (200 samples). The *N*-best candidates for each similarity moment criteria are obtained as the *N* top peaks of the similarity curve.

The N-best candidates together with the speech segment are processed to determine which pitch candidate can best facilitate the synthesis of the harmonic components of the speech signal.

The synthesis of harmonic part of the signal is performed in frequency domain. First, the time domain signal is transformed to frequency domain via FFT with a frequency resolution of 1 Hz. The next stage is to readjust the nominal values of the harmonic frequencies for each candidate kF_{0i} using a local peak search in the range $kF_{0i} \pm \Delta$ where Δ is typically 10 - 20% of F_{0i}.

Subsequently a Gaussian train of harmonic excitation pulses are placed at $kF_{0i} \pm \Delta$. Each Gaussian pulse is fitted to the shape of the speech harmonic pulse using the asymmetric Gaussian pulse model definition and estimation described in section 6.3.

For estimation of the envelope of the signal spectrum, the method described in section 6.4 is used. First the spectral bandwidth (of 4000 Hz) is divided into $N = 15$ segments and in each segment the two highest peaks are selected, here a peak is defined as a turning point that stands above neighbouring samples for a range of $\pm\Delta$ (typically 20) samples. Next, the most significant spectral peaks, subject to the constraints of the maximum allowable magnitude difference and the minimum allowable position (frequency) difference from the neighboring peaks, are obtained. This is an iterative peak pruning process. Finally, the spectral envelop is interpolated through the spectral peaks.

The spectral envelop and the harmonic excitation Gaussian pulses are multiplied to obtain the synthesized signal spectrum.

6.7.1 Analysis of the *N*-best Candidates Compared with True Pitch Value

This section provides justification for N-best strategy in the form of statistical analysis of the number of times each of the *N*-best candidates is closest to the true pitch value. For the purpose of this analysis the N-best are arranged in order of the increasing peak value

of the similarity moment criteria and hence the choice of the first candidate represents the conventional method of selection of the top peak for the pitch value.

For the 2nd moment and the 3rd moments criteria of pitch estimation, Figures 6.18 and 6.19 show the histograms of the errors and the mean squared errors (MSE) for the two cases when the pitch values are derived from the position of (a) the top peak of the moment curve and (b) the peak of the moment curve that is nearest to the true pitch value obtained from laryngograph. As expected case (b), that is the choice of the best peak of the curve, provides reduced pitch estimation error and a better distribution of errors in that there are less large more noticeable errors.

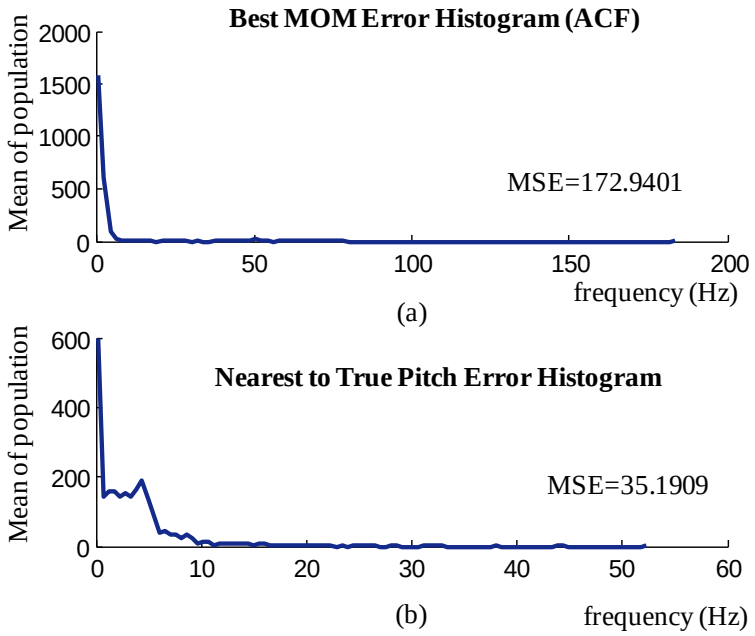


Figure 6. 18 - The histograms of the 2nd moment (ACF) pitch estimation error and the MSE values: (a) for the pitch estimate at the maximum peak of the moment curve, (b) for the pitch estimate at the moment peak that is nearest to the true pitch.

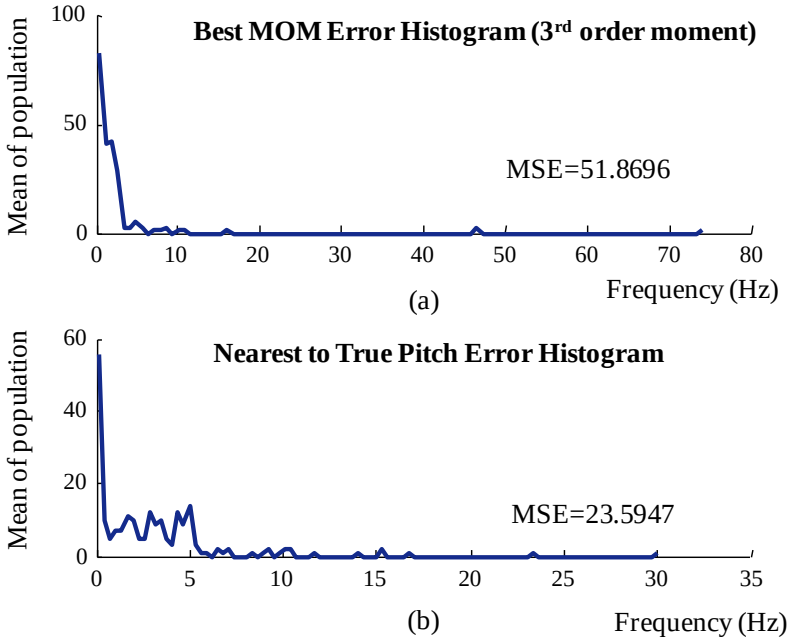


Figure 6. 19 - Illustration of the histogram of the 3rd order moment pitch estimation error and MSE values: (a) for the pitch estimate at the maximum peak of the moment curve, (b) for the pitch estimate at the moment peak that is nearest to the true pitch.

Figure 6.20 illustrate the percentage of frames of the best peak closest to the true pitch at the peak of n=1 to n=7 for ACF and third order moment of similarity criteria.

Next are the results finding of *N*-Best pitch estimation using three different distortion measures; (i) weighted signal-to-noise, WSNR, (ii) minimum mean squared error, MMSE , and (iii) combination of WSNR, MMSE, and weighted harmonicity distance, WHD. The comparison graphs of these three distortion measure are presented in Appendix A for further reference.

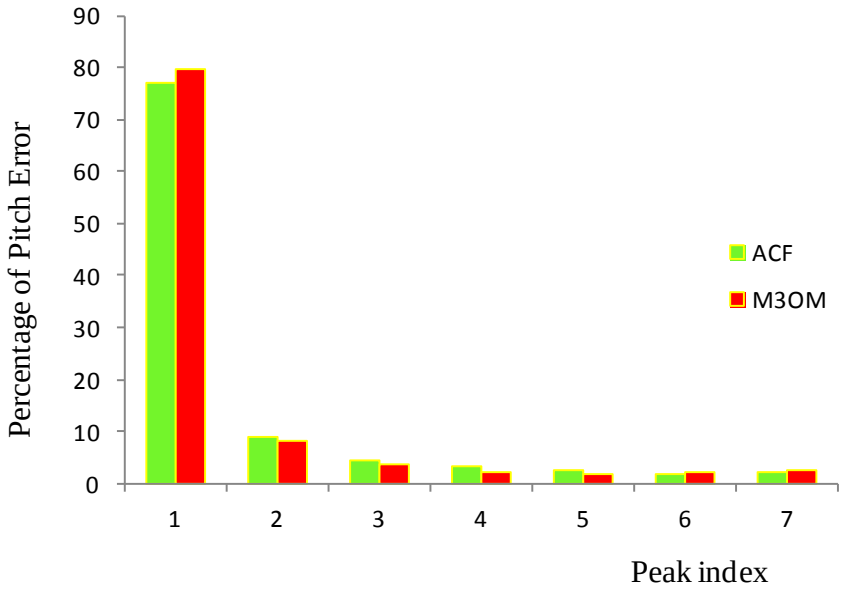


Figure 6. 20 - The % of frames for which the true pitch is closest to the n^{th} best peak where n=1:7 and the peaks are arranged in descending order.

i) Weighted Signal-to-Noise Ratio Distortion Measure

From Figure 6.21 displays the average of similarity criteria of pitch estimation methods using weighted signal-to-noise ratio, WSNR distortion measure.

The pitch errors decrease consistently with the increasing number of the N-Best candidates from *N*-Best =1 to 7 for the error criteria considered in this thesis.

For the modified third order moment, the pitch estimation error is reduced by 10% as *N*-Best increases from 1 to 7. Likewise, the error for autocorrelation and modified autocorrelation methods is reduced by 14% as *N*-Best increases from 1 to 7. The AMDF is improved by 30% with the increasing number of the *N*-Best.

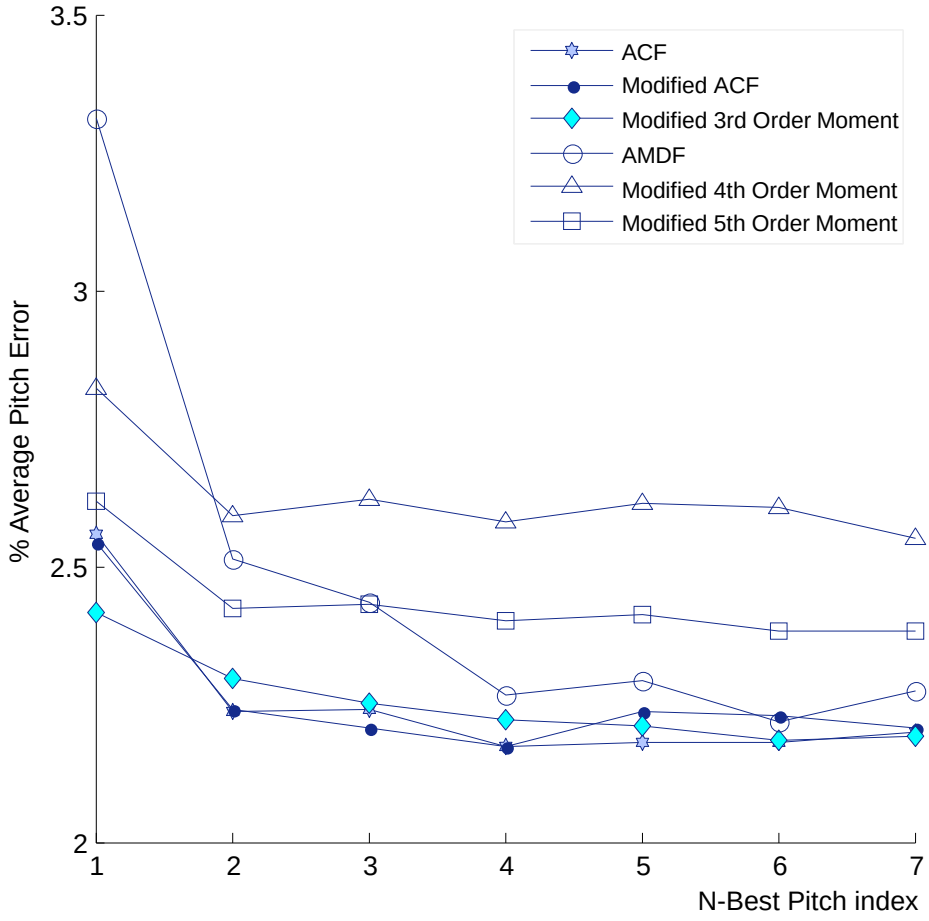


Figure 6. 21 - The % *N*-Best overall pitch error with weighted SNR distortion measure.

Figure 6.22 shows the values of the percentage gross pitch errors, with WSNR distortion measure using various similarity criteria for pitch estimation. In Figure 6.22, the *N*-Best pitch estimation based on conventional ACF and MHOMs result in a significantly lower percentage of population of the gross pitch errors (that are greater than 20%) compared to the bench mark YIN method. The pitch estimation error are reduced about 32% of ACF method, likewise 23% from modified 3rd order moment, and almost 64% from AMDF method as the *N*-Best index increases.

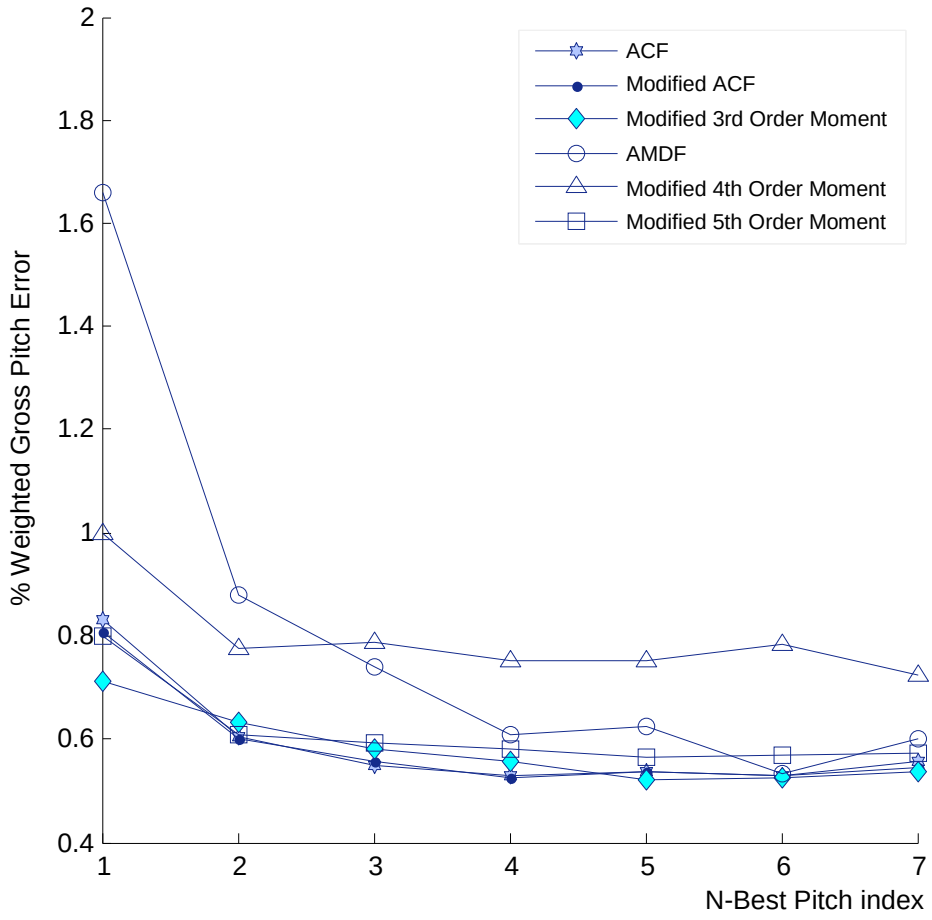


Figure 6. 22 - The % *N*-Best gross pitch error with the weighted SNR distortion measure as a function of *N*-Best pitch index.

The population gross pitch error using WSNR distortion measure as in Figure 6.23 shows the improvement in conventional and proposed similarity criteria as the *N*-Best pitch index increases.

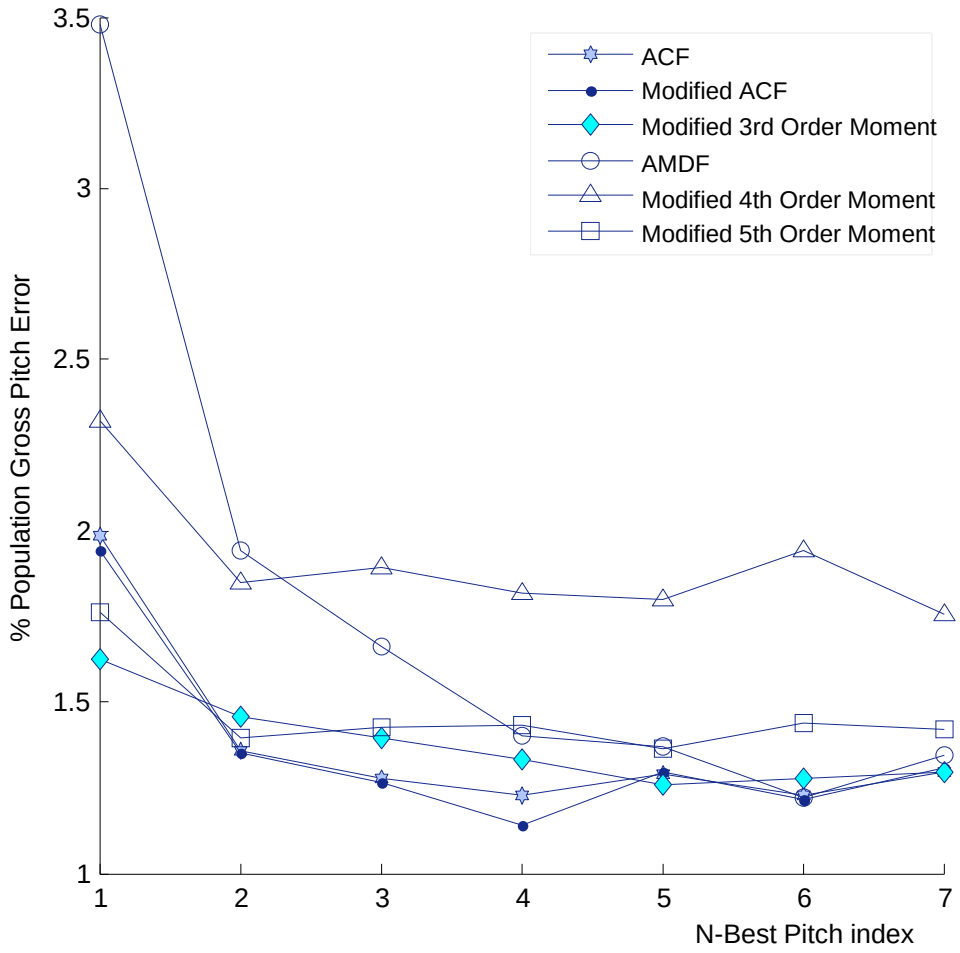


Figure 6. 23 - The % *N*-Best population gross pitch error with weighted SNR distortion measure.

The standard deviation of pitch estimation error as in Figure 6.24 shows the improvement in conventional and proposed similarity criteria as the *N*-Best pitch candidate increases from 1 to 7.

The smallest standard deviation is achieved for the modified third order moment criteria.

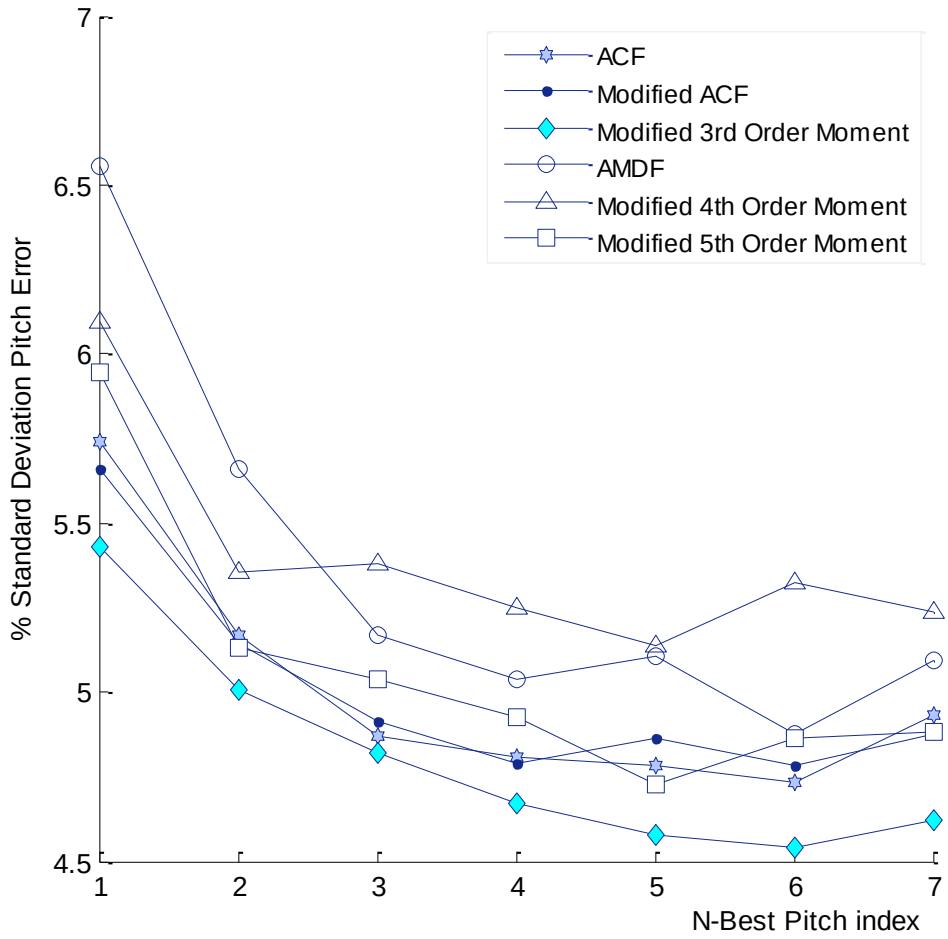


Figure 6. 24 - The % *N*-Best standard deviation of pitch error with the weighted SNR distortion measure.

ii) Fine - Tuning Around Best Choice of *N* Candidates

The *N*-Best strategy can be reused for fine-tuning around the best pitch value $F\_0 - F_{0_{Best}}$. In this method given the value of $F_{0_{Best}}$, we can generate a new set of N closely spaced candidate values around $F_{0_{Best}}$ within a predetermined range (such as the 1% of the best value). The range of $(F_{0_{Best}} \pm \Delta F_0)$ may be divided into N closely spaced values and the spectral analysis-synthesis method is used to select the best value.

An alternative fine tuning, is to increase apparent spectral ‘resolution’ by resampling the speech spectrum at a high rate and then recalculating the position of the peaks of the first M harmonics in a search range that is constrained to lie within say 1.5% of the spectral peaks corresponding to pitch estimate and its harmonics. The increase in spectral resolution can be obtained by zero-padding speech or by resampling the spectrum around the peaks using a linear interpolation method.

Then harmonic amplitude weightings are employed to combine the refined estimates. This method offers further reduction in pitch error as shown in Figure 6.25 and Figure 6.26. Figure 6.25 shows the reduction in pitch estimation error about 3.5% for ACF and in Figure 6.26 shows the improvement of 2.5 % for modified third order moment method.

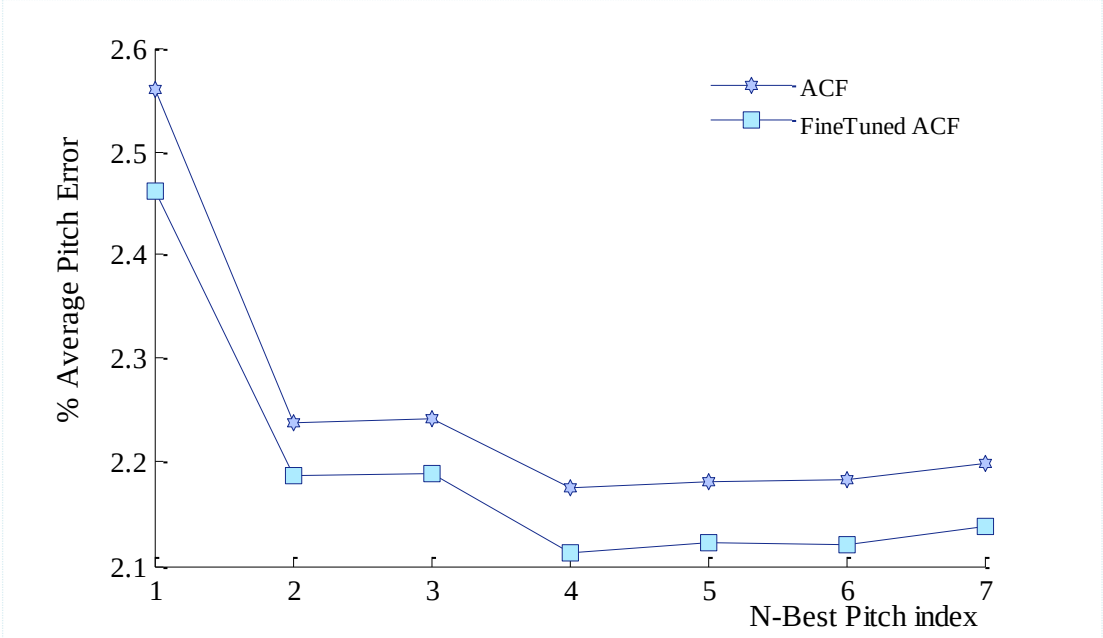


Figure 6. 25 - The ACF and fine-tuned ACF with weighted SNR as a function of N-Best pitch index.

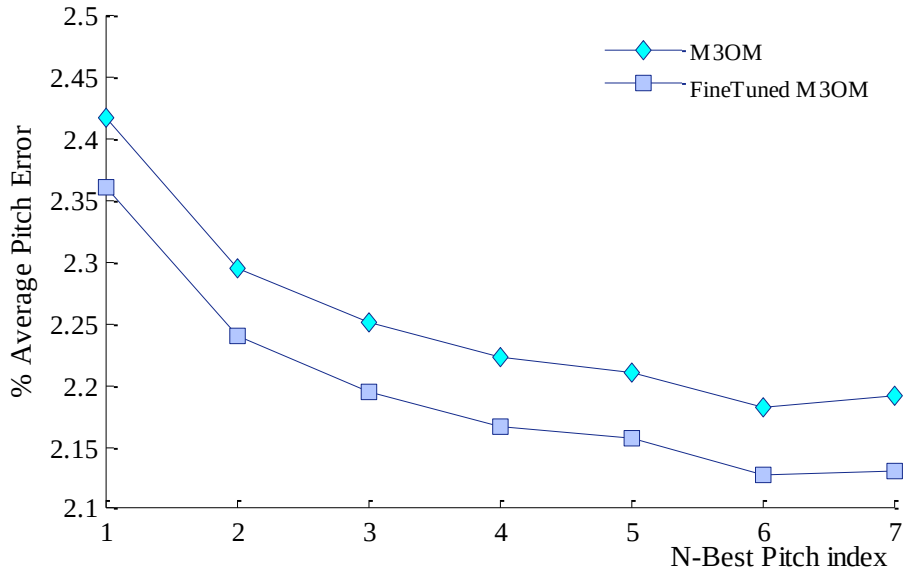


Figure 6. 26 - The 3rd moment and fine-tuned 3rd moment with weighted SNR distortion measure as a function of *N*-Best pitch index.

iii) Minimum Mean Squared Error (MMSE) Distortion Measure

Figure 6.27 to Figure 6.30 show the pitch estimation results of *N*-Best pitch error with minimum mean squared error (MMSE) distortion measure for the percentage overall pitch error, percentage gross pitch error, percentage population gross pitch error and standard deviation pitch error as the function of *N*-Best pitch candidate increases. These results show the improvement in pitch estimation as a function of the increasing number of *N*-Best candidate.

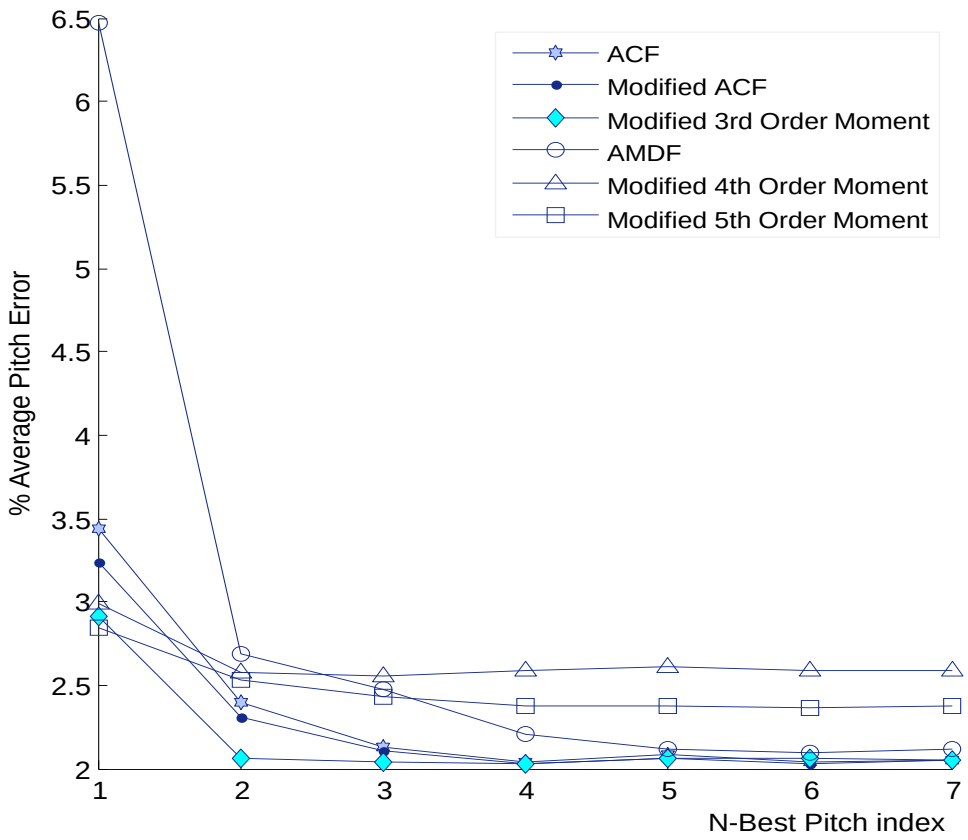


Figure 6. 27 - The % overall pitch error with MMSE distortion measure.

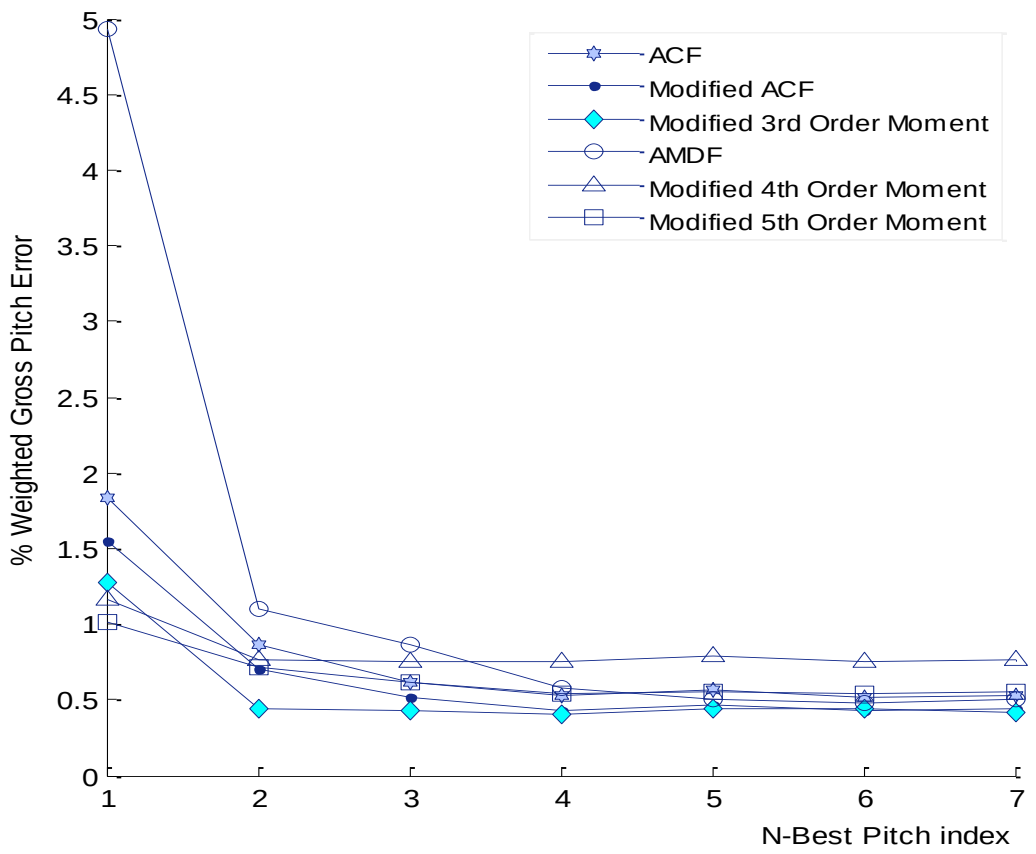


Figure 6. 28 - The % gross pitch error with MMSE distortion measure as a function of pitch index.

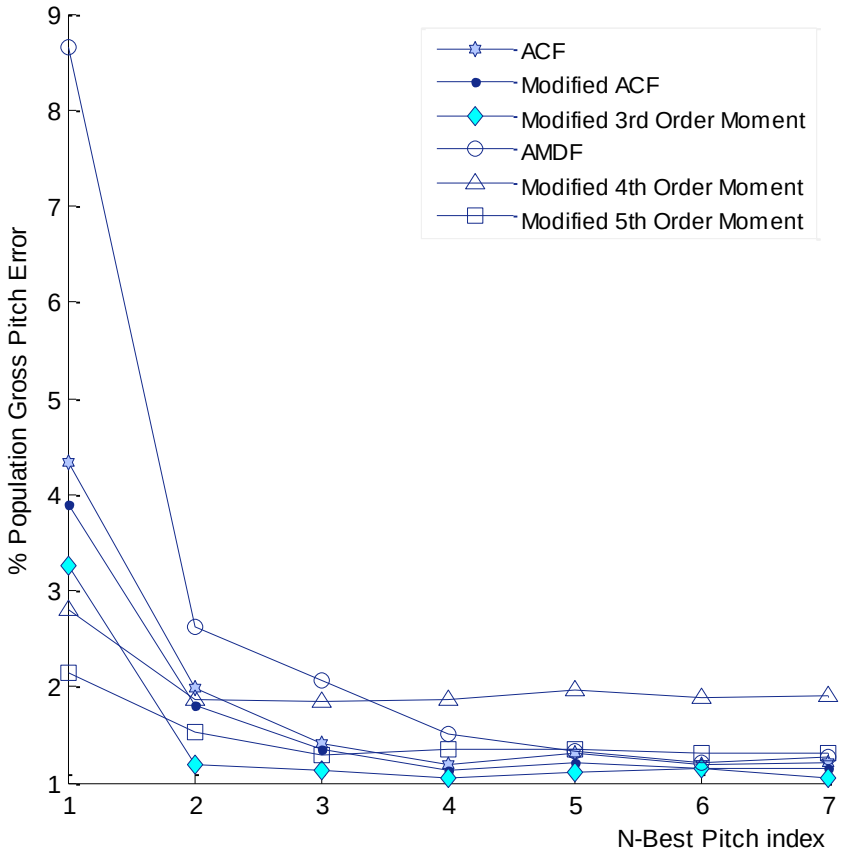


Figure 6. 29 - The % population gross pitch error with MMSE distortion measure as a function of pitch index.

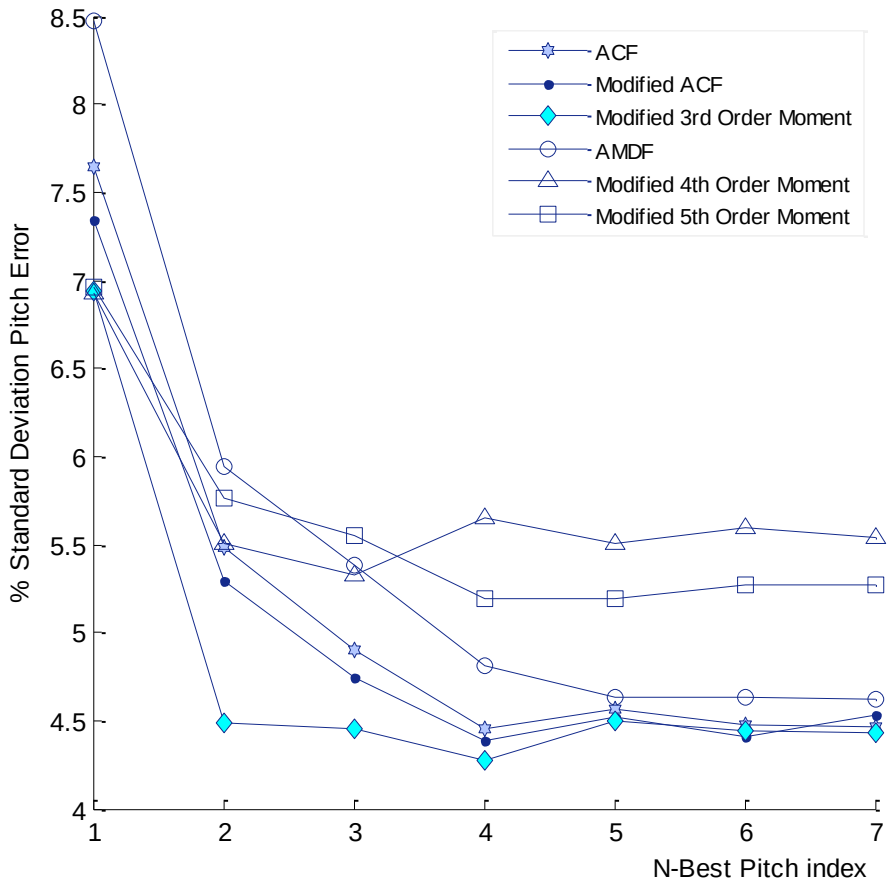


Figure 6. 30 - The % standard deviation pitch error with MMSE distortion measure as a function index.

6.8 CONCLUSION

In this chapter, we proposed an analysis-synthesis method for robust and smooth trajectory pitch or fundamental frequency estimation, where the N-Best pitch candidates are estimated. The proposed method reduces the double pitch and half pitch error challenges.

The selection of the best pitch value among the *N* top candidates is investigated. Generally, a similarity criteria yields *N* candidates taken from the peaks (or minima) of the criteria and the selection of the best candidate may be facilitated with such constraints as continuity and smoothness of the pitch trajectory curve. A novel approach is proposed in which each of the *N* pitch frequency candidates are used to synthesis the spectrum of the harmonic part of the speech. The candidate that achieves the minimum distortion synthesis is selected as the best candidate.

For this approach to work optimally, the synthesis method should be designed such that the resulting distortion increases with increasing error in pitch frequency. The spectral synthesis method is based on the product of spectral envelop derived from the speech signal and harmonic excitation composed of a series of asymmetric Gaussian pulses positioned at adjusted, refined, estimates of the fundamental frequency and its harmonics.

The choice of the spectral envelope is a critical aspect of the synthesis method. The envelope should be such that it passes through the major peaks at the fundamental harmonics but it should not pass through the spectral peaks that are in between the harmonics. To find an appropriate algorithm for the spectral envelope is a challenging nontrivial task, as the conventional spectral envelope extraction methods such as the linear prediction or cepstral method are unsuitable choices for a spectral synthesis method that ultimately aims to select the best pitch value based on the spectral synthesis distortion.

The synthesis of spectral excitation is also a challenging task. The shape of the excitation pulses, in frequency domain, are modelled by asymmetric Gaussian pulses whose

variances, on both sides, are derived such that the Gaussian pulse fits the actually pulse shape at the proposed harmonic frequency.

Three distortion measures are experimented for evaluation of the *N*-best pitch trajectory such as the weighted signal-to-noise ratio, WSNR, the weighted minimum mean squared error MMSE, and the combination of the WSNR, weighted MMSE and weighted harmonicity distance, WHD.

The results show the improvement of pitch estimation error for all distortion measures as the *N*-best pitch candidate increases.

7

CONCLUSIONS AND FURTHER WORK

The main contributions of this research are three folds: (1) an investigation of the variation of the pitch error with increasing window length for various similarity criteria, (2) an investigation of the influence of the choice of the similarity criteria on pitch estimation accuracy and (3) a novel method of spectral analysis-synthesis for selection of the best pitch value among *N* candidates.

The issue of window length is a relatively simple and yet somewhat under-explored issue. The experimental finding is that the pitch estimation error initially decreases sharply when the window length is increased from the conventional choice of 20 ms (standard in mobile phones) and then rate of decrease of pitch error levels off until it reaches a minimum which is at about 37 - 80 ms depending on the type of similarity criteria used. Increasing the window length, beyond the point at which minimum pitch error occurs,

results in some increase in error due to the non-stationary characteristics of speech and the time varying trajectory of pitch.

The second issue investigated in this thesis is the impact of the choice of the similarity criteria on pitch estimation error. Two distinct contributions were made in this respect; (1) exploring the use of the higher order moments methods in reducing pitch estimation error, (2) a new method of calculation of moments named modified higher order moments wherein the signal is split, rectified, into positive and negative halves, before the moments are calculated. Interestingly the modified higher order method for the 2^{nd} order moment performs better that the conventional correlation method.

The third major issue investigated in this thesis is the selection of the best pitch value among the *N* top candidates. Generally, a similarity criteria yields *N* candidates taken from the peaks (or minima) of the criteria and the selection of the best candidate may be facilitated with such constraints as continuity and smoothness of the pitch trajectory curve. A novel approach is proposed in which each of the *N* pitch frequency candidates are used to synthesis the spectrum of the harmonic part of the speech. The candidate that achieves the minimum distortion synthesis is selected as the best candidate.

For this approach to work optimally, the synthesis method should be designed such that the distortion increases with increasing error in pitch frequency. The spectral synthesis method is based on the product of the spectral envelop derived from the speech signal and harmonic excitation composed of a series of asymmetric Gaussian pulses positioned at adjusted, refined, estimates of the fundamental frequency and its harmonics.

The choice of the spectral envelope is a critical aspect of the synthesis method. The envelope should be such that it passes through the major peaks at the fundamental harmonics but it should not pass through the spectral peaks that are in between the harmonics. To find an appropriate algorithm for the spectral envelope is a challenging nontrivial task, as the conventional spectral envelope extraction methods such as the linear prediction or cepstral method are unsuitable choices for a spectral synthesis method that ultimately aims to select the best pitch value based on the spectral synthesis distortion.

The iterative method of spectral envelope estimation introduced in this thesis uses a number of constraints regarding the pivotal peak points through which the spectral envelop pass. These constrains include the total number of peak points, the number of peak points per spectral segment, the length of the spectral segments, the minimum and maximum distances of the spectral peaks and their associated frequencies. Ideally the envelope should go through the peaks at the fundamental and harmonic points which are actually the unknowns. In practice constrains can be chosen such that a trade-off is achieved between miss rate (when a spectral peak at a harmonic is missed) and a false alarm rate (when a spectral peak in between the harmonics is selected). For the purpose of pitch estimation in this thesis, it is less harmful to have a miss-rate than a false-alarm.

The synthesis of spectral excitation is also a challenging task. The shape of the excitation pulses, in frequency domain, is modelled by asymmetric Gaussian pulses whose variances, on both sides, are derived such that the Gaussian pulse fits the shape at the proposed harmonic frequency of speech.

Examination of the shape of speech spectrum at the harmonics shows that the spectrum, at the two sides of the peak at harmonic, are asymmetric, i.e. having different slopes and bandwidth of rate of decay. This asymmetry of harmonic shape is due to a number of reasons including the glottal pulse shape and the influence of resonances and anti-resonances of the vocal tract.

In this thesis a novel method has been used to model the harmonic pulse shapes. The asymmetric Gaussian pulses are obtained by merging of two halves of Gaussian pulse of different variances. The variance of the pulses are derived from the width of the speech harmonic at such points where the harmonic spectral amplitude drops to half the peak (maximum) value (alternatively other factors such as 0.75 or 0.25 of the maximum may be used)

The choice of distortion measure is important for the selection of the best pitch value, among the N-best likely candidates, as the one that facilitates the minimum distortion spectral synthesis of the harmonic part of speech. This work explored several weighted spectral-segmental distortion measures, a spectral segment consisting of the bandwidth of frequencies around a harmonic. The weighted segmental SNR and the weighted segmental MSE perform particularly well. The use of harmonicity as a further discriminator of distortion was also explored.

The overall impact of the spectral analysis-synthesis method proposed for N-best selection is a marked improvement in error, the error decreases by some 50% or more relative to the case when the top peak of the similarity criterion is selected.

In conclusion pitch extraction is a challenging problem that has been subject of 40 years of research. This thesis has made contribution in systematic investigation of the impact of window length, similarity criteria and method of selection of the best pitch among the *N* top candidate. Much more remains to be done however it will be as is the nature of research an incremental process towards increasing better pitch extraction systems.

FURTHER WORK

The three main issues considered in this thesis can be subject of further research investigations.

On the issue of windowing, one can consider the simultaneous use of several overlapping windows of different lengths having different time-frequency resolutions. The issues of the choice of window length, efficient combinations of the pitch estimates from different windows and the relationship to wavelet analysis are interesting challenges that may arise.

On the issue of similarity criteria, one can also investigate composite similarity criteria, i.e. those that are combinations of several different similarity criteria. We have shown that a new method of splitting the signal into positive and negative valued part and combining the similarity criteria for each part provides improve results. There is room for a more in-depth analytical and experimental exploration of such an approach and its variants.

On the issue of *N*-best, the spectral analysis-synthesis method and the distortion so calculated for each candidate can be a component of a composite cost function that may include other costs such as continuity of pitch trajectory within a Viterbi dynamic optimization network.

REFERENCES

[1] Haskins Laboratories THE SCIENCE OF THE SPOKEN. [Online]. http://www.haskins.yale.edu/CaseStatement/Haskinscase.pdf.

[2] J. L. Miller, "Interactions in Processing Segmental and Suprasegmental Features of Speech," *Perception & Psychophysic*, vol. 24, no. 2, pp. 175-180, 1978.

[3] Wang, L.; Ambikairajah, E.; and Choi, E.H.C., "Automatic Tonal and Non-Tonal Language Classification and Language Identification Using Prosodic Information," in *IEEE International Conference on Multimedia and Expo, 2007*, pp. 352-355, 2006.

[4] S. Vaseghi, *Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications*.: John Wiley, 2007.

[5] Ohala, J. J. and Ewan, W.G., "Speed of Pitch Change," *Acoustical Society of America*, vol. 53, p. 345(A), 1973.

[6] S. Amano, T. Nakatani and T. Kondo, "Fundamental Frequency of Infants` and Parents` Utterances in Longitudinal Recordings," *Acoustical Society of America*, vol. 119, no. 3, pp. 1636-1647, 2006.

[7] M. Nishio and S. Niimi, "Changes in Speaking Fundamental Frequency Characteristics with Aging," *International Journal of Phonatrics, Speech Therapy and Communication Phatology*, vol. 60, no. 3, pp. 120-127, 2008.

[8] Vocal Register. [Online]. http://en.wikipedia.org/wiki/Vocal_register.

[9] *ITU-T G.114 Telecommunication Standard Sector of ITU*., 05/2003.

[10] H. Traunmüller and A. Eriksson, "The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults," , 1994.

[11] Speech Disorder [Online].http://en.wikipedia.org/wiki/Speech_disorder.

[12] C. Manfredi, M. D'Aniello, P. Bruscaglioni, A. Ismaelli, "A Comparative Analysis of Fundamental Frequency Estimation Methods with Application to Pathological Voices," *Medical Engineering & Physics*, vol. 22, no. 2, pp. 135-147, 2000.

[13] L. R. Rabiner and R. W. Shafer, *Theory and Applications of Digital Speech Processing*, 1st ed.: Pearson, 2011.

[14] A. M. Kondoz, *Digital Speech*, 2nd ed.: John Wiley & Sons, Ltd, 2004.

[15] J. J. Dubnowski, R. W. Schafer and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 1, pp. 2-8, 1976.

[16] M. M. Sondhi, "New Methods of Pitch Extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, no. 2, pp. 262-266, June 1968.

[17] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGoneagal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 5, pp. 399-417, 1976.

[18] P. Boersma, "Accurate Short-Time Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of the Sampled Sound," in *in IFA Proceeding 17*, 1993, pp. 97-110.

[19] Y. Stylianou, "Modeling Speech Based on Harmonic Plus Noise Models," *Springer-Verlag Berlin Heidelberg*, pp. 244-260, 2005.

[20] R. Ahn and W. H. Holmes, "Harmonic-Plus-Noise Decomposition and its Application in Voiced/Unvoiced Classification," in *IEEE TENCON- Speech and Image Technologies for Computing and Telecommunications*, pp. 587-590, 1997.

[21] H. Hong, Z. Zhao, X. Wang and Z. Tao, "Detection of Dynamic Structures of Speech Fundamental Frequency in Tonal Languages," *IEEE Signal Processing Letters*, vol. 17, no. 10, pp. 843-846, 2010.

[22] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, pp. 24 - 33, 1977.

[23] D. Talkin, "A Robust Algorithm for Pitch Tracking," *Elsevier: Speech Coding and Synthesis*, pp. 495-518, 1995.

[24] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed.: Wiley, 2009.

[25] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Alan V. Oppenheim, Ed.: Prentice-Hall, 1978.

[26] D. J. Liu and C.T. Lin, "Fundamental Frequency Estimation Based on the Joint Time-Frequency Analysis of Harmonic Spectral Structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 609-621, 2001.

[27] G. Jovanovic, "A New Algorithm for Speech Fundamental Frequency Estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 626-630, 1986.

[28] C. Shahnaz, W-P. Zhu and M. O. Ahmad, "On the Estimation of Pitch of Noisy Speech Based on Time and Frequency Domain Representations," in *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2008)*, pp. 1819-1822, 2008.

[29] J. Darch and B. Milner, "A Comparison of Estimated and MAP-Predicted Formants and Fundamental Frequency with a Speech Reconstruction Application," in 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), pp. 542-545, 2007.

[30] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, 1992.

[31] S.- H. Chen and J.- F. Wang, "Extraction of Pitch Information in Noisy Speech Using Wavelet Transform with Aliasing Compensation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 89-92, 2001.

[32] X. Rodet, "Speech Analysis and Synthesis Methods Based on Spectral Envelopes and Voiced/Unvoiced Functions," in *European Conference on Speech Technology*, 1987.

[33] E. Chilton and B. G. Evans, "The Spectral Autocorrelation Applied to the Linear Prediction Residual of Speech for Robust Pitch Detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP`88)*, pp. 358-361, 1988.

[34] A. de Cheveigne´ and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Acoustical Society of America*, vol. 111, no. 4, pp. 1917-1930,

2002.

[35] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-20, no. 5, pp. 367-377, 1972.

[36] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 727-730, 1999.

[37] A. M. Noll, "Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 296-302, 1964.

[38] C. Shahnaz, W. P. Zhu, and M. O. Ahmad , "Robust Pitch Estimation at Very Low SNR Exploiting Time and Frequency Domain Cues," in *IEEE International Conference on Acoustics, Speech, and Signals Processing (ICASSP)*, pp. I-389 - I-392, 2005.

[39] J. O. Hong and P. J. Wolfe, "Model-Based Estimation of Instataneous Pitch in Noisy Speech," in *The IEEE International Symposium on Circuits and Systems (ISCAS)*, 2009.

[40] K. J. U. Ahmed and M. R. Khan, "Estimation of Pitch of Noisy Speech Using AR Model Based Inverse Filtering," in *4th International Conference on Electrical and Computer Engineering ICECE 2006*, pp. 447-450, 2006.

[41] V. Mahadevan and C. Y. Espy-Wilson, "Maximum Likelihood Pitch Estimation Using Sinusoidal Modeling," in *IEEE International Conference on Communications and Signal Processing (ICCSP 2011)*, pp. 310-314, 2011.

[42] B. Doval, X. Rodet, "Fundamental Frequency Estiamtion and Tracking Using Maximum Likelihood Harmonic Matching and HMMs," *IEEE International Conference on Acoustics, Speech, and Signal Processig (ICASSP-93)*, pp. 221-224, 1993.

[43] P. M. B. Gambino and I. S. Burnett, "Low Delay Pitch Detection Using Dynamic-Programming/Viterbi Techniques," in *International Symposium on Signal Processing and its Applications, (ISSPA)*, pp. 77-80, 1996.

[44] X Sun, "Pitch Determination and Voice Quality Analysis Using Subharmonic-to-Harmonic Ratio," in *IEEE International Conference on Acoustics, Speech, and*

Signal Processing, (ICASSP), pp. 333-336, 2002.

[45] C. Chandra, M. S. Mooreand S. K. Mitra , "An Efficient Method for the Removal of Impulse Noise From Speech and Audio Signals," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '98),* pp. 206 - 208, 1998.

[46] K. Abdullah-Al-Mamun, F. Sarker and G. Muhammad, "A High Resolution Pitch Detection Algorithm Based on AMDF and ACF," *Journal of Scientific Research*, vol. 1, no. 3, pp. 508-515, 2009.

[47] J.- W. Xu and J. C. Principe, "A Pitch Detector Based on a Generalized Correlation Function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1420-1432, Nov 2008.

[48] T. Shimamura and H. Kobayashi, "Weighted Autocorrealtion for Pitch Extraction of Noisy Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 727-730, 2001.

[49] D. W. Griffin, "Multi-Band Excitation Vocoder," Research Laboratory of Electronics, Massachusetts Institute of Technology, USA, PhD 1987.

[50] J. Rouat, Y. C. Liu and D. Morissette, "A Pitch Determination and Voiced/Unvoiced Decision Algorithm for Noisy Speech," *Elsevier- Speech Communication*, pp. 191-207, 1997.

[51] L. R. Rabiner and M. R. Sambur, "Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 4, pp. 338-342, 1977.

[52] S. A. Samad, A. Hussain and L. K. Fah, "Pitch Detection of Speech Signals Using the Cross-Correlation Technique," in *TENCON 200 Proceedings*, pp. 283-286, 2000.

[53] S. A. Zahorin and H. Hu, "A Spectral/Temporal Method for Robust Fundamental Frequency Tracking," *Acoustical Society of America*, vol. 123, no. 6, pp. 4559-4571, 2008.

[54] K. Kasi and S. A. Zahorin, "YET Another Algorithm for Pitch Tracking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I-361-I-364, 2002.

[55] K. Guangyu and G. Shize, "Improving AMDF for Pitch Period Detection," in *The Ninth International Conference on Electronics Measurement & Instruments*, pp. 4-283 - 4-286, 2009.

[56] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-22, no. 5, pp. 353-362, 1974.

[57] L. Hui, B-Q. Dai and L. Wei, "A Pitch Detection Algorithm Based on AMDF and ACF," in *IEEE Internatioanl Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, pp. I-377 -I-380, 2006.

[58] H. Maalem and F. Marir, "The Fourth Order Cumulant of Speech Signals Applied to Pitch Estimation," in *IEEE International Conference on Industrial Technology (ICIT),* pp. 1303-1306, 2004.

[59] A. Moreno and J. A. R. Fonollosa, "Pitch Determination of Noisy Speech Using Higher Order Statistics," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 133 - 136, 1992.

[60] D. H. Friedman, "Multichannel Zero-Crossing-Interval Pitch Estimation," in *IEEE International Conference Audio, Speech, and Signal Processing,* pp. 764-767, 1979.

[61] A. Lacroix and N. Hoptner, "Accurate Pitch Estimation Using Digital Filters," in *IEEE Interantional Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 319-322, 1977.

[62] H. Kameoka, N. Ono and S. Sagayama, "Speech Spectrum Modeling for Joint Estimation of Spectral Envelope and Fundamental Frequency," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1507-1516, 2010.

[63] F. J. Charpentier, "Pitch Detection Using the Short-Term Phase Spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 113-116, 1986.

[64] T. En-najjary, O. Rosec and T. Chonavel, "A New Method for Pitch Prediction from Spectral Envelope and its Application in Voice Conversion," in *8th European Conference on Speech Communication on Technology (EUROSPEECH`03)*, pp. 1753-1756, 2003.

[65] H. Ding, B. Qian, Y. Li and Z. Tang, "A Method Combining LPC-Based Cepstrum and Harmonic Product Spectrum for Pitch Detection," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pp. 537-540, 2006.

[66] C. Nadeu, J. Pascual and J. Hernando, "Pitch Determination Using the Cepstrum of the One-Sided Autocorrelation Sequence," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3677-3680, 1991.

[67] P. Martin, "Comparison of Pitch Detection by Cepstrum and Spectral Comb Analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 180-183, 1982.

[68] D. G. Childers, D. P. Skinner and R. C. Kemerait, "The Cepstrum: A Guide to Processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428 - 1443, 1977.

[69] A. M. Noll, "Cepstrum Pitch Determination," *The Journal of the Acoustical Society of America* , vol. 42, no. 2, pp. 293-309, 1966.

[70] A. M. Noll, "Clipstrum Pitch Determination," *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1585- 1591, July 1968.

[71] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring Speech Representations Using a Pitch-adaptative Time-frequency Smoothing and an Instateneous-Frequency-Based F0 Extraction: Possible role of a repetitve structure in sounds," *ELSEVIER: Speech Communication*, vol. 27, pp. 187-207, 1999.

[72] T. Abe, T. Kobayashi and S. Imai, "Robust Pitch Estimation with Harmonics Enhancement in Noisy Environments Based on Instantenous Frequency," in *Fourth International Conference on Spoken Language (ICSLP 96)*, pp. 1277-1280, 1996.

[73] T. Abe, T. Kobayashi, and S. Imai, "Harmonics Tracking and Pitch Extraction Based on Instantenous Frequency," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 756-759, 1995.

[74] H. Yang, L Qiu, and S-N. Koh, "Application of Instantenous Frequency Estimation for Fundamental Frequency Detection," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 616-619, 1994.

[75] S. Gonzalez and M. Brookes, "A Pitch Estimation Filter Robust to High Levels of

Noise (PEFAC)," in *19th European Signal Processing Conference (EUSIPCO 2011)*, pp. 451-455, 2011.

[76] H.-T. Hu, C. Yu and C.- H. Lin, "Usefulness of the Comb Filtering Output for Voiced/Unvoiced Classification and Pitch Detection," in *International Conference on Signal Processing System*, pp. 135-139, 2009.

[77] J. Lienard, F. Signol and C. Barras, "Speech Fundamental Frequency Estimation Using the Alternate Comb," in *Conference of the International Speech Communication Association, INTERSPEECH 2007*, pp. 2773-3776, 2007.

[78] J.- H. Chang, N. S. Kim and S. K. Mitra, "Pitch Estimation of Speech Signal Based on Adaptive Lattice Notch Filter," *Elsevier, Signal Processing*, vol. 85, pp. 637-641, 2005.

[79] M. Gainza, R. Lawlor and E. Coyle, "Multi Pitch Estimation by Using IIR Comb Filters," in *47th. International Symposium focused on Multimedia Systems and Applications (ELMAR), 2005*.

[80] K. Nishi and S. Ando, "An Optimal Comb Filter for Time-Varying Harmonics Extraction," in *IEICE Transactions Fundamentals*, pp. 1622-1627, 1998.

[81] J. A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-22, no. 5, pp. 330-338, 1974.

[82] M. G. Amin, "A Frequency-Domain LMS Comb Filter," *IEEE Transactions on Circuits and Systems*, vol. 38, no. 12, pp. 1573-1576, 1991.

[83] Y. Tadokoro, T. Morita and M. Yamaguchi , "Pitch Detection of Musical Sounds Noticing Minimum Output of Parallel Connected Comb Filters ," in *IEEE TENCON 2003*, pp. 380-383, 2003.

[84] Z. N. Milivojevic and M. D. Mirkovic, "Estimation of the Fundamental Frequency of the Speech Signal Modeled by the SYMPES method," *ElSEVIER, International Journal of Electronics and Communications (AEU)*, vol. 63, pp. 200-208, 2009.

[85] M. T. Nagy, G. Rozinaj and A. Palenik, "A Hybrid Pitch Period Estimation Method Based on HNM Model," in *ELMAR*, pp. 175- 178, 2007.

[86] A. Shah, R. P. Ramachandran and M. A. Lewis, "Robust Pitch Estimation Using an

Event Based Adaptive Gaussian Derivative Filter," in *IEEE International Symposium on Circuits and Systems, ISCAS2002,* pp. II-843 - II-846, 2002.

[87] R. J. McAulay, "Maximum Likelihood Pitch Estimation Using State-Variable Technique," in *IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP)'78,* pp. 12-14, 1978.

[88] Y. H. Gu, "HMM-Based Noisy-Speech Pitch Contour Estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 21-24, 1992.

[89] J. Tabrikian, S. Dubnov and Y. Dickalov, "Maximum A-Posteriori Probability Pitch Tracking in Noisy Environments Using Harmonic Model," *IEEE Transactions on Speech and Audio Proceeding,* vol. 12, no. 1, pp. 76-87, 2004.

[90] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum Likelihood Pitch Estimation ," *IEEE Transaction on Acoustics, Speech, Signal and Processing* , vol. ASSP-24, no. 5, pp. 418-423, 1976.

[91] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Transactions on Signal Processing,* vol. 39, no. 2, pp. 411-423, 1991.

[92] J. L. Flanagan, *Speech Analysis Synthesis and Perception,* Second edition ed.: Springer-Verlag, 1972.

[93] wiki. Lung volumes. [Online]. http://en.wikipedia.org/wiki/Lung_volumes.

[94] The Trachea and Bronchi. [Online]. http://education.yahoo.com/reference/gray/subjects/subject/237.

[95] The Vocal Track and Larynx. [Online]. http://www.phon.ox.ac.uk/jcoleman/phonation.htm.

[96] T. Wilde. Principle Voice Production. [Online]. http://www.ehow.com/info_8114873_principles-voice-production.html.

[97] Cari Cole. 5 Ways to Stop Shreding Your Vocal Cords. [Online]. http://www.caricole.com/blog/2012/04/10/vocal-health-stop-shredding-your-vocal-cords/.

[98] Vocal Folds. [Online]. http://en.wikipedia.org/wiki/Vocal_folds.

[99] I. R. Titze, "The Physics of Small-Amplitude Oscillation of the Vocal Folds,"

Journal Acoustical of America, vol. 83, no. 4, pp. 1536-1552, 1988.

[100] U. G. Goldstein, "An Articulatory Model for the Vocal Tracts of Growing Children," Massachusetts Institute of Technology, Cambridge, MA, Ph.D. dissertation 1980.

[101] Washington Voice Consotium. Vocal Fold Scarring. [Online]. http://www.voiceproblem.org/disorders/vfscarring/diagnosis.php#.

[102] Jonathan Harrington. Acoustic Phonetics. [Online]. http://www.phonetik.uni-muenchen.de/~jmh/research/papers/acoustics.pdf.

[103] G. Fant, J. Liljencrants, and Q. Lin, "A Four-Parameter Model of Glottal," *STL-QPSR*, vol. 26, no. 4, pp. 1-13, 1985.

[104] Paul A. Taylor, *A Phonetic Model of English Intonation, PhD Thesis*, Edinburg University, 1992.

[105] D. Reynolds, "Gaussian Mixture Model," MIT Lincoln Laboratory.

[106] Gaussian function. [Online]. http://en.wikipedia.org/wiki/Gaussian_function.

[107] Moving average. [Online]. http://en.wikipedia.org/wiki/Moving_average.

[108] J. M. Mendel, "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoritical Results and Some Applications," *Proceedings of the IEEE*, vol. 79, no. 3, pp. 278-305, 1991.

[109] C. L. Nikias and J. M. Mendel, "Signal Processing with Higher-Order Spectra," *IEEE Signal Processing Magazine*, pp. 10-37, 1993.

[110] X. Jiang, "Fundamental Frequency Estimation by Higher Order Spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00*, pp. 253-256, 2000.

[111] E. Nemer, R. Goubran and S. Mahmoud., "TheThird-Order Cumulant of Speech Signals with Application to Reliable Pitch Estimation," in *Ninth IEEE SP Workshop on Statistical Signal and Array Processing*, pp. 427-430, 1998.

[112] T. Takagi, N. Seiyama and E. Miyasaka, "A Method for Pitch Extraction of Speech Signals Using Autocorrelation Functions Through Multiple Window Lengths," Electronic Communication Japan, Part III, vol. 83, no. 2, pp. 67-79, 2000.

[113] K. Hirose, H. Fujisaki and S. Seto, "A Scheme for Pitch Extraction of Speech Using

Autocorrelation Function with Frame Length Proportional to Time Lag," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 149–152, 1992.

[114] K. Oh,and C. Un, "A Performance Comparison of Pitch Extraction Algorithms for Noisy Speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP`84*, pp. 85-88, 1984.

[115] Speech Noises. [Online]. http://www.utdallas.edu/~loizou/speech/noizeus/.

[116] http://www.festvox.org/cmu_artic.

[117] http://www.cstr.ed.ac.uk/research/projects/fda.

[118] G. F. Meyer. Keele Pitch Database. [Online]. fttp://ftp.cs.keele.ac.uk/pub/pitch.

[119] The Centre for Speech Technology Research. [Online]. http://www.cstr.ed.ac.uk/research/projects/artic.

[120] Alan W. Black. festvox. [Online]. http://www.festvox.org/dbs_kdt.html.

[121] Wolfgang Hess, and Helge Indefrey, "Accurate Time-Domain Pitch Determination of Speech Signals by Means of a Laryngograph," *Elsevier- Speech Communication*, vol. 6, pp. 55-68, 1987.

[122] Electroglottograph. [Online]. http://en.wikipedia.org/wiki/Electroglottograph.

[123] N. C. Geckinli and D. Yavuz, "Algorithm for Pitch Extraction Using Zero-Crossing Interval Sequence," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 6, pp. 559-564, 1977.

[124] C. Shahnaz, W.- P. Zhu and M. O. Ahmad, "Pitch Estimation Based on a Harmonic Sinusoidal Autocorrelation Model and a Time-Domain Matching Scheme," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 322-325, 2012.

[125] I. A. Atkinson, A. M. Kondoz, and B. G. Evans, "Pitch Detection of Speech Signals Using Segmented Autocorrelation," *Electronics Letter*, vol. 31, no. 7, pp. 533-535, 1995.

[126] M. Ghulam, "Noise Robust Pitch Detection Based on Extended AMDF," in *IEEE International Symposium on Signal processing and Information Technology, 2008 (ISSPIT)*, pp. 133-138, 2008.

[127] X. Gang and T. Liang-Rui, "Speech Pitch Period Estimation Using Circular AMDF," *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, (PIMRC 2003)*, pp. 2452-2455, 2003.

[128] Y.- M. Zeng, Z.-Y. Wu, H.- B. Liu and L. Zhou, "Modified AMDF Pitch Detection Algorithm," in *International Conference on Machine Learning and Cybernetics*, pp. 470-473, 2003.

[129] W. Zhang, G. Xu and Y. Wang, "Pitch Estimation Based on Circulation AMDF," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I-341 - I-344, 2002.

[130] D. Tuffelli, "A Pitch Detection Algorithm with Hypothesis and Test Strategy by Means of Fast Surface AMDF," in *IEEE Internatioanl Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 81-84, 1984.

[131] E. Ambikairajah, M. J. Carey and G. Tattersall, "Estimating the Pitch Period of Voiced Speech," *Electronics Letters*, vol. 16, no. 12, pp. 464-466, 1980.

[132] C. K. Un and S-C. Yang, "A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 6, pp. 565-572, 1977.

[133] G. S. Ying, L. H. Jamieson and C. D. Michell, "A Probability Approach to AMDF Pitch Detection," *Forth International Conference on Spoken Language Proceedings, ICSLP 96*, pp. 1201-1204, 1996.

[134] C. Shahnaz, W.- P. Zhu and M. O. Ahmad, "A Robust Pitch Estimation Algorithm in Noise," in *IEEE Internatioanl Conference on Acoustics, Speech, and Signals Processing,(ICASSP)*, pp. IV-1073 - IV-1076, 2007.

[135] A. Pawi, S. Vaseghi, B. Milner and S. Ghorshi, "Fundamental Frequency Estimation Using Modified Higher Order Moments and Multiple Windows," in *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011*, pp. 1965-1968, 2011.

[136] A. Pawi, S. Vaseghi and B. Milner, "Pitch Extraction Using Modified Higher Order Moments," in *IEEE International Conference on Acoustics, Speech,and Signal Processing (ICASSP)*, pp. 5078 - 5081, 2010.

[137] L. Baghai-Ravary , G. Kochanski and J. Coleman, "Precision of Phoneme

Boundaries Derived Using Hidden Markov Models," in *10th Annual Conference of the International Speech Communication Assosciation, INTERSPEECH 2009*, pp. 2879-2882, 2009.

[138] J. Laroche, Y. Stylianou and E. Moulines, "HNM: A Simple, Efficient Harmonic + Noise Model for Speech," IEEE workshop on Applications of Signal Processing to Audio and Acoustics, pp. 169_172, 1993.

[139] W. Zhang, H.- S. Kim and W. H. Holmes, "Investigation of the Spectral Envelope Estimation Vocoder and Improved Pitch Estimation Based on the Sinusoidal Speech Model," in *International Conference on Information, Communications and Signal Processing, ICICS'97*, pp. 513-516, 1997.

[140] E. Godoy, O. Rosec and T. Chonavel, "Speech Spectral Envelope Estimation Through Explicit Control of Peak Evolution in Time," in *10th International Conference on Information Sciences Signal Processing and Their Applications (ISSPA)*, pp. 209-212, 2010.

[141] R. Vich and M. Vondra, "Speech Spectrum Envelope Modeling," *Springer-Verlag Berlin Heidelberg*, pp. 129-137, 2007.

[142] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion Measure for Speech Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 367-376, 1980.

[143] Full width at half maximum. [Online]. http://en.wikipedia.org/wiki/Full_width_at_half_maximum.

[144] D. B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 4, pp. 786- 794, 1981.

[145] J. Bartosek, "A Pitch Detection Algorithm for Continuous Speech Signals Using Viterbi Traceback with Temporal Forgetting," *Journal of Advance Engineering, ACTA Polytechnica*, vol. 51, no. 5, pp. 8-13, 2011.

[146] D. Gerdhard, "Pitch Extraction and Fundamental Frequency: History and Current Techniques," Department of Computer Science, University of Regina, Canada., Technical Report TR-CS 2003-06 ISSN 0828-3494; ISBN 0 7731 0455 0, 2003.

[147] R. Kumaresan and A. Rao, "Model-Based Approach to Envelope and Positive Instantenaous Frequency Estimation of Signals with Speech Applications," *Journal*

Acoustical Society America, vol. 105, no. 3, pp. 1912-1924, 1999.

[148] T. Akgul and A. El-Jaroudi, "Discrete All-Pole Modeling Using Higher-Order Spectra," in *IEEE International Conference on Audio, Speech and Signal Processing, ICASSP 1991,* pp. 3493-3496, 1991.

[149] F. J. Harris, "On the Use of Window for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE,* vol. 66, no. 1, pp. 51-83, 1978.

[150] Edward P. Neuburg, "On Estimating Rate of Change of Pitch," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 355-357, 1988.

[151] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusiodal Speech Model," in *IEEE International Conference on Acoustic, Speech, and Signal Processing,ICASSP'90.*, pp. 249-252, 1990.

[152] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Transaction on Speech and Audio Processing*, vol. 9, no. 1, pp. 21-29, 2001.

[153] T. Nakatani and T. Irino, "Robust and Accurate Fundamental Frequency Estimation Based on Dominant Harmonic Components," *Journal Acoustical Society of America*, vol. 116, no. 6, pp. 3690-3700, 2004.

APPENDIX A

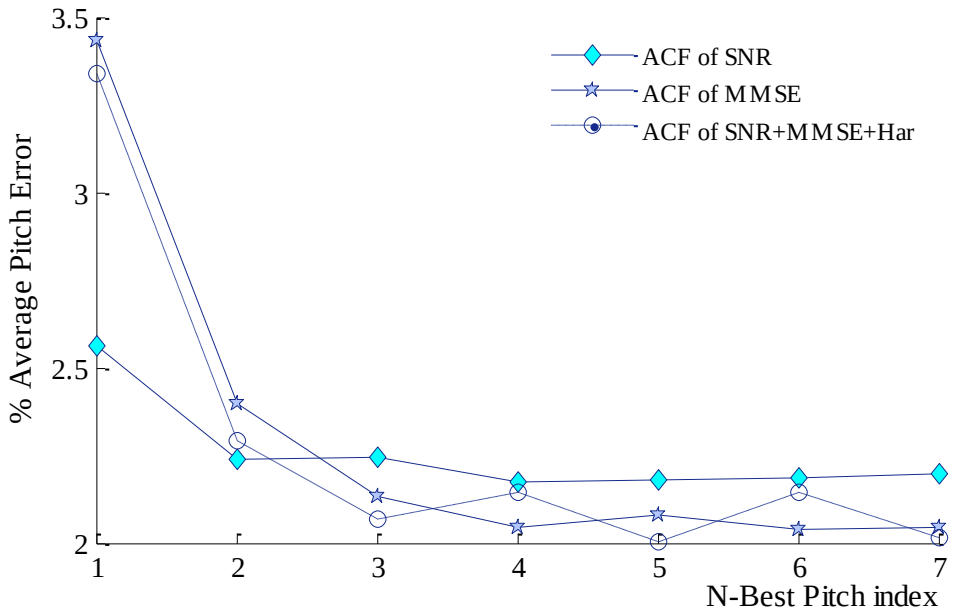


Figure A.1 – Comparison % overall pitch error of three distortion measures: ACF with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index.

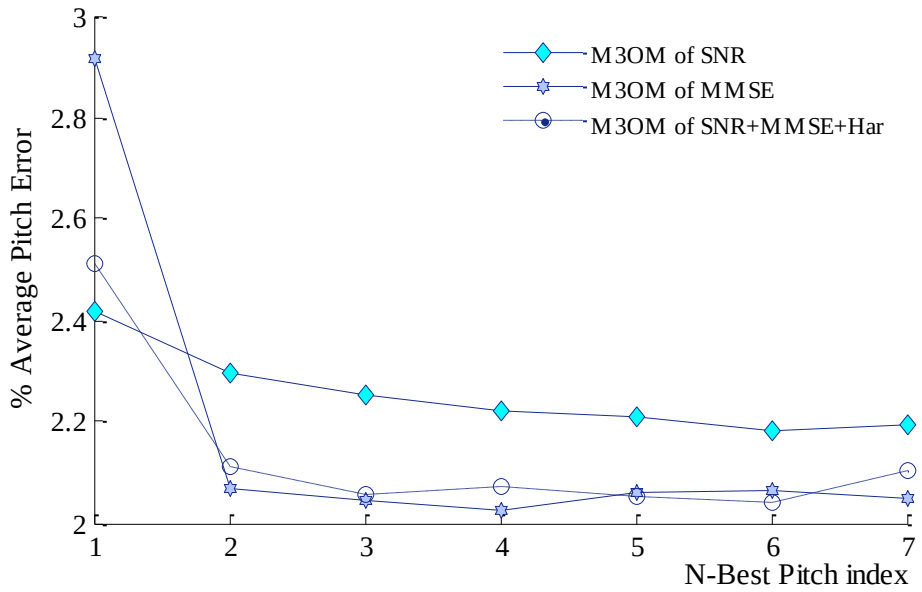


Figure A.2 - Comparison % overall pitch error of three distortion measures: Modified 3rd order moment with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index

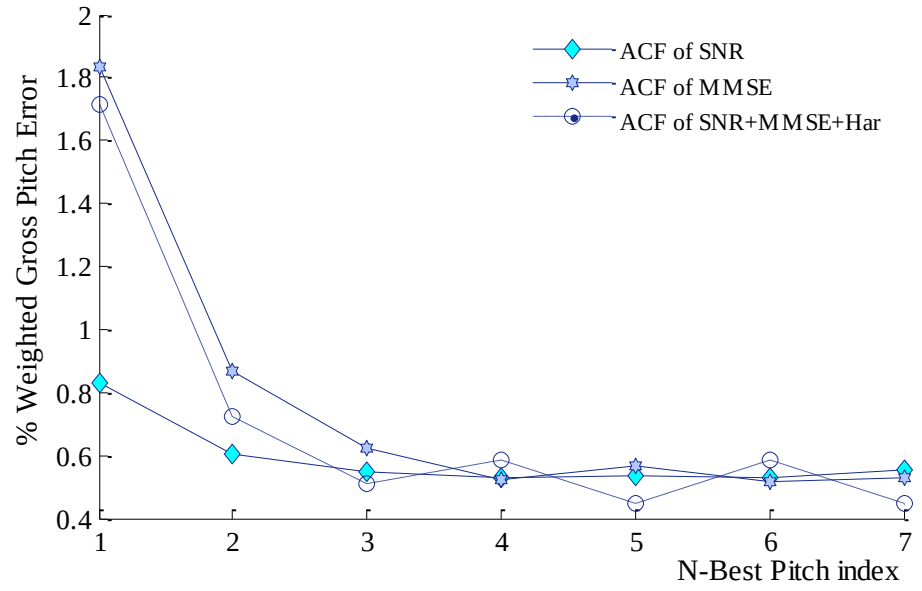


Figure A.3 - Comparison % weighted gross pitch error of three distortion measures: ACF with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index.

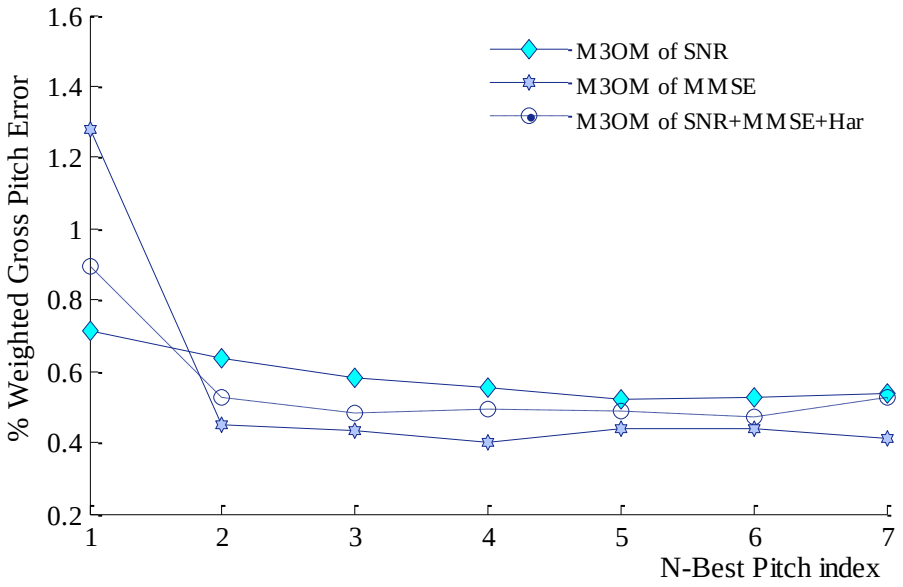


Figure A.4 - Comparison % weighted gross pitch error of three distortion measures: Modified 3rd order moment with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index

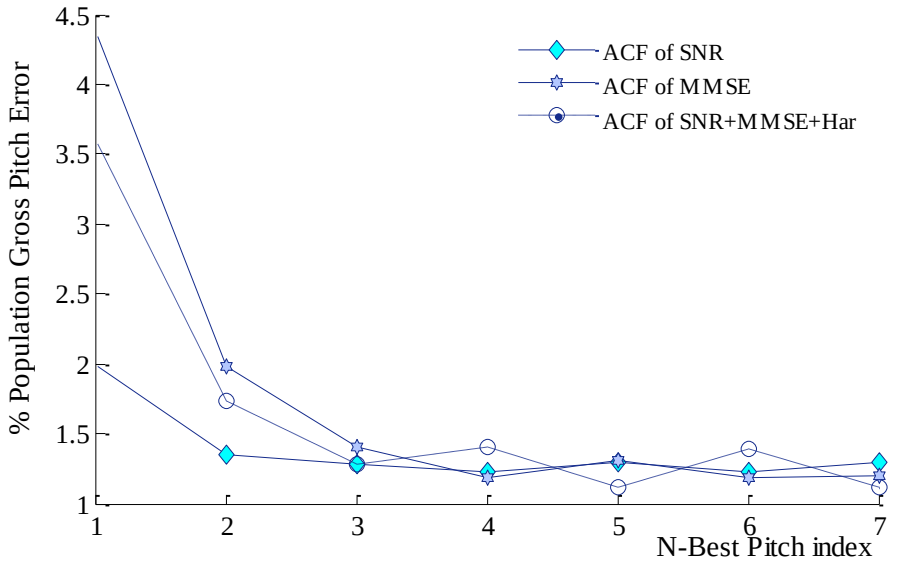


Figure A.5 - Comparison % population gross pitch error of three distortion measures: ACF with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index

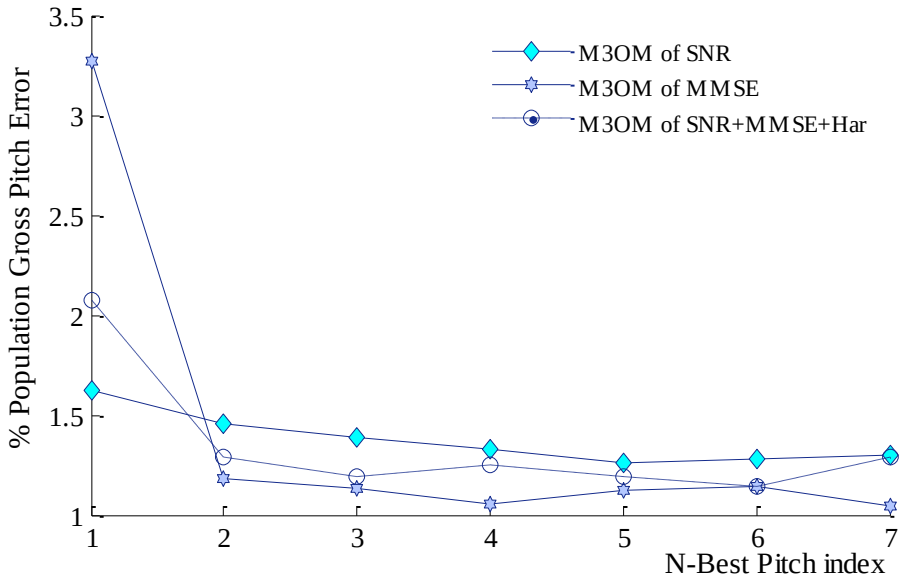


Figure A.6 - Comparison % population gross pitch error of three distortion measures: ACF with weighted SNR, weighted MMSE, and combination (i.e. SNR + MMSE + Harmonicity) as a function of *N*-Best index