Prosody Modifications for Voice Conversion

Jitendra Kumar Dhiman

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

Department of Electrical Engineering

July 2013

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

Jitendra

(Signature)

(Jitendra Kumar Dhiman)

EE11M04

(Roll No.)

Approval Sheet

This Thesis entitled Prosody Modifications for Voice Conversion by Jitendra Kumar Dhiman is approved for the degree of Master of Technology from IIT Hyderabad

(Dr. C. S. Sastry) Examiner
Department of Mathematics
IITH

(Dr. Sumohana Channappayya) Examiner
Department of Electrical Engineering
IITH

(Dr. K. Sri Rama Murty) Adviser
Department of Electrical Engineering
IITH

(Dr. Siva Rama Krishna Vanjari) Chairman
Department of Electrical Engineering
IITH

Acknowledgements

.

Dedication

.

Abstract

Generally defined, speech modification is the process of changing certain perceptual properties of speech while leaving other properties unchanged. Among the many types of speech information that may be altered are rate of articulation, pitch and formant characteristics.Modifying the speech parameters like pitch, duration and strength of excitation by desired factor is termed as prosody modification. In this thesis prosody modifications for voice conversion framework are presented.Among all the speech modifications for prosody two things are important firstly modification of duartion and pauses (Time scale modification) in a speech utterance and secondly modification of the pitch(pitch scale modification).Prosody modification involves changing the pitch and duration of speech without affecting the message and naturalness.In this work time scale and pitch scale modifications of speech are discussed using two methods Time Domain Pitch Synchronous Overlapped-Add (TD-PSOLA) and epoch based approach.In order to apply desired speech modifications TD-PSOLA discussed in this thesis works directly on speech in time domian although there are many variations of TD-PSOLA.The epoch based approach involves modifications of LP-residual.

Among the various perceptual properties of speech pitch contour plays a key role which defines the intonation patterns of speaker.Prosody modifications of speech in voice conversion framework involve modification of source pitch contour as per the pitch contour of target.In a voice conversion framework it requires prediction of target pitch contour. Mean/ variance method for pitch contour prediction is explored.

Sinusoidal modeling has been successfully applied to a broad range of speech processing problems. It offers advantages over linear predictive modeling and the short-time Fourier transform for speech analysis/ synthesis and modification. The parameter estimation of sinusoidal modeling which permits flexible time and frequency scale voice modifications is presented. Speech synthesis using three models sinusoidal, harmonic and harmonic-plus-residual is discussed.

Contents

Chapter 1

Introduction

1.1 Motivation

Voice conversion (VC) is the process of modifying a source speaker's speech to make it sound like that of a different target speaker. Due to its wide range of applications, there has been a considerable amount of research effort directed at this problem in the last few years. As an end in itself, it has use in many anonymity and entertainment applications. For example, voice conversion can be used in the film dubbing industry and/ or automatic translation systems to maintain the identity of the original speakers when translating from the original language to another. It can also be applied to transform an ordinary voice singing karaoke into a famous singer's voice. Another application could be to mask the identity of a speaker who wants to remain anonymous on the telephone. Computer aided language learning systems can also benefit from voice conversion, by using converted utterances as feedback for the learner. Another important application is the customization of text-to-speech (TTS) systems. Typically, unit selection and concatenation TTS synthesis requires the recording of large speech corpora by professional speakers. Because of the high cost involved in recording a separate database for each new speaker, commercial text-to-speech implementations only generate speech by a few speakers. Voice conversion can be exploited to economically synthesize new voices from previously recorded and already available databases. In dialog systems, voice conversion technology can be used to adapt speech outputs to different situations and make man-machine interactions more natural. Systems can mimic human-human communication using emotion conversion techniques to transmit extra information to the user on how the dialog is going by generating confident, doubtful, neutral, happy, sad or angry utterances for example. They can also modify the focus to indicate more precisely the informational item in question. Finally, as an attempt to separate speaker identity and message, voice conversion can be very useful in core technologies such as speech coding, synthesis and recognition to achieve very low bandwidth speech coding, lead to more accurate speech models and improve the performance of state-of-the-art speech recognizers.

1.2 Thesis objective

The objective of this work is to study techniques for time and pitch scale speech modifications.

The objective of time scale modification is to alter the speaking rate without changing the spectral content of the original speech. Considering the source-filter model of speech, this means that the time evolution of the excitation signal and the vocal tract filter needs to be time scaled.

The objective of pitch scale modification is to alter the fundamental frequency without affecting the spectral envelope or the formant structure of the signal.

In general time scale and pitch scale speech modifications are not independent of each other.Simple attempt for time scale modification such as playing speech sentence at faster or slower rate interacts with pitch.Pitch increases if a speech utterance is played at faster rate than the original sampling rate of utterance.Pitch decreases if speech utterance is played at slower rate than the original sampling rate of the utterance. In this process naturalness of the speech is not preserved.This work has been directed to apply prosody modifications(particularly time scale and pitch scale) while at the same time preserving naturalness of the speech.

1.3 Outline

The remainder of the thesis is organized as follows:

- Chapter 2: Time Domain Pitch Synchronous Overlap -Add (TD-PSOLA) – It Presents simple algorithm for pitch scale and time scale modification of speech. The algorithm details and results for pitch and time scale modification are compared. Finally few drawbacks of the algorithm are listed.
- Chapter 3: Prosody modifications of speech using epoch based approach – This chapter presents mapping of pitch contour in voice conversion framework. A technique using epoch based approach for LP-residual modification is discussed.
- Chapter 4: Sinusoidal analysis/synthesis – It presents the basics of model based approach for speech modifications with analysis and synthesis of speech using sinusoidal modeling.
- Chapter 5: This concludes with a summary of work presented in this dissertation.

Chapter 2

Time Domain Pitch Synchronous Overlap -Add (TD-PSOLA)

2.1 Introduction

Prosody modifications of speech require pitch scale and time scale modifications independently. The usual technique such as playing the original signal for pitch or time scale modification at faster or slower rate does not work because by this method the pitch and duration can not be modified independent of each other. There is a need for a technique that can be used to modify pitch and duration independently.To achieve this TD-PSOLA method can be used and is discussed in this chapter.

2.2 Definitions

2.2.1 Time scale modification

The goal of time scale modification is to change the apparent rate of articulation without affecting the perceptual quality of the original speech. This requires the pitch contour to be stretched or compressed in time and formant structure to be changed at slower or faster rate than the rate of the input speech but otherwise not modified (contour shape is preserved).

2.2.2 Pitch scale modification

The goal of pitch modification is to alter the fundamental frequency in order to compress or expand the spacing between the harmonic components in the spectrum preserving the short time envelope (the locations and bandwidths of the formants) as well as the time evolution. In contrast to time-scale modifications, in this case the pitch contour is modified without modifying time resolution of the pitch contour.

2.3 TD–PSOLA

This algorithm was proposed by F. Charpentier and E. Moulines [3]. TD–PSOLA is a digital signal processing technique used for speech processing and more specifically speech synthesis. It can be used to modify the pitch and duration of a speech signal. TD–PSOLA works by dividing the speech waveform in small overlapping segments. To change the pitch of the signal, the segments are moved further apart (to decrease the pitch) or closer together (to increase the pitch). To change the duration of the signal, the segments are then repeated multiple times (to increase the duration) or some are eliminated (to decrease the duration). The segments are then combined using the overlap add technique.The idea of the method for time scale modification is shown in the Fig.2.1 below. Figure 2.1 shows that the length of the original speech signal can be doubled by repeating each frame twice without changing the pitch of original speech signal.

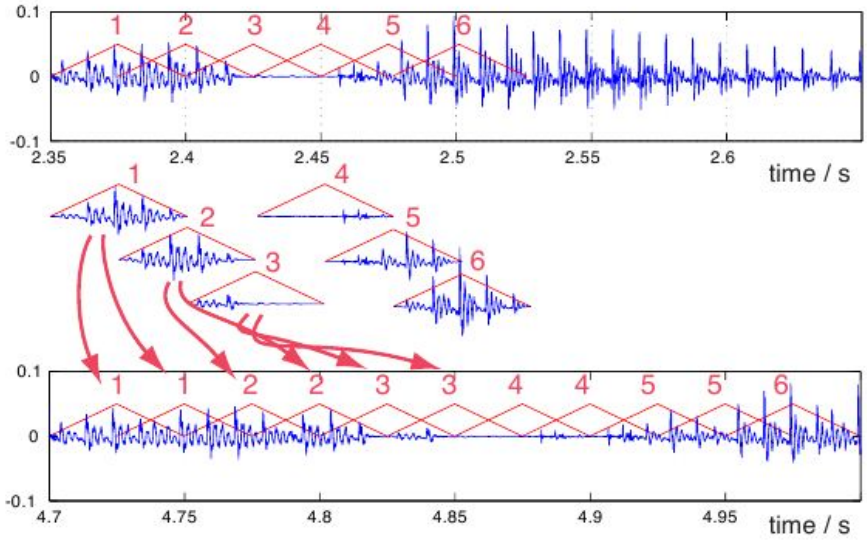

Figure 2.1: TD–PSOLA concept

The prerequisite for PSOLA is the analysis time instants.The speech signal is divided into overlapping frames by placing analysis window at these instants. These analysis time instants are obtained at pitch synchronous rate for voiced parts of speech and for unvoiced parts the locations of analysis time instants are random or can be set at constant rate till the next voiced segment begins. In PSOLA framework for voiced part of speech the analysis time instants are also known as pitch marks. In this project the pitch marks are the glottal closure instants (GCI) which are obtained by using group delay analysis [1]. The determination of GCIs is briefly explained in sec. 2.5. In this case PSOLA can exploit knowledge of the pitch to correctly synchronize the time segments, avoiding pitch discontinuities. For the speech analysis in this project Hamming window is used.

2.4 Algorithm(How does TD-PSOLA work)

2.4.1 Time scale modification

When we perform time stretching of an input sound, the time variation of the pitch period $p(t)$ should be stretched accordingly. If $\tilde{t} = \alpha t$ describes the time-scaling function or time-warping function that maps the time t of the input signal into the time of the output signal, the local pitch period of the output signal $\tilde{p}(\tilde{t})$ will be defined by $\tilde{p}(\tilde{t}) = \tilde{p}(\alpha t) = p(t)$. More generally, when the scaling factor is not constant, a nonlinear time–scaling function can be defined as $\tilde{t} = T(t) = \int_0^t \alpha(\tau)d\tau$ and used instead of $\tilde{t} = \alpha t$. The algorithm is composed of two phases: the first phase analyzes and segments the input sound and the second phase synthesizes a time-stretched version by overlapping and adding time segments extracted by the analysis algorithm.

- Analysis algorithm:
  1. Determination of the pitch period $p(t)$ of the input signal and of time instants (pitch marks) t_i. These pitch marks are in consistent with the glottal closures instants at a pitch-synchronous rate during the periodic part of the sound and at a random rate during the unvoiced portions. In practice $p(t)$ is considered constant $p(t) = p(t_i) = t_{i+1} - t_i$ on the time interval (t_i, t_{i+1}) .
  2. Extraction of a segment centered at every pitch mark t_i by using a Hamming window with length $L_i = 2p(t)$ (two pitch periods). In each iteration window centre is placed at the current analysis instant t_i covering one pitch period left and one pitch period right of the current analysis instant t_i.

- Synthesis algorithm: For every synthesis pitch mark $\tilde{t_k}$ –
  1. Choice of the corresponding analysis segment (identified by the time mark) is obtained by minimizing the time distance $|\alpha t_i - \tilde{t}_k|$.
  2. Overlap and add the selected segment. Notice that some input segments will be repeated for $\alpha > 1$ (time expansion) or discarded when $\alpha < 1$ (time compression).
  3. Determination of the time instant $\tilde{t}_{k+1}$ where the next synthesis segment will be centered, in order to preserve the local pitch, by the relation $\tilde{t}_{k+1} = \tilde{t}_k + \tilde{p}(\tilde{t}_k) = \tilde{t}_k + p(t_i)$.

2.4.2 Pitch scale modification

The algorithm for pitch scale modification is same as time scale modification except that the equation for deriving synthesis time instants is modified as follows $\tilde{t}_{k+1} = \tilde{t}_k + \tilde{p}(\tilde{t}_k) = \tilde{t}_k + p(t_i)/\beta$ where β is pitch modification factor.

2.5 Determination of analysis time instants for TD-PSOLA

The determination of the pitch and the position of pitch marks (analysis time instants) is not a trivial problem and could be difficult to implement robustly. The sound quality of the modification results

of the PSOLA algorithm essentially depends on the positioning of the pitch marks, since the pitch marks provide the centers of the segmentation windows of the PSOLA. A method which determines the pitch marks of a complete voiced frame is proposed in [1]. As mentioned above, in this project the GCIs are taken as analysis time instants (pitch marks). In this section the method of extracting the GCIs from the LP residual is briefly discussed. The group-delay analysis is used to derive the instants of significant excitation from the LP residual [1]. The analysis involves computation of the average slope of the unwrapped phase spectrum (i.e., average group-delay) for each frame. If $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the windowed signal $x(n)$ and $nx(n)$, respectively, then the group-delay function $\tau(\omega)$ is given by the derivative of the phase function $\theta(\omega)$ of $X(\omega)$ and is given by

$$\tau(\omega) = \theta'(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + Y_R^2} \tag{2.1}$$

where $X_R + jX_I = X(\omega)$, and $Y_R + jY_I = Y(\omega)$.

The justification and proof for the above formula is as follows: The group delay is defined as the negative of the derivative of unwrapped phase of the signal. In MATLAB, group delay is calculated using Fourier transform rather than differentiation. Note that all the Fourier transforms are implemented using the discrete Fourier transform. In amplitude phase representation $X(\omega)$ can be represented as

$$X(\omega) = A(\omega)e^{j\theta(\omega)} \tag{2.2}$$

where $A(\omega)$ is the absolute amplitude and $\theta(\omega)$ is the phase of $X(\omega)$.

Taking the differentiation,

$$X^{'}(\omega) = A^{'}e^{j\theta(\omega)} + A(\omega)e^{j\theta(\omega)}(j\theta^{'}(\omega))\frac{X^{'}(\omega)}{X(\omega)} = \frac{A^{'}(\omega)}{A(\omega)} + j\theta^{'}(\omega) \tag{2.3}$$

The group delay $\tau(\omega)$ is the negative of the imaginary part in this equation, i. e.,

$$\tau(\omega) = -\theta^{'}(\omega) = -\mathcal{IM}\left[\frac{X^{'}(\omega)}{X(\omega)}\right] \tag{2.4}$$

Now using the Fourier transform property $-jnx(n) \overset{F.T.}{\leftrightarrow} X^{'}(\omega)$ above equation can be written as

$$\tau(\omega) = -IM\left[\frac{F.T.(-jnx(n))}{F.T.(x(n))}\right] = \mathrm{Re}\left[\frac{F.T.(nx(n))}{F.T.(x(n))}\right] \tag{2.5}$$

We have $nx(n) \overset{F.T}{\longleftrightarrow} Y(\omega) = Y_R + jY_I$ and $x(n) \overset{F.T.}{\longleftrightarrow} X(\omega) = Y_R + jY_I$ Then

$$\tau(\omega) = \mathrm{Re}\left[\frac{Y_R + jY_I}{X_R + jX_I}\right] = \left[\frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2}\right] \tag{2.6}$$

Any isolated sharp peaks in $\tau(\omega)$ are removed by using a 3-point median filtering. The average value $\overline{\tau}$ of the smoothed $\tau(\omega)$ is the value of the phase slope function for the time instant corresponding to the center of the windowed signal $x(n)$.The phase slope function is computed by shifting the analysis window by one sample at a time. The instants of positive zero-crossings of the phase slope function correspond to the instants of significant excitation. Figure 2.2 illustrates the results of extraction of the GCIs for voiced segment.

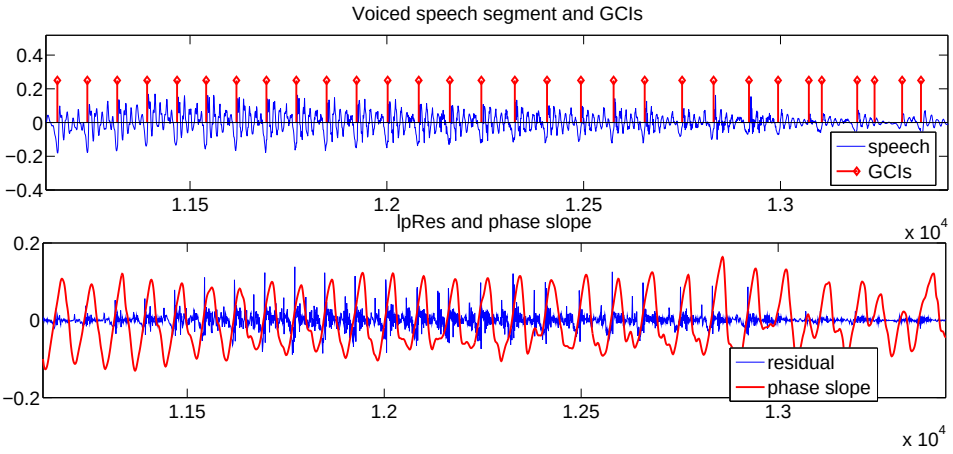

Figure 2.2: Extraction of the GCIs for voiced segment

For generating the fig. 2.2, a 10^{th}–order Linear Prediction (LP) analysis is performed using a frame size of 20 ms and a frame shift of 5 ms. Throughout this study a signal sampled at 8 kHz is used. The signal in the analysis frame is multiplied with a Hamming window to generate a windowed signal. It can be noticed from the fig. 2.2 that the GCIs are in consistent with the positive zero crossing of phase slope function.

2.6 Results and conclusions

Results for time scale modification factor using TD-PSOLA:

- Results are shown for the following cases

  1. Time scale modification factor 2
  2. Time scale modification factor 0.5
  3. Pitch scale modification factor 2
  4. Pitch scale modification factor 0.5

- **Time scale modification for modification factor 2**: Applying TD-PSOLA , in this case, stretches the original speech signal by twice of its length without modifying the pitch. Table 2.1 shows how do the analysis time instants are mapped into synthesis time instants for this case.

  It is noticed from the above Table 2.1 that the frames are repeated for time stretching. Corresponding to each synthesis time mark in the second row the frames located at analysis time mark in the first row are selected and overlap added to synthesize time stretched output speech. Figure 2.3 shows the original speech utterance and time stretched speech utterance.

Table 2.1: Time scale modification for modification factor 2

Analysis Time Instants	1	2	2	3	3	4	4	5	5	5	6	6	7	7	8	9	9	10	10	11
Synhesis Time Instants	220	296	344	392	445	498	537	576	647	718	789	886	983	1093	1203	1261	1308	1355	1413	1471

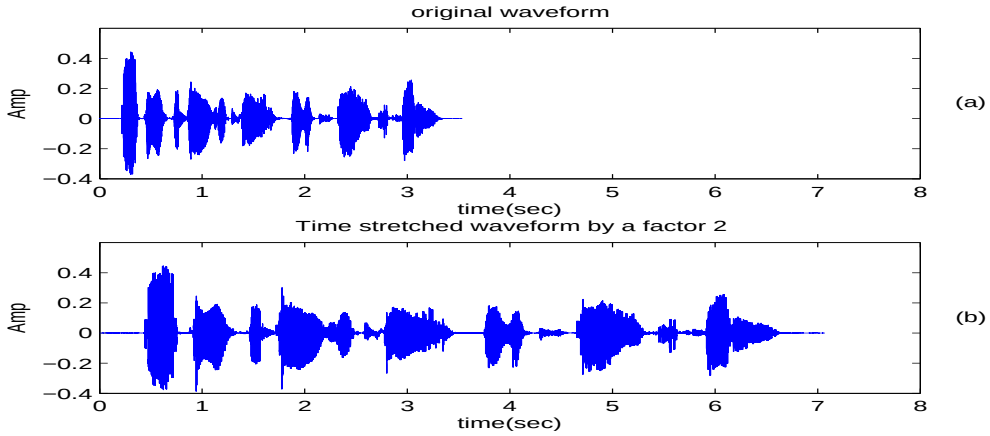

Figure 2.3: (a)Original waveform(b)Time stretchhed waveform by a factor 2

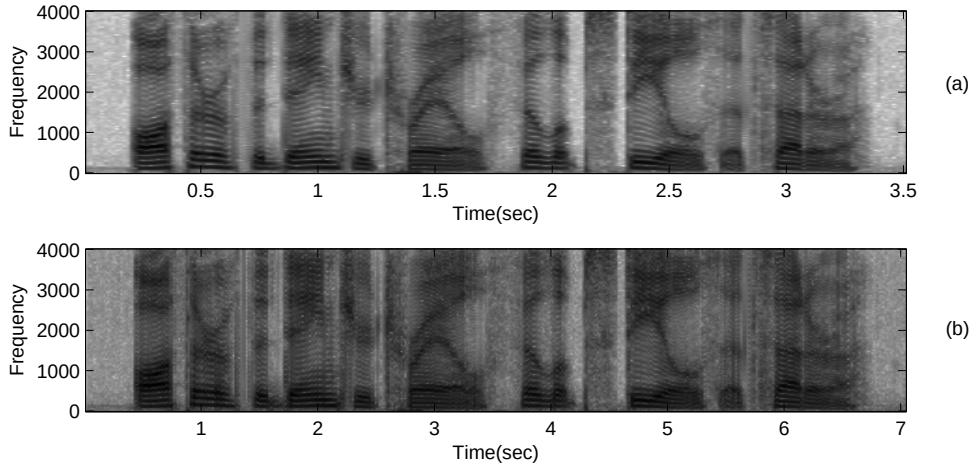

Figure 2.4: (a)Spectrogram for original utterance(b)Spectrogram for time stretched utterance

Comparison of spectrograms shows that formants evolution on time scale is stretched but otherwise not modified.

- **Time Scale Modification Factor** 0.5:

  Applying TD-PSOLA , in this case, compresses the original speech signal by half of its length without modifying the pitch. Table 2.2 shows how do the analysis time instants are mapped into synthesis time instants for this case.

  It is noticed from the table2.2 that the some of the frames are discarded for time compression.

Table 2.2: Time Scale Modification Factor **0.5**

Analysis Time Instants	1	4	5	7	10	12	14	16	17	19	21	23	24	26	29	31	33	35	37	39
Synhesis Time Instants	55	131	170	241	351	409	455	498	556	640	711	762	813	874	953	1013	1076	1141	1202	1264

Figure 2.5 shows the original speech utterance and time compressed speech utterance.

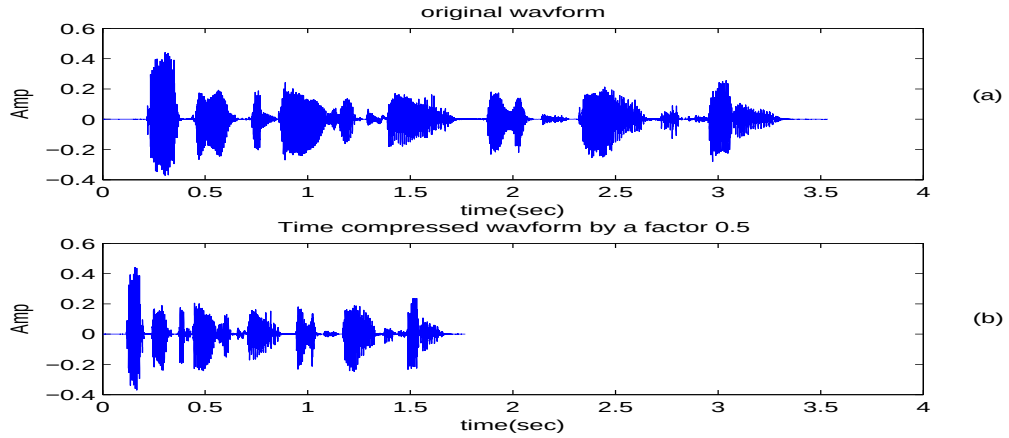

Figure 2.5: (a)Original waveform(b)Time compressed waveform by a factor 0.5

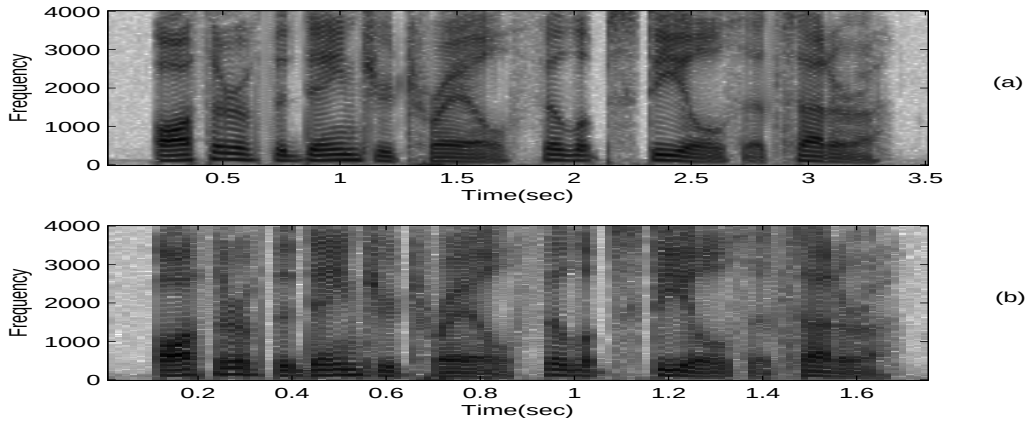

Figure 2.6: (a)Original spectrogram(b)Time compressed spectrogram

Comparison of spectrograms shows that formants evolution on time scale is compressed but otherwise not modified.

- **Pitch scale modification factor 2:** Applying TD-PSOLA, in this case, increases the pitch to double of its original pitch. Figure 2.7 shows original utterance and pitch modified utterance without changing the duration. It is more evident by looking at the narrow band spectrograms in the figures below.

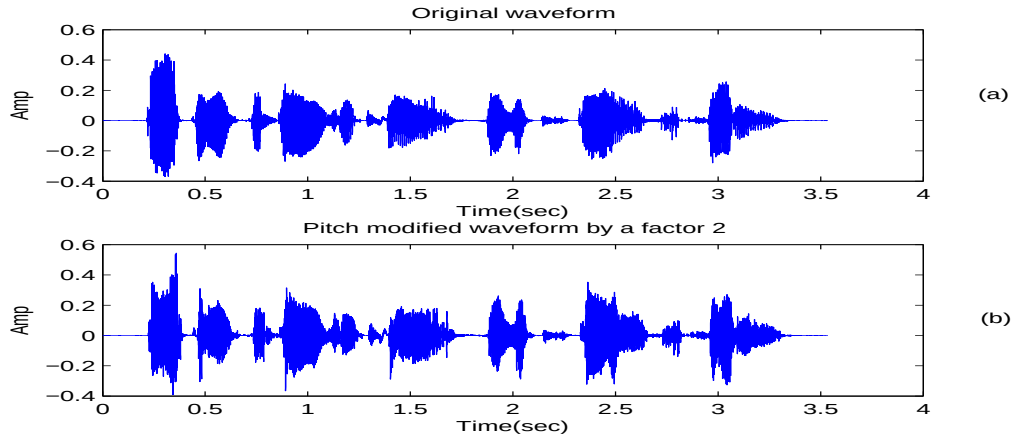

Figure 2.7: (a)Original(b)Pitch modified waveform by factor 2

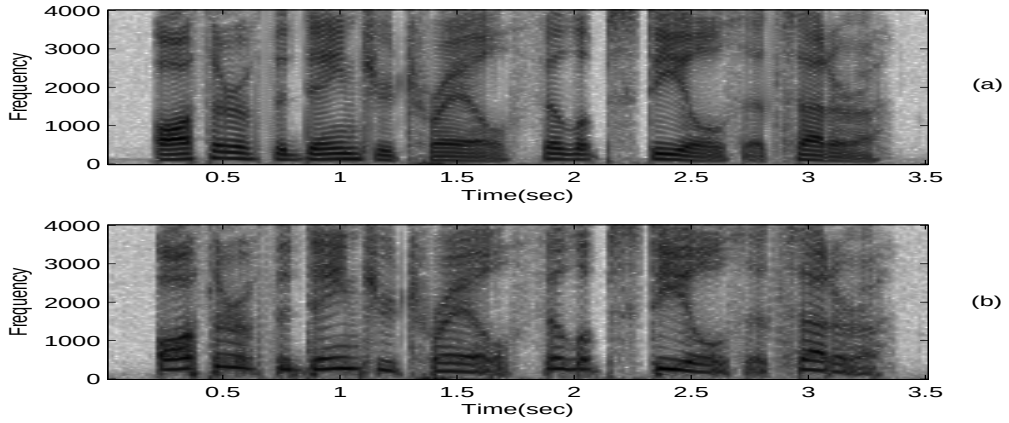

Figure 2.8: (a)Narrow band spectrogram of original utterance(b)Narrow band spectrogram for pitch modified utterance by a factor 2

Spectrograms comparison shows that the gap between horizontal striations increases as the pitch increases.

- **Pitch scale modification factor 0.5:** Applying TD-PSOLA, in this case, decreases the pitch to half of its original pitch.Figure 2.9 shows the original and pitch modified waveforms.

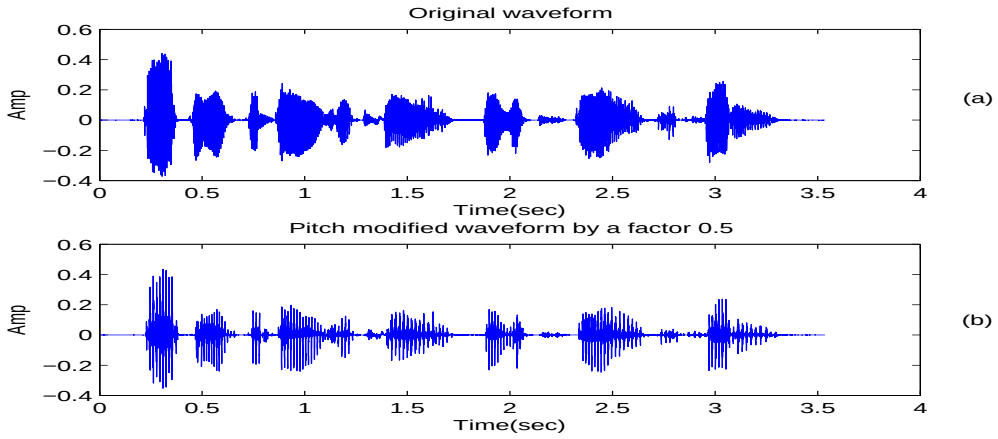

Figure 2.9: (a)Original waveform(b)Pitch modified waveform by a factor 0.5

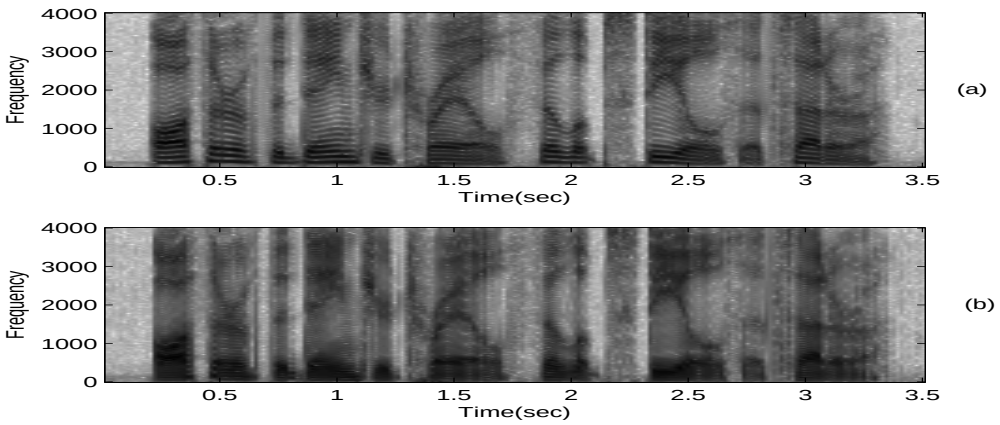

Figure 2.10: (a)Narrow band spectrogram of original utterance(b)Narrow band spectrogram for pitch modified utterance by a factor 0.5

Spectrograms comparison shows that the gap between horizontal striations decreases as the pitch decreases.

- **Conclusions:**

  1. The range of modification factors is typically from 0.5 to 2. Modification factors out of this range cause degradation in synthesized speech quality.
  2. Performance for time scale modifications is better than the pitch scale modification in terms of naturalness of synthesized speech. Pitch modified speech sounds robotic.

3. Although TD-PSOLA has several drawbacks but it is simple and computationally fast for implementation because TD-PSOLA directly operates on speech in time domain.

2.7 Drawbacks of TD-PSOLA

1. It modifies the pitch and duration only by constant factors. In general time varying speech modifications are required.
2. The accuracy of TD-PSOLA depends on the locations of pitch marks and the window chosen for segmentation.
3. If unvoiced parts repeat a number of times (due to large modification factors) then synthesized utterance sounds with audible buzziness.
4. Proper concatenation of the frames at frame boundaries is required otherwise pitch distortion and phase mismatch at frame boundaries occur which further degrade the speech quality.

Chapter 3

Prosody modifications of speech using epoch based approach

3.1 Introduction

Voice conversion algorithms aim to modify the utterance of a source speaker to sound as if it was uttered by a target speaker. Since speaker identity consists weave of factors including short-term spectral characteristics, prosody and linguistic style, it is important to transform each such factor as successfully as possible. There have been a substantial amount of work in the literature that focuses on the conversion of spectral parameters which are related to the timbre, i.e. how the voice itself sounds [7]. Pitch is arguably the most expressive manifestation of speaker-dependent prosody, which also includes factors such as phone duration, loudness and pause locations [8]. For instance, different speakers have different pitch ranges (e.g. women's mostly higher than men's), which can be represented by calculating the mean pitch and pitch variance for each speaker. In fact, the simplest and most widely used way of converting one speaker's pitch contour to another is to modify pitch values on a frame by frame basis using a linear transformation based on the mean and variance of each of the speakers [9]. The goal of this chapter is to explore pitch and duration modification of speech which falls under the more general area of prosody transformation. Voice conversion framework requires prediction of target pitch contour in order to impose the prosody characteristics of target on the source utterance. In this chapter mean-variance method for target pitch contour prediction is explored. To impose the predicted contour on source utterance LP-residual modification using epoch (instants of significant excitation) based approach is discussed which allows to synthesize speech with the predicted prosody characteristics.

3.2 Background

3.2.1 What is a Pitch Contour?

A pitch contour refers to the rise and fall of the fundamental frequency, f_0 , over time. Since f_0 is only defined for voiced sections of a speech signal, the pitch contour has positive pitch values for voiced intonation groups separated by gaps for unvoiced regions of the speech signal. The unvoiced

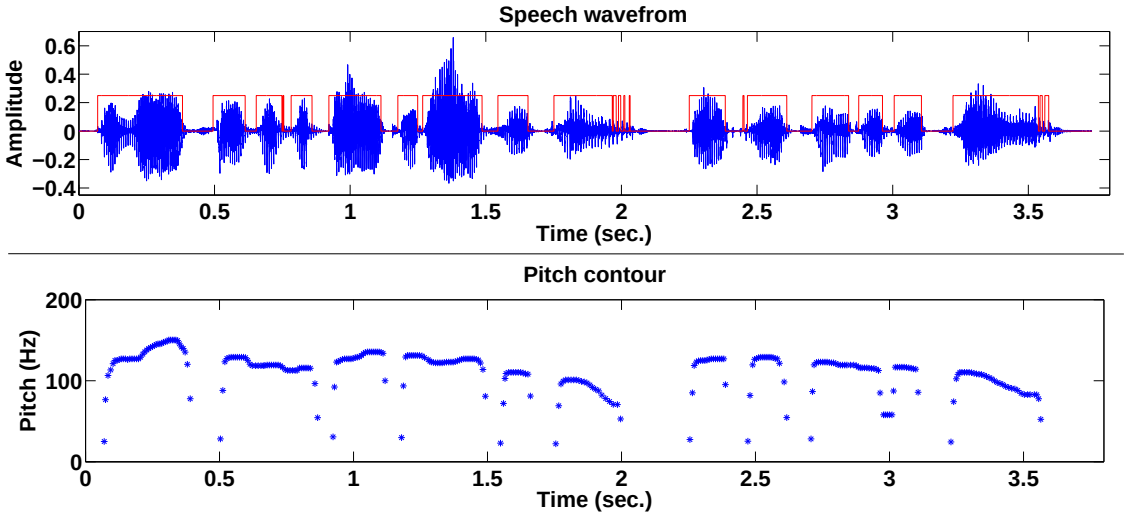

Figure 3.1: Example of pitch contour

regions corresponds to the condition where the vocal chords are not vibrating, such as the production of certain plosives or fricatives. Pitch contours reflect expressions of emotion as well as linguistic features such as the sentence type. For instance, most declaratives have an overall declining pitch contour. On the other hand, yes-no questions usually have an extreme final upturn in the last intonational phrase which sometimes causes the overall contour to incline. While such patterns are more or less speaker-independent, there may be intonational trends specific to a speaker or a group of speakers. Figure 3.1 shows example of pitch contour with the speech waveform label with voiced section.

3.2.2 Pitch extraction

In order to get pitch contour for a given speech utterance the utterance is processed for GCI detection. GCIs are taken as anchor points for pitch determination. The difference between the next GCI location to current GCI location over a voiced section of speech yields the pitch value associated to the current GCI location. For example take two consecutive GCI locations at t_k and t_{k+1} (in samples). The pitch vale at instant t_k is given by $f_0 = \frac{t_{k+1}-t_k}{f_s}$ in Hz; where f_s is the sampling frequency of speech waveform. Pitch values are determined for each GCI and plotted against corresponding GCI. A 10 point Smoothing filter on the pitch values is applied to make the pitch contour smooth. Chapter 2 describes how the GCIs are derived using group delay analysis.

3.2.3 Speech corpus

A speech corpus consists of a collection of recordings from a number of speakers. Constructing a corpus that is appropriate for the task at hand is essential since the corpus is the cornerstone of all experiments. Not only is the data used for training the various transformation methods but also the availability of a representative range of test utterances directly influences the success of the evaluation procedure. The speech corpus for this study contains data from CMU arctic database.

3.3 Pitch conversion algorithm

3.3.1 Mean/variance linear transformation

Substantial work has been undertaken by [7, 9, 10] on the area of spectral conversion, which involves mapping between spectral envelopes of two speakers at the segmental level. To handle the pitch transformation, all three works employ the mean-variance linear conversion method, with the assumption that "average pitch frequency already carries a great deal of the speaker specic information" [9]. This method involves converting the source f_0 values such that the converted contour matches the average pitch value and the pitch range of the target speaker, while maintaining the intonation pattern of the source. The underlying assumption is that each speaker's f_0 values belong to a Gaussian distribution with a specific mean and variance. A linear transformation can then be defined as follows:

$$t = h(s) = as + b \tag{3.1}$$

where is the instantaneous target pitch value and is the instantaneous source pitch value. The goal is to come up with a and b in terms of the mean and variance values of the two Gaussian distributions.

If the target pitch values have a pdf of p_t with parameters $(\mu_{t,}\sigma_t)$ and the source pitch values have a pdf with parameters $(\mu_{s,}\sigma_s)$, the pdf of the linear transformation can be written as follows[11]:

$$\left[p_t(t) = \frac{\partial h^{-1}(t)}{\partial t} p_s(h^{-1}(t))\right] \tag{3.2}$$

Taking the inverse of the linear transformation in eq.(3.2)

$$h^{-1}(t) = \frac{t-b}{a} \tag{3.3}$$

$$\left[\frac{\partial h^{-1}(t)}{\partial t} = \frac{1}{a}\right] \tag{3.4}$$

Therefore eq.(3.2) can be rewritten as,

$$p_t(t) = \frac{1}{a} p_s(\frac{t-b}{a}) \tag{3.5}$$

Replacing both sides by the expression for a Gaussian distribution, we get

$$\left[\frac{1}{\sqrt{2\pi\sigma_t^2}} \exp(-\frac{(t-\mu_t)^2}{2\sigma_t^2}) = \frac{1}{a}\frac{1}{\sqrt{2\pi\sigma_s^2}} \exp(\frac{(\frac{t-b}{a}-\mu_s)^2}{2\sigma_s^2})\right] \tag{3.6}$$

Taking the logarithm of both sides we end up with

$$\log(\frac{a\sigma_s}{\sigma_t}) - \frac{(t-\mu_t)^2}{2\sigma_t^2} = -\frac{(\frac{t-b}{a}-\mu_s)^2}{2\sigma_s^2} \tag{3.7}$$

The second order terms can then be equated to produce an expression for a in terms of the

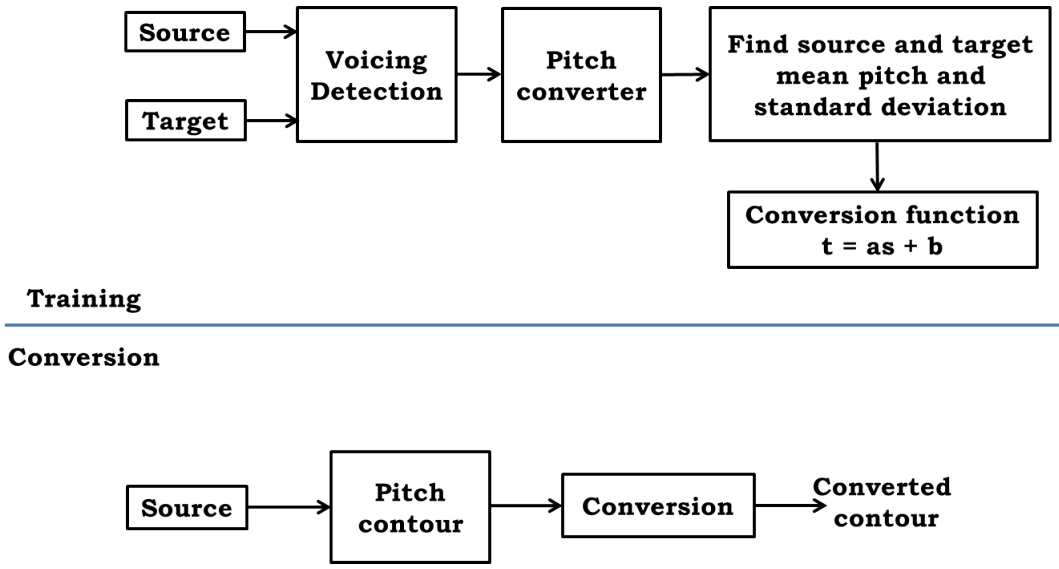

Figure 3.2: Block diagram for pitch contour mapping using mean/ variance method

variances of the source and target pitch distribution.

$$\frac{1}{2\sigma_t^2} = -\frac{1}{2a^2\sigma_s^2} \tag{3.8}$$

$$a = \frac{\sigma_t}{\sigma_s} \tag{3.9}$$

Substituting the value of a into eq.(**??**) we get

$$b = \mu_t - \frac{\sigma_t \mu_s}{\sigma_s} \tag{3.10}$$

Putting the values of a and b in eq.(3.1), we get the linear transformation conversion as

$$t = (\frac{\sigma_t}{\sigma_s})s + (\mu_t - \frac{\sigma_t \mu_s}{\sigma_s}) \tag{3.11}$$

The flow diagram of the pitch conversion system can be seen in Fig. 3.2:

- **Training**

The first step is to acquire a set of training sentences spoken by both the source and target. The source mean pitch value is acquired by averaging the pitch values of all voiced sections in training utterances spoken by source. In the same way target mean pitch value is obtained. The source pitch standard deviation value is calculated using the formula

$$\sigma_s = \sqrt{\frac{1}{N_s - 1}\sum_{n=1}^{N_s}(s_n - \mu_s)^2} \tag{3.12}$$

where is the total number source utterances in training database and is instantaneous source pitch value. Thirty utterances of each speaker were used to extract the mean pitch and pitch standard

deviation over all voiced frames. These values were set aside for each speaker. This concludes the training phase of mean/variance method.

- **Conversion**

In the conversion stage, the pitch contour of the new input source utterance was extracted as detailed in sec 3.2.2. Pitch values of each voiced frame is then converted using eq. 11.

3.4 Mean/Varience method Results

Figure 3.3 and fig. 3.4 show the source contour and target contour for the utterance *"It seemed nearer to him since he had seen and talked with Gregson."* Source is a male speaker and target is a female speaker. Note that pitch range for source speaker is roughly 100 to 150 Hz and for target speaker it is roughly 200 to 220 Hz. Figure 3.5 shows the source and converted contours. As a result of mean/ variance linear transformation from the figures it can be observed that this technique fails in capturing the local rises and falls of the original target contour although the pitch range is more or less converted to target pitch range. Note that if the source parameters were to be resynthesized with the new contour at this point, the resulting speech signal would retain all the original voice characteristics of the source speaker except for the pitch contour which has been converted to the target. This results in a very awkward speech signal, which does not sound similar to either the source or the target and it is very difficult to evaluate objectively or subjectively.This further motivates for exploring alternative algorithms in pitch conversion.

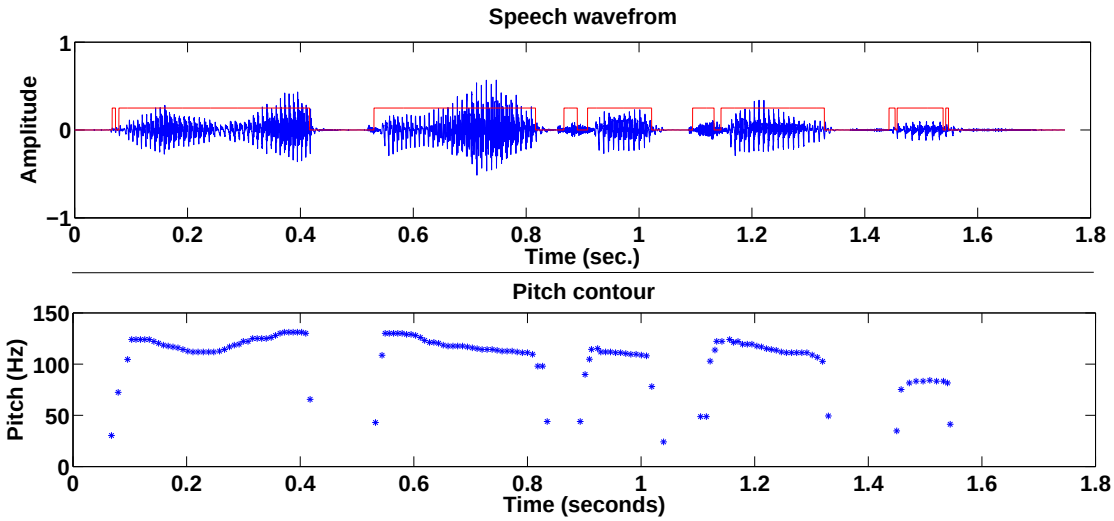

Figure 3.3: Source contour

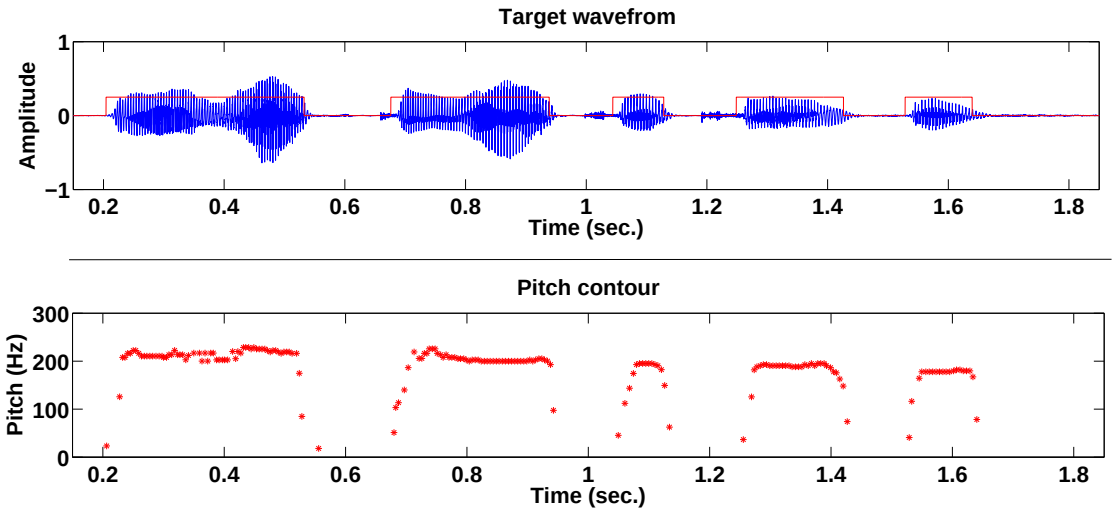

Figure 3.4: Target contour

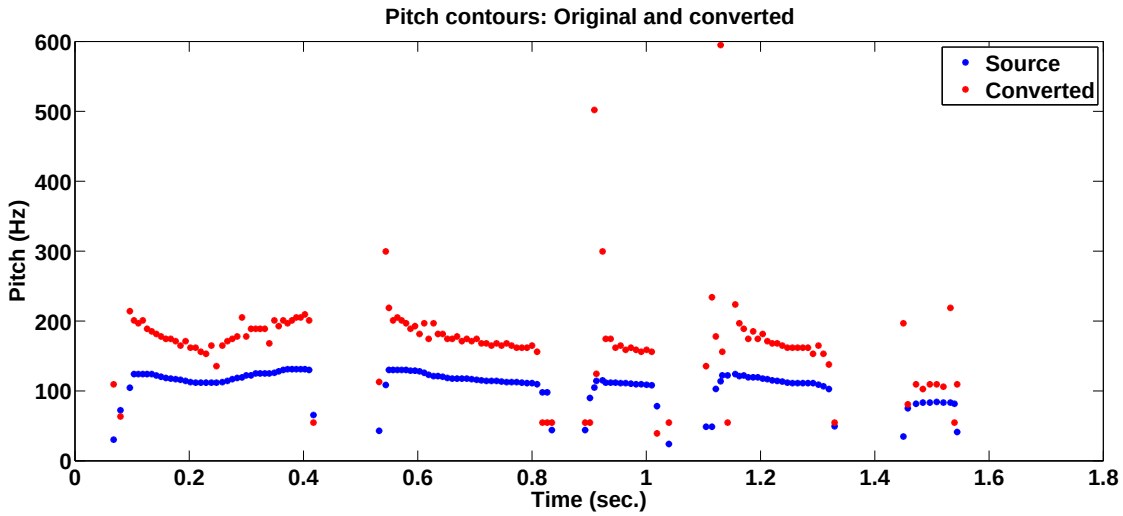

Figure 3.5: Source and converted pitch contours

3.5 Prosody modification using instants of significant excitation

3.5.1 Introduction

Once a target pitch contour has been predicted the next step is to change the source prosody as per the target pitch contour. In this section a method to impose a given pitch contour on the source utterance is presented. The method makes use of the instants of significant excitation. The instants of significant excitation refer to the instants of glottal closure in the voiced region and to some random excitations like the onset of burst in the case of non-voiced regions. The instants of significant excitation are also termed as epochs. The proposed method does not distinguish between voiced and non-voiced regions in the implementation of the desired pitch contour modification.The

method manipulates the source LP-residual using the epochs which are derived from the given pitch contour to be imposed.

3.5.2 Basis for the Method

The proposed method for pitch contour manipulation makes use of the properties of the excitation source information. The residual signal in the Linear Prediction (LP) analysis is used as an excitation signal. This is because while extracting the LP residual from speech using LP analysis, the second order correlations are removed from speech. The residual signal is manipulated by using resampler either for increasing or decreasing the number of samples required for the desired pitch contour modification. The residual manipulation is likely to introduce less distortion in the speech signal synthesized using the modified LP residual and LP coefficients (LPCs). The time-varying vocal tract system characteristics are represented by the LPCs for each analysis frame. Since the LPCs carry the information about the short-time spectral envelope, they are not altered in the proposed method for pitch contour modification. LP analysis is carried out over short segments (analysis frames) of speech data to derive the LP coefficients and the LP residual for the speech signal. There are four main steps involved in the pitch contour manipulation:

1. Deriving the instants of significant excitation (epochs) from the LP residual signal.
2. Deriving a modified (new) epoch sequence according to the desired pitch contour.
3. Deriving a modified LP residual signal from the modified epoch sequence, and
4. Synthesizing speech using the modified LP residual and the LPCs. The method of extracting the instants of significant excitation was discussed in section of chap in this thesis.

Throughout this study a tenth-order LP analysis is performed using a frame size of 20 ms and a frame shift of 5 ms. Speech signal sampled at 8 kHz is used. The signal in the analysis frame is multiplied with a Hamming window to generate a windowed signal. Note that for nonvoiced speech, the epochs occur at random instants, whereas for voiced speech the epochs occur in the regions of the glottal closure, where the LP residual error is large. The time interval between two successive epochs corresponds to the pitch period for voiced speech. With each epoch, three parameters, namely, time instant, epoch interval, and LP residual are associated. These parameters can be called as epoch parameters.

The prosody manipulation involves deriving a new excitation (LP residual) signal by incorporating the desired modification in the duration and pitch period for the utterance. This is done by first creating a new sequence of epochs from the original sequence of epochs. For this purpose, all the epochs derived from the original signal are considered, irrespective of whether they correspond to a voiced segment or a non-voiced segment. The methods for creating the new epoch sequence for the desired prosody modification are discussed in nesxt sections. For each epoch in the new epoch sequence, the nearest epoch in the original epoch sequence is determined, and, thus, the corresponding epoch parameters are identified. The original LP residual is modified in the epoch intervals of the new epoch sequence, and, thus, a modified excitation (LP residual) signal is generated. The modified LP residual signal is then used to excite the time varying all-pole filter represented by the LPCs. For pitch period modification, the filter parameters (LPCs) are updated according to the

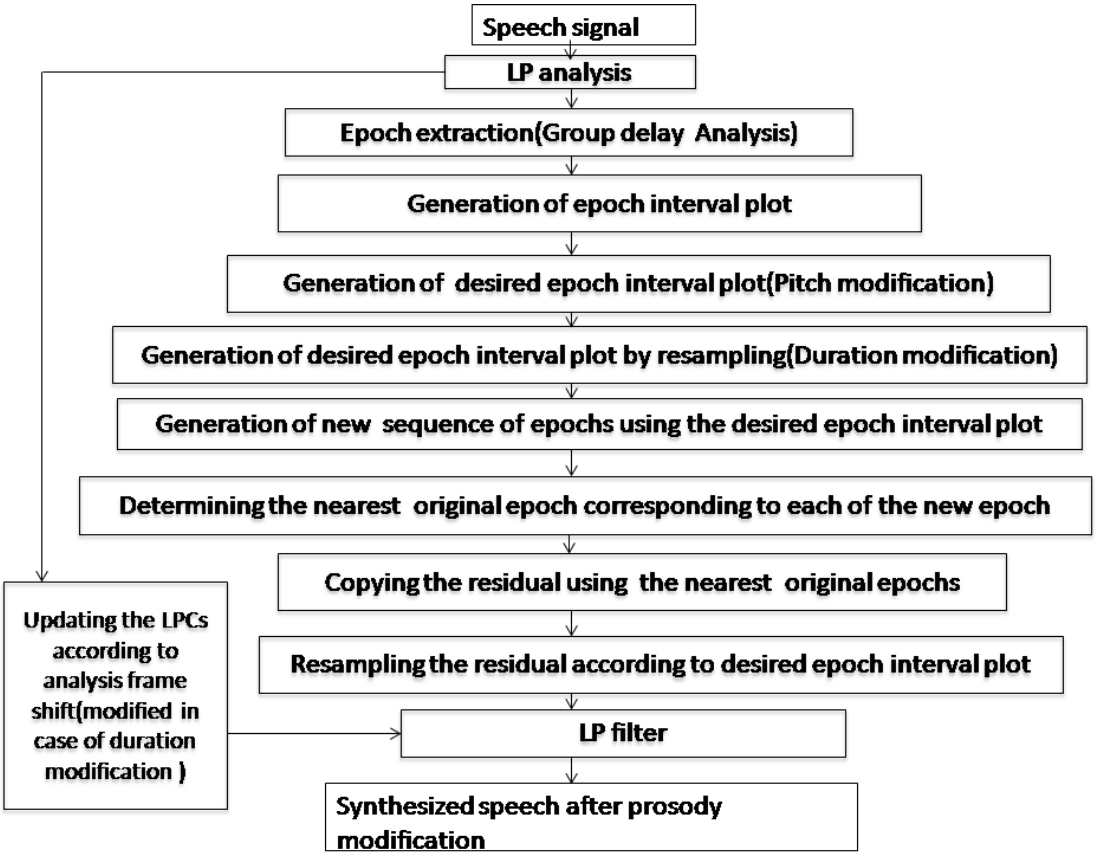

Figure 3.6: Prosody modification using epoch based approach

frame shift used for analysis of the original signal. For duration modification, the LPCs are updated according to the modified frame shift value. Generation of the modified LP residual according to the desired pitch period and duration modification factors is described in sec.3.5.5 and the speech synthesis procedure in sec.3.6. Figure 3.6 shows the block diagram indicating various stages in prosody modification.

3.5.3 Pitch period modification

The objective is to generate a new epoch sequence, and then a new LP residual signal according to the desired pitch period modification factor. Note that for generating this signal, all epochs in the sequence are considered, without discriminating the voiced or non-voiced nature of the segment to which the epoch belongs. Thus, it is not necessary to identify the voiced, unvoiced and silence regions of speech. Figure 3.7 illustrates the prosody modification where the pitch period is reduced by a factor $\alpha = 1.5$. The original epoch interval plot is shown by the solid curve, and the desired epoch interval plot is shown by the dotted curve, which is obtained by multiplying the solid curve by α. The original epochs are marked by circles ('o') on these two curves. The epoch interval at any circle is the spacing (in number of samples) between this and the next circle. The solid and dotted curves are obtained by joining the epoch interval values. Starting from the point A, the new epoch interval value is obtained from the dotted curve, and this value is used to mark the next new epoch B along the x–axis. The epoch interval at this instant on the dotted curve is used to generate the next new epoch C, and so on. The new epoch sequence is marked as '$\times$' along the dotted curve, and also along the x–axis. The nearest original epoch for each of these new epochs is also marked as a sequence of circles ('o') along the x–axis. The procedure for the generation of new epoch sequence, when the pitch period is scaled up, is similar to the one used for the case of fig. 3.7, except that the new epoch interval values are obtained from the scaled up plot. Note that in the above discussion for pitch period modification, the random epoch intervals in the nonvoiced regions are modified by the same pitch period modification factor. As we will see later in sec. 3.6, this will not have any effect on the synthesized speech.

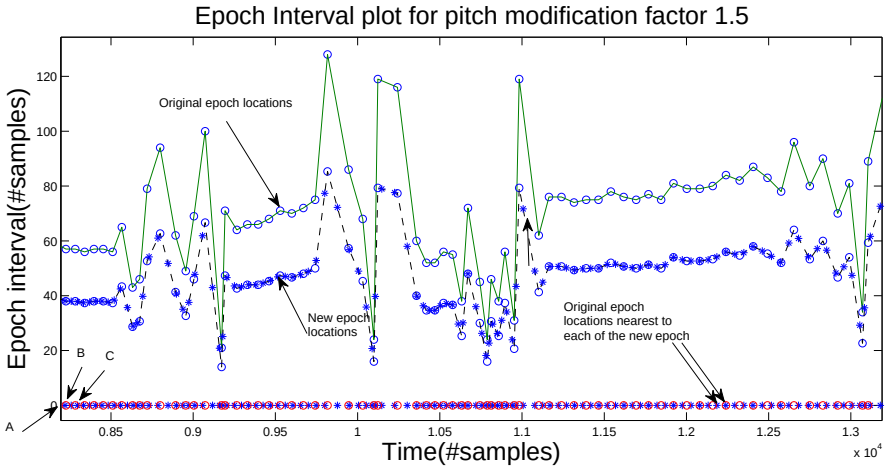

Figure 3.7: Generation of new sequence of epochs for the modification of pitch period by a factor α=1.5

3.5.4 Duration Modification

For generating the desired epoch interval plot for duration modification, the original epoch interval plot is re-sampled according to the desired modification factor. The generation of new epochs using the re-sampled epoch interval is done in the same way as it is done in the case of pitch modification.

3.5.5 Modification of LP residual

After obtaining the modified epoch sequence, the next step is to derive the excitation signal or LP residual. For this, the original epoch (represented by a 'o') closest to the modified epoch '$\times$' is determined from the sequence of 'o' and '$\times$' along the desired epoch interval curve. As mentioned earlier, with each original epoch, i.e., the circles ('o') in the plots, there is an associated LP residual sequence of length equal to the value of the original epoch interval for that epoch. The residual samples are placed starting from the corresponding new epoch. Since the value of the desired epoch interval (M) is different from the value of the corresponding original epoch interval (N), there is a need either to delete some residual samples or append some new residual samples to fill the new epoch interval. Increasing or decreasing the number of LP residual samples for pitch period modification can be done in two ways. In the first method, all the residual samples are used to resample them to the required number of new samples. While there is no distortion perceived in the synthetic speech, the residual samples are expanded or compressed even in the crucial region around the instant of glottal closure. In the second method, a small percentage of the residual samples within a pitch period are retained (i.e., they are not modified), and the rest of the samples are expanded and compressed depending on the pitch period modification factor. The residual samples to be retained are around the instant of glottal closure, as these samples determine the strength and quality of excitation. Thus, by retaining these samples, we will be preserving the naturalness of the original voice. The percentage of samples to be retained around the instant of glottal closure may not be critical, but if we use a small number (say less than 10% of pitch period) of samples, then we may miss some crucial information in some pitch periods, especially when the period is small. On the other hand, if we consider large number (say about 30%) of samples, then we may include the complete glottal closure region, which will not change in proportion when the pitch period is modified. In this study 20% of the residual samples are used. In this study, we resample the residual samples instead of deleting or appending the samples. The first $0.2N$ (nearest integer) residual samples are retained and the remaining $N - 0.2N$ residual samples are resampled to generate $M - 0.2N$ new samples.

3.6 Generating the synthetic signal

The modified LP residual signal is used as an excitation signal for the time varying all-pole filter. The filter coefficients are updated for every P samples, where P is the frame shift used for performing the LP analysis. In these studies, a frame shift of 5 ms and a frame size of 20 ms are used for LP analysis. Thus, the samples correspond to 5 ms when the prosody modification does not involve any duration modification. On the other hand, if there is a duration modification by a scale factor β, then the filter coefficients (LPCs) are updated for every P samples corresponding to 5β ms. Since the LP residual is used for incorporating the desired prosody modification, there is no significant distortion due to resampling the residual samples both in the voiced and in the non-voiced regions.

This is because there is less correlation among samples in the LP residual compared to the correlation among the signal samples.

3.7 Results and discussion

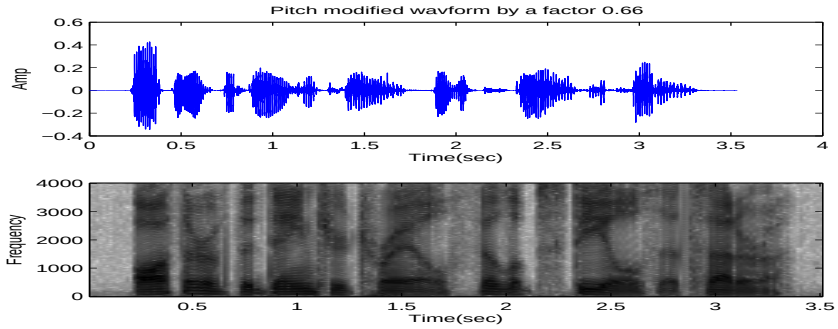

Figure 3.8: Pitch modification by a factor 0.66

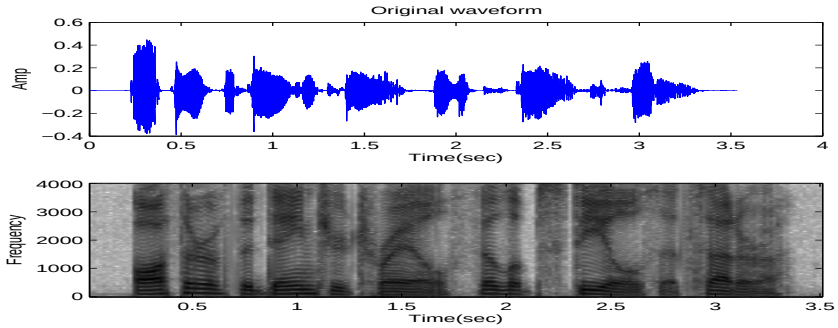

Figure 3.9: Original male voice

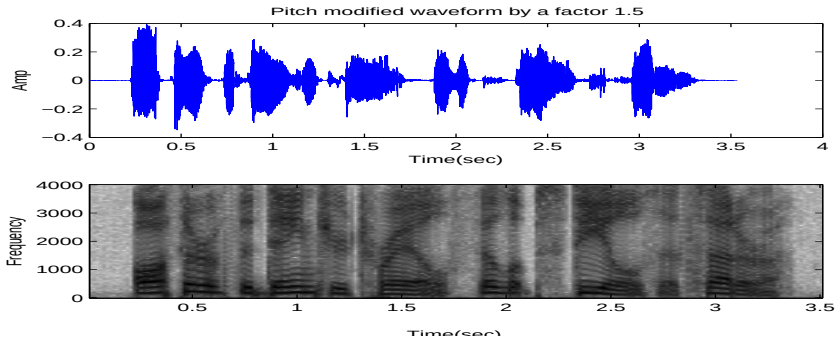

Figure 3.10: Speech signal and its narrowband spectrogram for the utterance Author of the danger trail, Philip Steels, etc. (a)Pitch period modification factor = .66 (pitch decreases in this case), (b) original, and (c) pitch period modification factor = 1.5. (pitch increases in this case).

Figures 3.8 and 3.9 show pitch modified waveforms with narrowband spectrograms and with original waveform in the middle. Epoch based prosody approach provides flexible pitch and duration

modifications by manipulating LP residual. Since the prosody modification is done on the residual, the spectral features are not modified. Thus, there are no spectral distortions. But there will be some degradation in the naturalness of the synthesized speech when large pitch period modification factors are involved, typical range of modification factors observed for this method is from 0.5 to 2.

Figure 3.11 shows modified LP residual for pitch modification factor 0.66 and 1.5 on the same time scale. It can be noticed from the figure that the spacing between the consecutive peaks in LP residual increases in case of $\alpha = 0.66$ (That causes pitch to decreases) and the spacing decreases in case of $\alpha = 1.5$ (That causes pitch to increase).

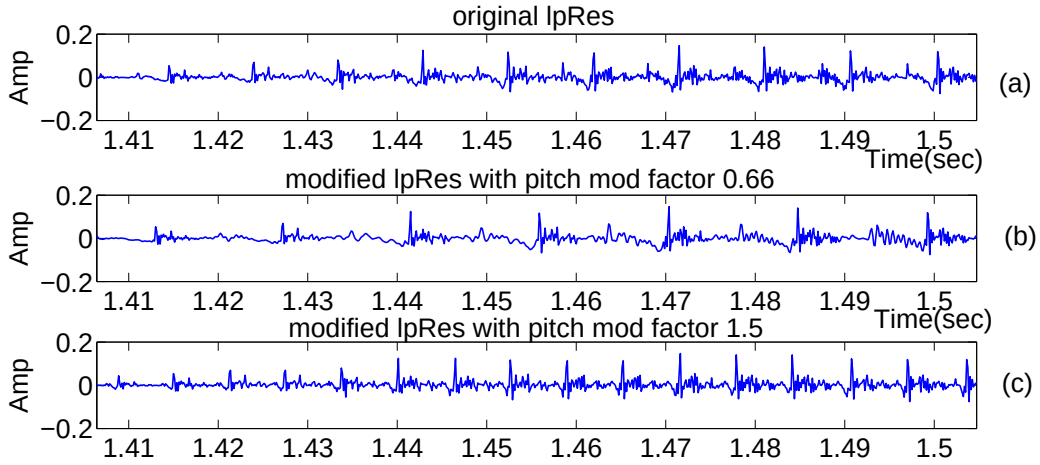

Figure 3.11: Fig (a) Original LP residual, (b) modified LP residual for pitch modification factor 0.66, (c) modified LP residual for pitch modification factor 1.5

Chapter 4

Sinusoidal analysis/synthesis

4.1 Introduction

This chapter presents speech synthesis using model based approach for speech signals. The sinusoidal model of speech waveform for producing synthetic speech is discussed which was first introduced by [4]. The sinusoidal model is based on extracting amplitudes, frequencies and phases of the component sine waves from the short time Fourier transform and using them for the production of synthetic speech. The use of a spectral representation of a sound yields a perspective that is sometimes closer to the one used in a sound-engineering approach. Frequency-domain analysis is a somewhat similar process to the one performed by the human hearing system, it yields fairly intuitive intermediate representations.

4.2 Basics of sinusoidal model

In the linear speech production model, the continuous-time speech waveform $s(t)$ is assumed to be the output of passing a source excitation waveform $u(t)$ through a linear time-varying filter as

$$s(t) = \int_0^t h(t, t-\tau)u(\tau)d\tau \tag{4.1}$$

where the excitation is convolved with a different impulse response at each time t. It is proposed that the excitation $u(t)$ be represented by a sum of sinusoids of various amplitudes, frequencies and phases [5].

$$u(t) = Re \sum_{k=0}^{K(t)} a_k(t) exp[j\phi_k(t)] \tag{4.2}$$

and where phase function

$$\phi_k(t) = \int_0^t \Omega_k(\sigma)d\sigma + \phi_k \tag{4.3}$$

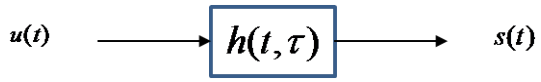

Figure 4.1: $s(t)$ is the output of $h(t, \tau)$

and $K(t)$ is the number of sine wave components at time t.

For the k^{th} sine wave component, $a_k(t)$ and $\Omega_k(t)$ represents the time varying amplitude and frequency, and ϕ_k is a fixed offset to account for the fact that at time $t = 0$, the sine waves are generally not in phase. Now, the vocal tract transfer function in terms of its time-varying magnitude $M(t, \Omega)$ and phase $\Phi(t, \Omega)$ components is written as $H(t, \Omega) = M(t, \Omega)exp\left[j\Phi(t, \Omega)\right]$. Then if the parameters of the excitation, $a_k(t)$ and $\Omega_k(t)$, are constant over the duration of the impulse response of the vocal tract filter, the speech waveform can be written as

$$s(t) = Re \sum_{k=1}^{K(t)} a_k(t) M\left[t, \Omega_k(t)\right] exp\left[j(\int_0^t \Omega_k(\sigma)d\sigma + \Phi\left[t, \Omega_k(t)\right] + \phi_k)\right] \tag{4.4}$$

by combining the excitation and vocal tract amplitudes and phases, the above equation can be written more concisely as

$$s(t) = \sum_{k=1}^{K(t)} A_k(t) exp\left[j\theta_k(t)\right] \tag{4.5}$$

where

$$A_k(t) = a_k(t)M(t, \Omega_k(t))$$

$$\theta_k(t) = \phi_k(t) + \Phi\left[t, \Omega_k(t)\right] = \int_0^t \Omega_k(\sigma)d\sigma + \Phi\left[t, \Omega_k(t)\right] + \phi_k$$

Equation 4.5 is the basic sinewave model that can be thought of as speech-independent, i.e., the model can be applied to any signal. The next step is to develop a robust procedure for extracting the amplitudes, frequencies, and phases of the component sinewaves from the speech waveform.

4.2.1 Estimation of Sine wave Parameters

To obtain a flavor for the analysis and synthesis problem, a simple example of analyzing and synthesizing a single sinewave is considered. Consider the discrete-time counterpart to the continuous-time sinewave model In particular consider a single discrete-time sinewave of the form $x(n) = Acos(\omega_0 n)$ derived from a 500 Hz continuous-time signal at 10000 samples/s. We are motivated to form an estimate of the amplitude and frequency of the sinewave as the amplitude and frequency of the spectral maximum, denote these estimates by $\hat{A}$ and $\hat{\omega_0}$, respectively. We can then synthesize sinusoid as $\hat{x}(n) = \hat{A}cos(\hat{\omega}_0 n)$. Figure 4.2 shows the sinusoid and its Fourier transform. Figure 4.3 shows the estimated peak and location from Fourier transform and synthesized sinusoid back.

The general analysis/ synthesis problem is to take a speech waveform, extract the parameters that represent a quasi-stationary portion of that waveform, and use those parameters to reconstruct an approximation that is "as close as possible" to the original speech. The general estimation

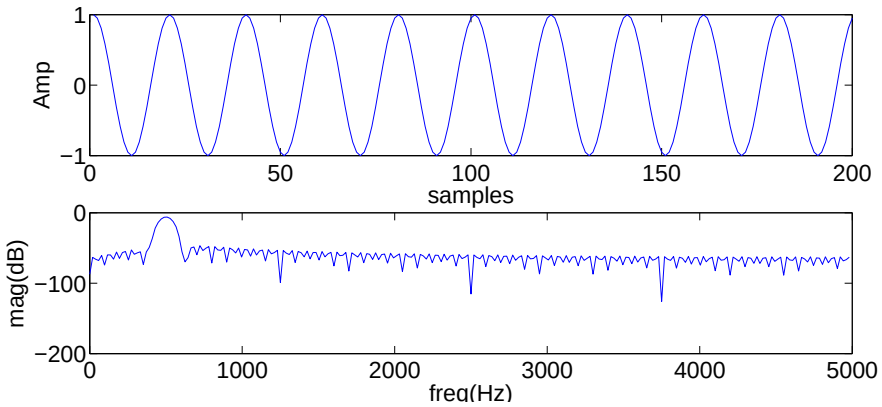

Figure 4.2: Sinusoid and its Fourier transform

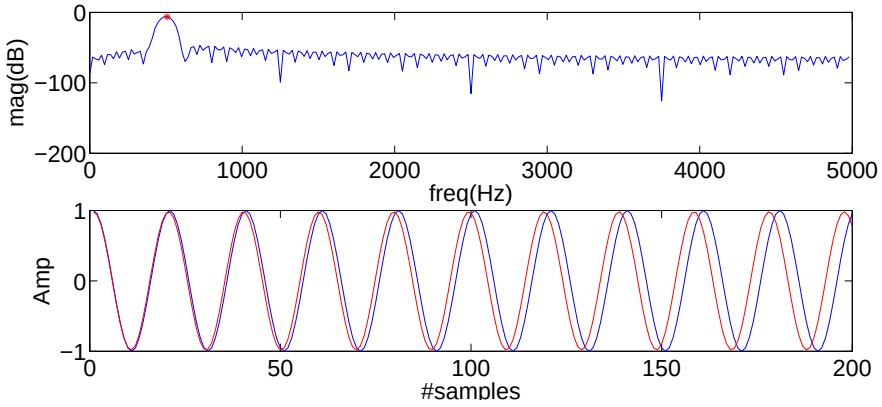

Figure 4.3: The sinusoid (red) is synthesized by picking peak in the Fourier transform

problem in which the speech signal is to be represented by a sum of sinewaves is a difficult one to solve analytically. An estimator is derived based on a Set of idealized assumptions; then, once the structure of the ideal estimator is understood, modifications are made as the assumptions are relaxed to better match the real speech waveform. Looking ahead to fig. 4.4 which shows Fourier transform of a voiced segment of speech. The transform consists several peaks indicating that voiced segment can be reconstructed as a sum of sinusoids by estimating the amplitudes, frequencies and phases from the Fourier transform. The frequency locations and amplitudes of peaks are estimated from absolute magnitude spectrum. Phase values are determined from the unwrapped phase spectrum at frequency locations of the peaks.

For a speech waveform the above procedure for parameter detection is repeated for each frame and the frames are overlap added to reconstruct the speech. The block diagram (fig. 4.6) shows the sinusoidal analysis/ synthesis system.

The accuracy of the peak detection method is limited by the frequency resolution of the FFT. In order to improve the implementation of the sinusoidal model, we have to use more refined methods for estimating the sinusoidal parameters. An efficient spectral interpolation scheme to better measure peak frequencies and magnitudes is to interpolate FFT spectrum, by using samples immediately surrounding the maximum-magnitude sample. This is illustrated in fig.4.7 , where magnitudes are

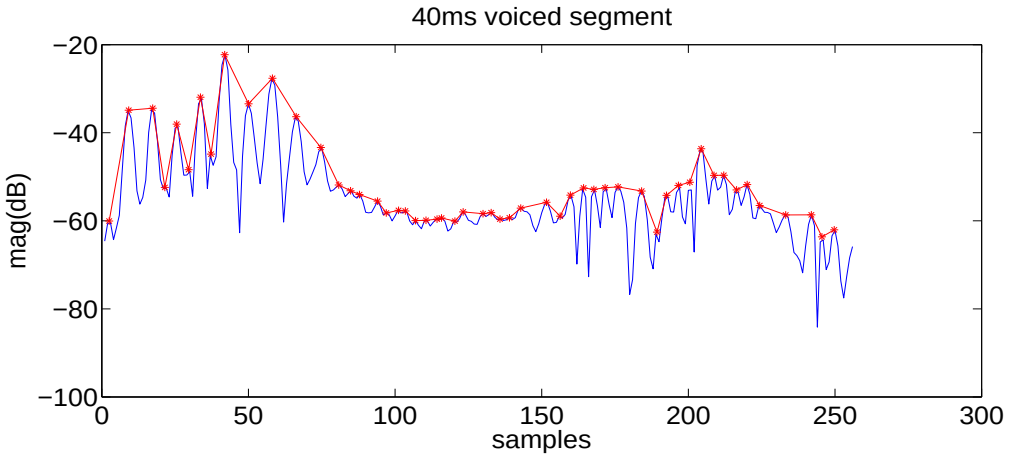

Figure 4.4: Fourier transform of a voiced speech segment and estimated peaks

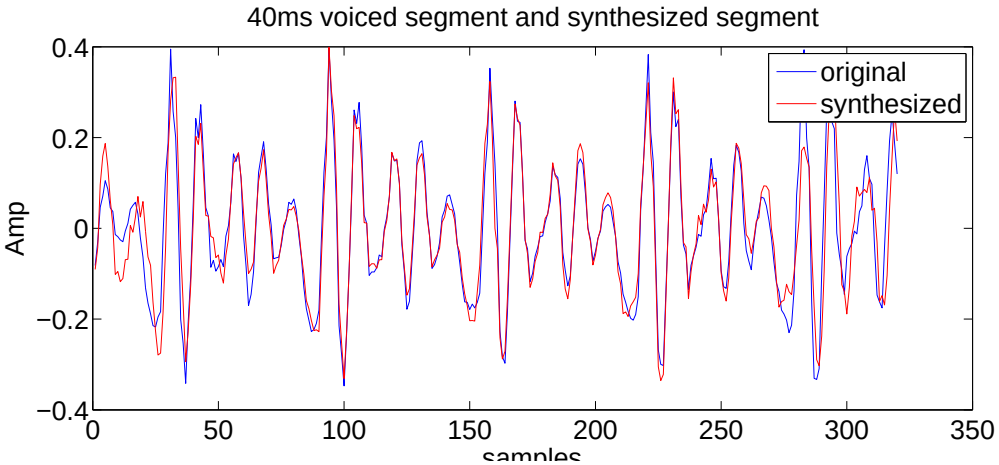

Figure 4.5: Original (blue) and synthesized voiced segment(red) using sinusoidal modeling

expressed in dB. For a spectral peak located at bin k_p , let us define $\left[\alpha = X_w^{dB}(k_p - 1), \beta = X_w^{dB}(k_p)\right]$ and $\gamma = X_w^{dB}(k_p + 1)$. The center of the parabola in bins is $\hat{k}_p = k_p + (\alpha - \gamma)/2(\alpha - 2\beta + \gamma)$, and the estimated amplitude $\hat{a}_p = \beta - (\alpha - \gamma)^2/8(\alpha - 2\beta + \gamma)$. Then the phase value of the peak is measured by reading the value of the unwrapped phase spectrum at the position resulting from the frequency of the peak. The type of window used also has a very strong effect on the qualities of the spectral representation we will obtain two features of the transform of the window are especially relevant to whether a particular function is useful or not: the width of the main lobe, and the main to highest side lobe relation.

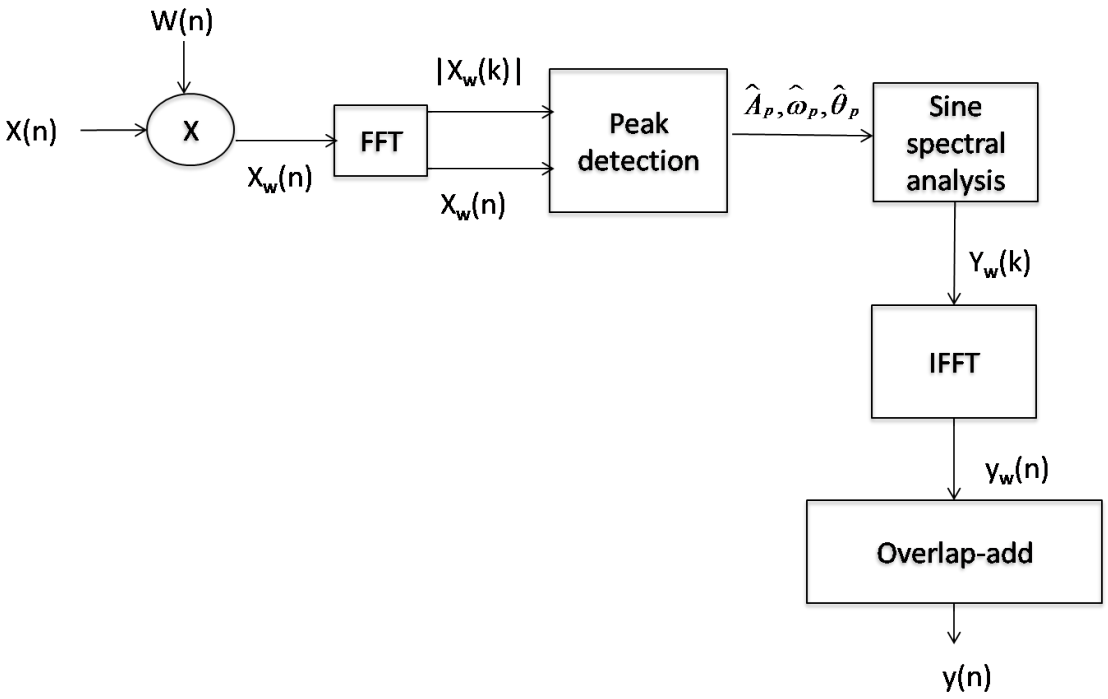

Figure 4.6: Sinusoidal analysis/ synthesis system

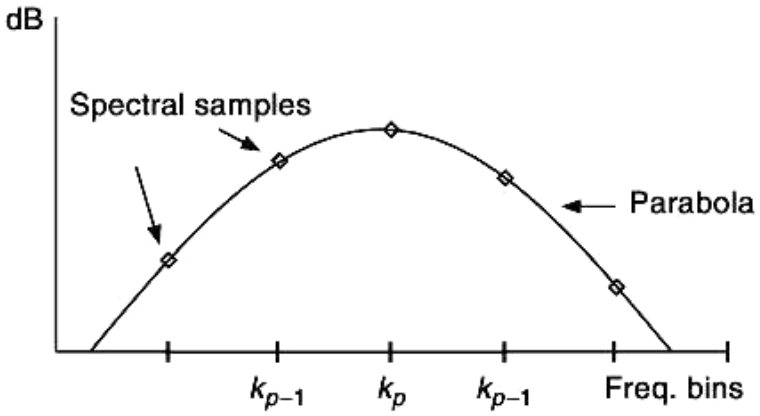

Figure 4.7: Parabolic interpolation of peaks

The main lobe bandwidth is expressed in bins (spectral samples) and, in conjunction with the window size, defines the ability to distinguish two sinusoidal peaks (see fig. 4.8).

The following formula expresses the relation that the window size, M, the main lobe bandwidth, B_s , and the sampling rate, f_s , should meet in order to distinguish two sinusoids of frequencies f_k and f_{k+1}:

$$M \geq B_s \frac{f_s}{|f_{k+1} - f_k|} \tag{4.6}$$

Common windows that can be used in the analysis step are: rectangular, triangular, Kaiser-Bessel,

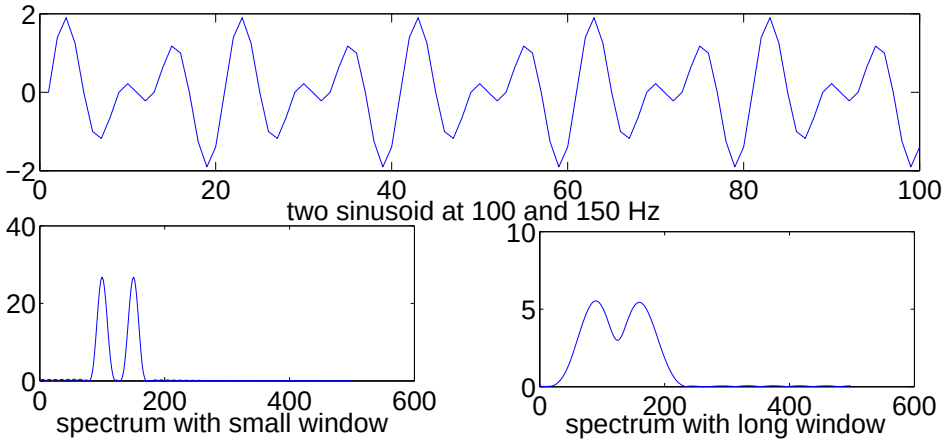

Figure 4.8: Effect of window size in distinguishing between two sinusoids

Hamming, Hanning and Blackman-Harris. The fig. 4.9 shows whole speech utterance synthesized using sinusoidal modeling.

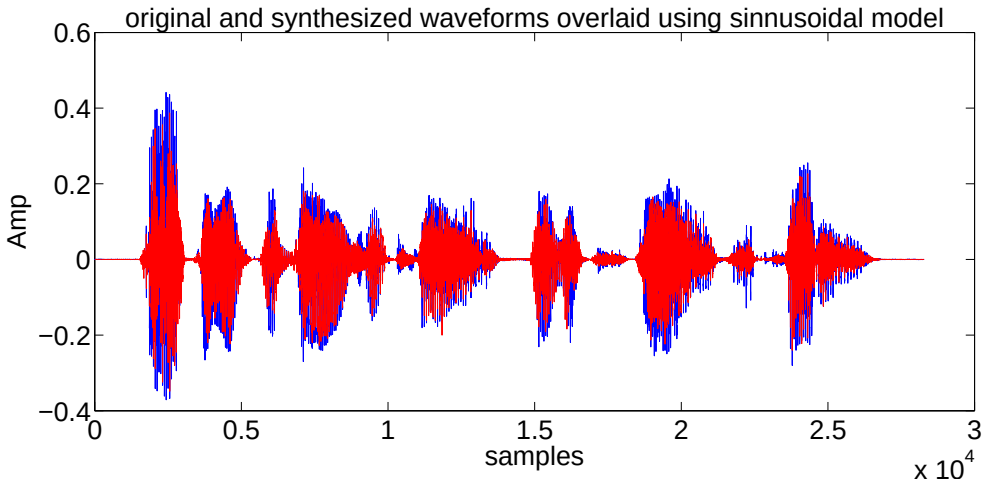

Figure 4.9: The orginal(blue) and synthesized speech(red) using sinusoidal modeling

4.3 Spectral Harmonics

4.3.1 From sinusoidal model to harmonic model

A very useful constraint to be included in the sinusoidal model is to restrict the sinusoids to being harmonic partials, thus to assume that the input sound is monophonic and harmonic. With this constraint it should be possible to identify the fundamental frequency, $F0$, at each frame, and to have a much more compact and flexible spectral representation. When a relevant fundamental frequency is identified, we can decide to what harmonic number each of the peaks belongs and thus restrict the sinusoidal components to be only the harmonic ones. Such a model is called Harmonic model. The diagram of the complete system is shown in fig.4.10. In this project two way mismatch algorithm

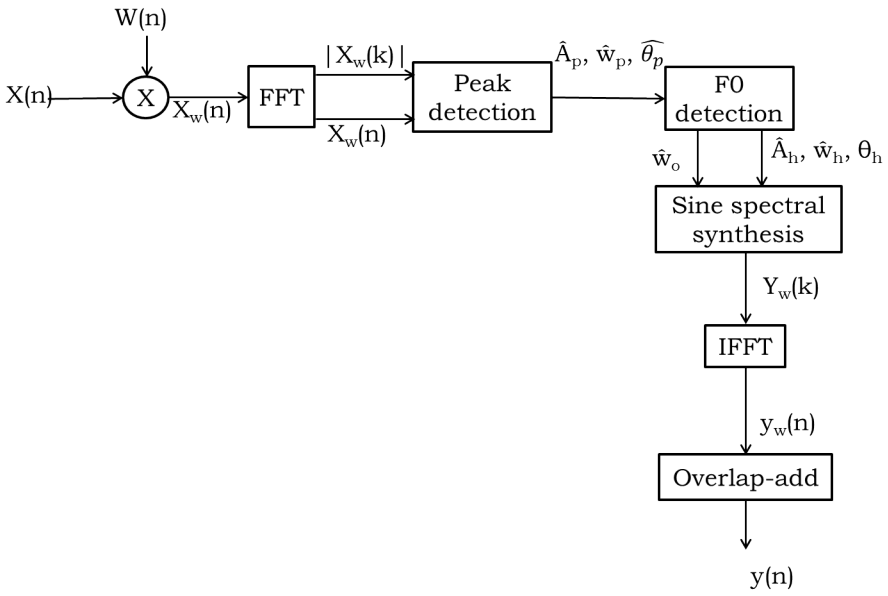

Figure 4.10: Block diagram of an analysis/ synthesis system based on the harmonic model

[6] was used for $F0$ detection. The fig. 4.11 shows synthesized speech using harmonic model.

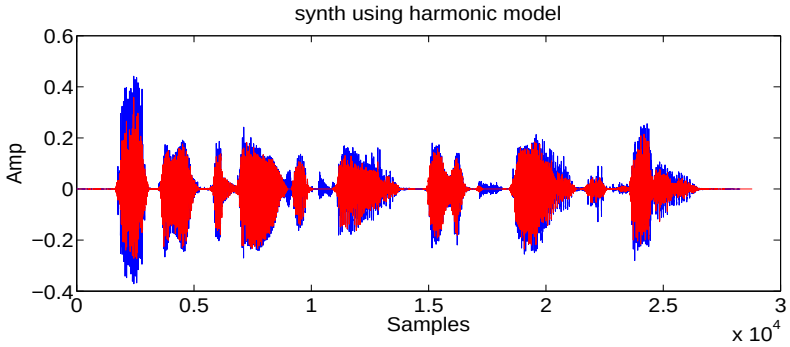

Figure 4.11: Original (blue) and synthesized speech (red) using Harmonic modeling overlaid

4.3.2 Spectral Harmonics plus residual model

Once we have identified the harmonic partials of a sound, we can subtract them from the original signal and obtain a residual component. It is shown in the block diagram fig. 4.12.The harmonic and residual waveforms are shown in 4.13.The sinusoidal subtraction in the spectral domain is in many cases computationally simple and hence residual spectrum is obtained by subtracting harmonics spectrum from original spectrum then IFFT is applied on residual spectrum to get residual in time domain. It is worth to mention here that the sinusoidal information obtained from the analysis is very much under-sampled, since for every sinusoid we only have the value at the tip of the peaks, and thus we have to re-generate all the spectral samples that belong to the sinusoidal peak to be subtracted. The synthesized signal, $y(n)$, is the sum of the harmonic and residual components. The function also returns the harmonic and residual components as separate signals.

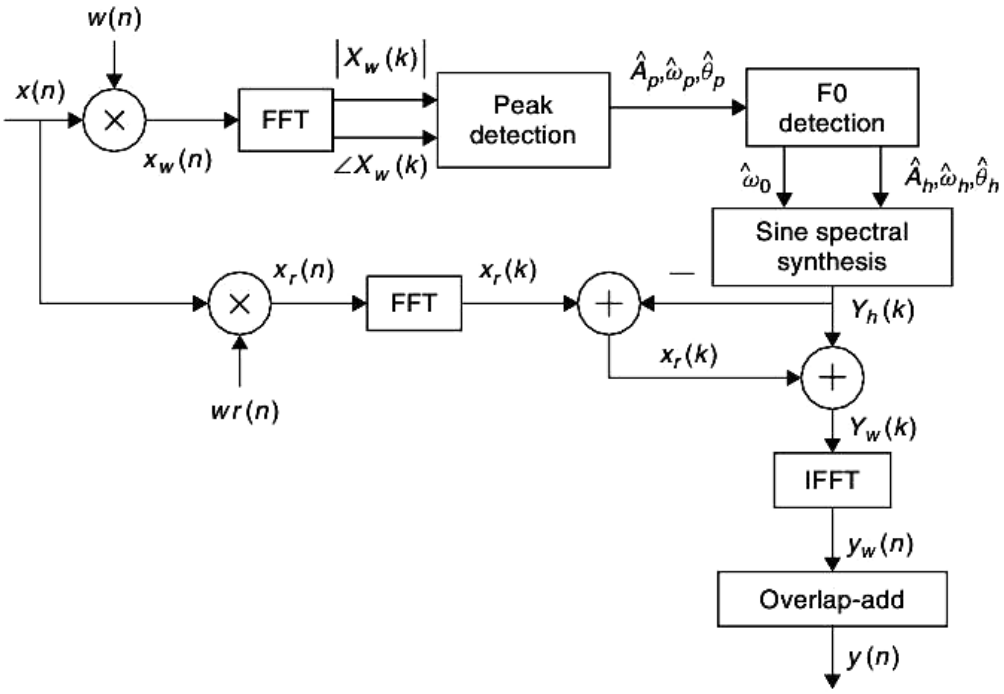

Figure 4.12: Block diagram of harmonics plus residual analysis/synthesis system

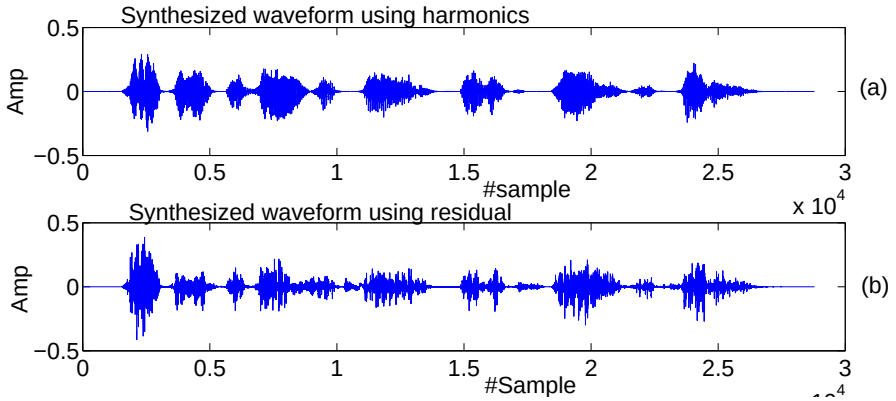

Figure 4.13: synthesized waveforms using(a)Harmonics(b)Residual

Once we have, either the residual spectrum or the residual time signal, it is useful to study it in order to check how well the partials of the sound were subtracted and therefore analyzed. If partials remain in the residual, the possibilities for transformations will be reduced, since these are not adequate for typical residual models. In this case, we should re-analyze the sound until we get a good residual, free of harmonic partials. Ideally, for monophonic signals, the resulting residual should be as close as possible to a stochastic signal.

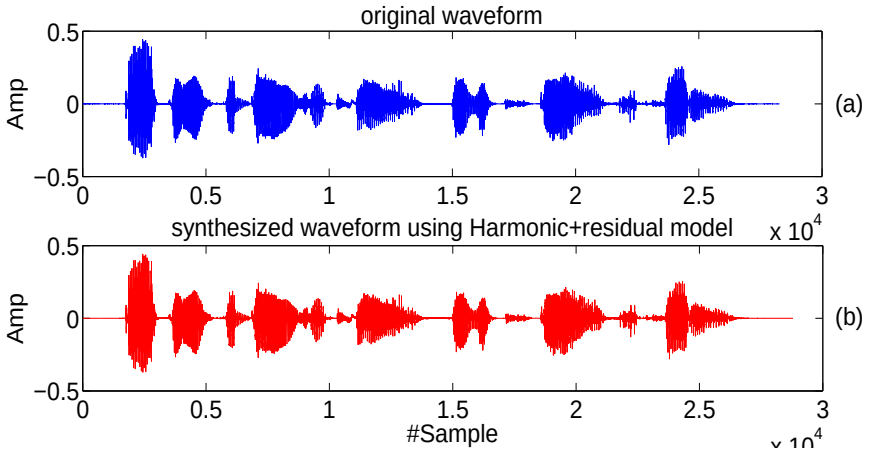

Figure 4.14: (a)Original and (b)synthesized speech using Harmonic plus residual model

4.4 Sinusoidal analysis/synthesis Conclusions

In this chapter, an analysis/synthesis system based on a sinusoidal model has been presented. The resulting representation is characterized by the amplitudes, frequencies, and phases of the component sine waves. The sinusoidal modeling of speech is widely used for voice transformation. Harmonic plus residual model is a simplification applied to sinusoidal model. In harmonic plus residual model the residual can further be analyzed for the acoustic cues present in it. In sinusoidal modeling the estimation of parameters is of crucial importance. It appears that the resulting quality is strongly influenced by the way the parameters are extracted from the data. Here the model parameters were extracted in frequency domain from FFT spectrum; new methods in time domain for model parameter estimation can be explored.

Chapter 5

Summary and Conclusions

Speech modifications require development of basic understanding of speech production mechanism. Source/ Filter theory for speech production mechanism provides good framework for voice conversion as it allows to represent speech as a composite consequence of source signal (excitation) and system.In order to achieve high quality of synthesized speech voice conversion techniques aim to find the answers of two questions: What can we transform? and How to transform? Voice modification techniques can be applied for speech modifications at source level and system level separately and modified components can be combined to synthesize modified speech.

5.1 Interpretation of results

This work was directed to study speech modifications at source level.At source level the prosody modifications are of foremost importance in a voice conversion framework.In this dissertation pitch and time scale modifications were implemented using two methods TD-PSOLA and linear prediction residual modification using epoch based approach.The TD-PSOLA algorithm is computationally simple and quite straight forward to implement but it has several drawbacks.By listening the modified speech wavforms it was noticed that the naturalness of the modified speech was preserved for both the methods.In terms of naturalness of synthesized speech for both of the methods the time scale modification provided perceptually good quality of synthesized speech as compared to pitch scale modification.At the same time it was also observed that if the same speech utterance is modified using TD-PSOLA and epoch based method separately while keeping the modification factor same,synthesized wavforms sounded perceptually same but the advantage of epoch based approach over TD-PSOLA is that epoch based method can be useful to modify pitch in an unconstrained manner as in epoch based approach speech is synthesized by modifying LP residual with the help of new epochs on an arbitrary pitch contour which may be a predicted pitch contour.

The prediction of target pitch contour is always required in voice conversion framework.For this purpose the simplest approach mean/varience method was discussed.This method shifts only pitch range values but does not capture local intonational patterns.

The TD-PSOLA and epoch based approach are non-parametric approaches as no speech model is required for speech modifications in these two approaches.Sinusoidal speech modeling is a parametric approach which is also useful for speech modifications.The estimation of sinusoidal parameters

using short time fourier transform and speech analysis/synthesis was discussed.As a simplification of sinusoidal speech model the speech was sythesized without any modification using harmonic and Harmonic plus residual models.The model based approach can give more flexibility for speech modifications at the cost of comutational efficiency.This work does not discuss speech modification using sinusoidal modeling.

5.2 Directions for future work

In this work the prosody modifications were carried out for constant factors in general modifications should be applied in a time varing manner over the speech utterance.
This work does not discuss modifications at system level.The system response has the information about the speaker identity.In order to develope the complete voice conversion system modifications at system level are required.
There is a need to work upon synthesis step because the speech synthesis techniques play a key role to maintain the quality of final modified speech.

References

[1] R. Smits, B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function", *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.

[2] E. Moulines, J. Laroche, "Non-Parametric Techniques for Pitch-Scale Modification of Speech", *Speech Communication*, vol. 16, pp. 175–205, 1995

[3] E. Moulines, F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, vol. 9, no. 5/6, pp. 453–467.

[4] R. J. McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. Acoust.,Speech ,Signal Processing*, vol. 34, no 4, pp. 744–754, Aug. 1986.

[5] T. F. Quatieri, "Discrete-Time Speech Signal Processing", *Prentice Hall*, 2002.

[6] R. C. Maher, J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure", *J. Acoust. Soc. Am.*, vol. 95, no. 4, pp. 2254–2263, 1994.

[7] Y. Stylianou, O. Capp, E. Moulines, "Continuous Probabilistic Transforms for Voice Conversion", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[8] X. Huang, A. Acero, H. W. Hon, "Spoken Language Processing – A Guide to Theory, Algorithm and System Development", *Prentice Hall*, 2001.

[9] L. Arslan, 'Speaker Transformation Algorithm using Segmental Codebooks (STASC)", *Speech Communication*, 1999.

[10] A. Kain, M. W. Macon, "Spectral Voice Conversion For Text-To-Speech Synthesis", *Proc. IEEE ICASSP.*

[11] J. A. Rice, "Mathematical Statistics and Data Analysis", *Duxbury Press*, 1995.

[12] K. S. Rao, B. Yegnanarayana, "Prosody Modification Using Instants of Significant Excitation", *IEEE trans. on audio, speech and language processing*, vol. 14, no. 3, May 2006

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

[14] S. R. M. Prasanna, C. S. Gupta, B. Yegnarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", *Speech Commun.*, vol. 48, pp. 1243–1261, 2006.

[15] S. R. M. Prasanna, P. K. Murthy, B. Yegnanarayana, "Speech enhancement using source features and group delay analysis", *IEEE INDICON*, pp. 19–23, Dec. 2005.

[16] K. S. Rao, "Unconstrained Pitch Contour Modification Using Instants of Significant Excitation", *Circuits Syst. Signal Process.*, vol. 31, pp. 2133–2152, 2012.