

Development of a Voice Conversion System

Minor Project Report

Submitted in the partial fulfillment of the requirements for the degree of

Bachelor of Technology

In

ELECTRONICS AND COMMUNICATIONS ENGINEERING

By

RAJVI SHAH [Roll No. 05BEC076]

PRIYA VAYA [Roll No. 05BEC093]

Under the guidance of

Mr. Akash Mecwan



Department of Electrical Engineering
Electronics & Communication Engineering Branch
Institute of Technology
Ahmedabad 382 481
October 2008

CERTIFICATE

This is to certify that the Minor Project Report entitled “**Development of a Voice Conversion System**” submitted by the following students as the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Electronics & Communication of Institute of Technology **Nirma University** is the record of work carried out by them under our supervision and guidance. The work submitted has in our opinion reached a level required for being accepted for the examination. The results embodied in this minor project work to the best of our knowledge have not been submitted to any other University or Institution for award of any degree or diploma.

Roll No.	Name of the Student
05BEC076	RAJVI SHAH
05BEC093	PRIYA VAYA

Date:

Mr. Akash Mecwan
Project Guide.

Prof. A.S.Ranade
HOD, (EE).

ACKNOWLEDGEMENT

We would like to take the opportunity to express our sincere gratitude to our Project Guide, **Mr. Akash Mecwan**, Lecturer, Electronics and Communication Department, Institute of Technology, Nirma University, for first of all agreeing to guide us for the project. Without his constant motivation and support, this work may not have reached the level that it has. Also, with heart-felt gratitude, we would like to thank all the Speech Researchers across the globe for making their research available on the internet and hence providing the new-comers a basic platform.

We are also thankful to the Library, Institute of Technology, Nirma University for providing access to the valuable resources.

Lots of love and gratitude to our parents - for their constant love and support.

TABLE OF CONTENTS

Chapter No.	Title	Page No.
	LIST OF FIGURES.....	VI
	LIST OF TABLES.....	VIII
	ABSTRACT.....	IX
1	INTRODUCTION.....	1
1.1	Motivation.....	1
1.2	Voice Conversion Introduction.....	2
1.3	Applications of Voice Conversion.....	3
1.4	General Framework.....	4
2	SPEECH SIGNAL ANALYSIS.....	6
2.1	Human Speech Production System.....	7
2.2	Modeling the Speech Signal.....	9
2.3	Source-Filter Decomposition.....	10
2.3.1	<i>Pre-Emphasis and Framing.....</i>	<i>10</i>
2.3.2	<i>Modeling Source and Filter using LP Analysis.....</i>	<i>11</i>

2.4	Mapping Parameters for voice conversion.....	14
2.4.1	<i>Mapping Vocal Tract Coefficients</i>	<i>15</i>
2.4.2	<i>Mapping or Modifying Excitation Component</i>	<i>17</i>
3	PSOLA BASED APPROACH.....	18
3.1	Aim of the Approach.....	19
3.2	TD-PSOLA.....	19
3.2.1	<i>Pre-processing.....</i>	<i>19</i>
3.2.2	<i>Analysis.....</i>	<i>20</i>
3.2.3	<i>Modification and Synthesis.....</i>	<i>21</i>
4	SPEECH SYNTHESIZER APPROACH.....	23
4.1	Aim of the Approach.....	24
4.2	Pitch Detection using Auto-Correlation.....	24
4.2.1	<i>Auto-Correlation based approach.....</i>	<i>24</i>
4.2.2	<i>Fast Auto-Correlation.....</i>	<i>26</i>
4.3	Excitation Generation and Synthesis.....	27
5	VOICE CONVERSION DEMO.....	29
5.1	Demo 1.....	30
5.2	Demo 2.....	32

5.3	Demo3.....	34
6	RESULTS AND DISCUSSION.....	35
6.1	Quality Measurement of Conversion Process.....	35
6.2	Improvement and Future Work.....	36
	CONCLUSION.....	37
	REFERENCES.....	38
	ABBREVIATIONS.....	39

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1.2.1	Prosodic Dependencies of Human Speech	3
1.4.1	General Framework of a Voice Conversion System	4
2.1.1	Human Speech Production System	7
2.1.2	Frequency Response of Vocal Tract	8
2.2.3	Block Diagram of the Speech Signal Production Process	9
2.3.1	Spectrum of a Speech Segment before and after Pre-Emphasis	10
2.3.2	Hamming Window (left) and Frequency Response (Right)	11
2.3.3	Response of filter characterized by LP coefficients	13
2.3.4	Excitation Components of voiced and unvoiced segments of speech	14
2.4.1	Block Diagram of Pitch and Vocal Tract Mapping Process	15
2.4.2	Spectra of target, source and filtered signal with filter response	16
3.2.1	A pitch marked speech segment	20
3.2.2	Time Scale Expansion(left) and Compression(Right)	21
3.2.3	Pitch scale Compression (Left) and Expansion (Right)	21
4.2.1	Representation of Autocorrelation at a particular shift	25

4.2.2	Autocorrelation function of a speech signal	26
4.3.1	Synthesis Process controlled by voicing detection	28
5.1.1	Screenshot of Demo 1 GUI program	30
5.1.2	Screenshot of Demo 1 GUI program in operation	32
5.2.1	Screenshot of demo 2 in operation	33
5.2.2	Screenshot of Play Menu	33
5.3.1	Result of Speech Synthesizer based approach	34

LIST OF TABLES

Table. No.	Title	Page No.
5.1.1	Demo 1 GUI Components and their functions	31
6.1.1	Performance Measure for PSOLA Based Approach	36
6.1.2	Performance Measure for Speech Synthesizer Approach	36

ABSTRACT

Voice Conversion is a technique which can be used to convert or change the speech uttered by a source speaker in such a manner that it is heard as if spoken by another target speaker. Here, an approach for static voice conversion is developed and implemented. Static speech parameters are the parameters over which speaker has least control such as vocal tract structure, natural pitch of speech etc. Here, two main parameters are considered Vocal Tract Structure and Pitch. Also two different approaches are studied and implemented in MATLAB. In the first approach, source and target speeches are resolved into excitation component and filter component using LPC based source-filter technique and pitch modification is achieved using a method called PSOLA (Pitch Synchronous Overlap-Add). Whereas in the second approach is based on speech generation model governed by voicing detection. For voiced frames pitch is estimated using autocorrelation method and the excitation component is generated using a set of signal generators driven by voicing detection flag. Filter coefficients are modified to approach target speaker coefficients. Finally, a user friendly demo using MATLAB GUI is developed which demonstrate the idea behind the system. This field of Speech Technology can contribute greatly to the Entertainment Industry as well as can significantly reduce the database size for multiple speaker TTS (*Text-to-Speech*) Systems making them more convenient to implement on portable devices.

Chapter 1

Introduction

- 1.1 Motivation
- 1.2 Voice Conversion Introduction
- 1.3 Applications of Voice Conversion
- 1.4 General Framework

1.1 Motivation

Though the humans lack the ability to fly like birds and swim like fishes and though they are physically feeble as compared to many other animals, they have proven themselves to be the fittest in the process of survival. Two striking features made it possible, the human ability of logical thinking and a means to propagate it. That is the fairly advanced auditory system enabling most complex and distinct mechanism of speech production in humans. The idea of studying this great human ability to communicate with others motivated us to pursue the work in the field of Speech Signal Processing. A survey of the current scenario of Speech Technology revealed the main concentration being on Text-to-speech and automatic speech recognition techniques. With little work done in the field of voice conversion, it is yet an undeveloped and naïve field with attractive applications, giving enough of challenge and space for new research. This gave us final directions to put our first step in the field of Speech Technology in terms of Development of a Voice Conversion System.

1.2 Voice Conversion Introduction

Voice Conversion is a process of transforming the parameters of a source voice to those of a target voice. Source voice is a recorded speech whereas the target voice can be either another recorded speech or more general descriptors like pitch or formant frequencies, prosody. These general descriptors can be specified indirectly in terms of age, gender and speaking style.

The parameters mentioned above can be broadly classified into two types,

- 1 **Static Parameters** – These are the parameters over which the speaker has least control. They are natural pitch and inherent vocal tract structure. These parameters can be modeled more accurately and remains more or less same for the static (flat) speech.
- 2 **Dynamic Parameters** – These are the parameters which are controlled by the speaker itself and they can coarsely be termed as prosody of the speaker. Prosody refers to mainly the speaking style and reveals intonation, breath pause, emotions etc.

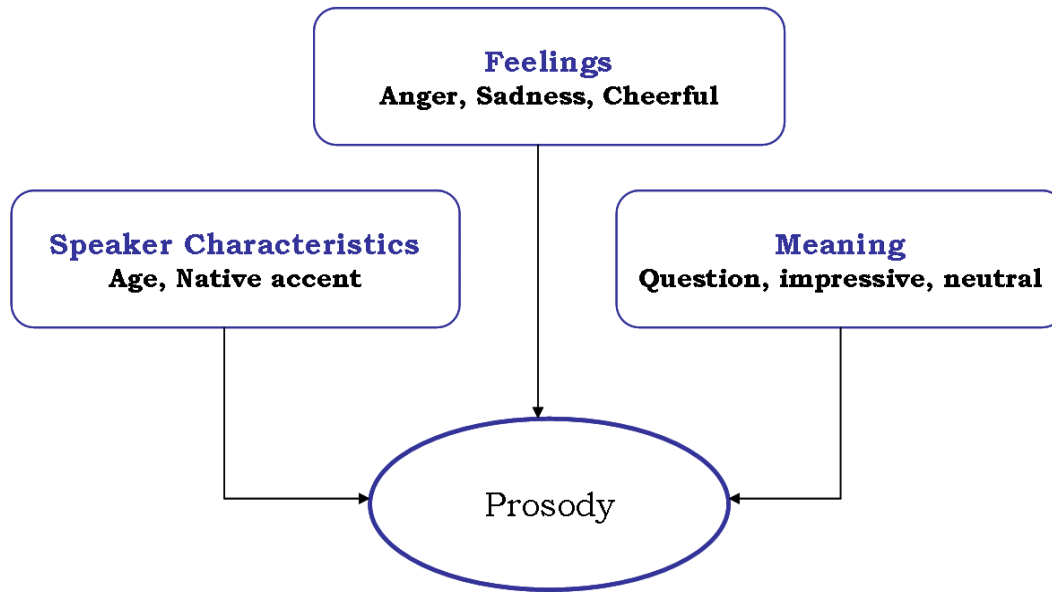


Figure 1.2.1 Prosodic Dependencies of Human Speech

1.3 Applications of Voice Conversion

- 1 Voice conversion has application as a co-module in multiple speaker TTS systems to reduce the database size drastically as with the help of only one speaker's database rest of the voices can be generated with stored parameters.
- 2 Voice conversion has potential applications in many Entertainment Industries.
 - Cross-language voice conversion can be used in dubbing industry preserving original actors' voices [2].
 - Text-independent voice conversion systems can be used to preserve voices of great artists and singers for years [6].
 - Voice Conversion can be used for gaming Avatars giving them unique voices as well as for creating many different voices by a single speaker for animation movies.

1.4 General Framework

The general framework of such a system is shown in figure 1.4.1,

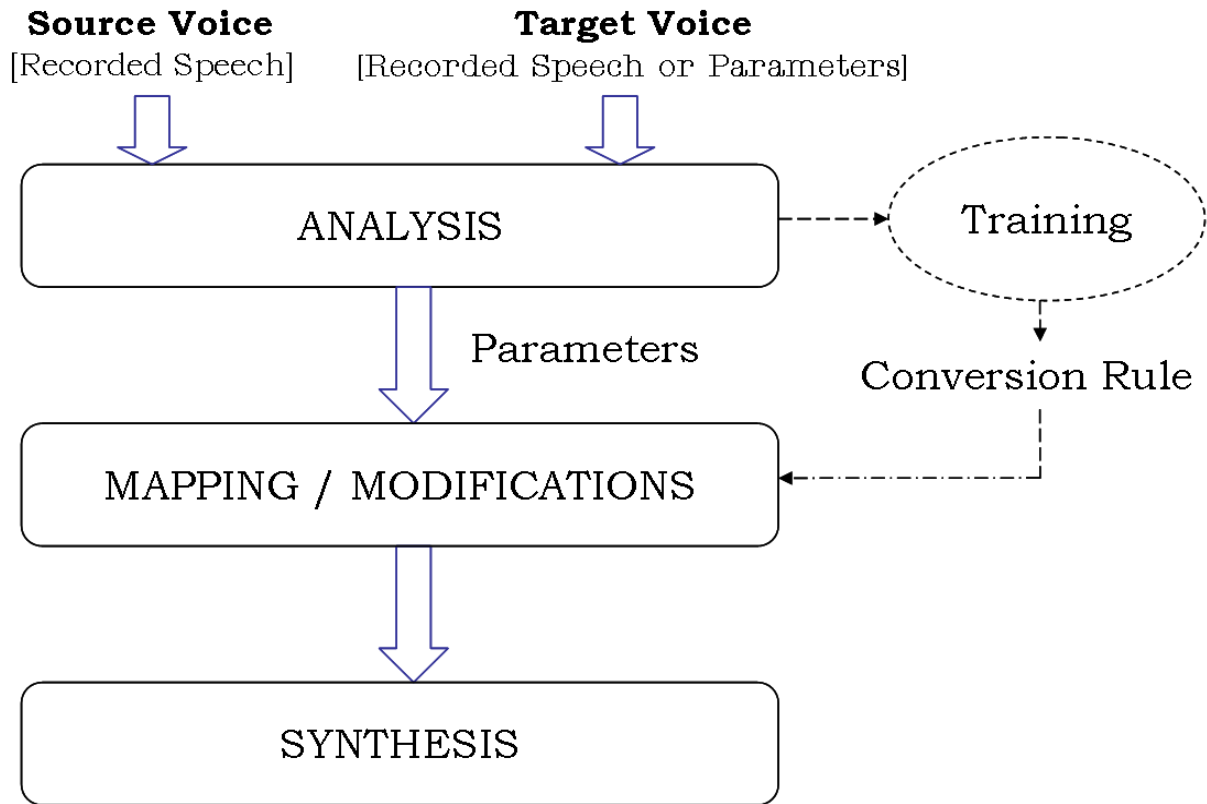


Figure 1.4.1 General Framework of a Voice Conversion System

The main phases involved in any voice conversion system are,

- 1 **Analysis:** In this phase voice parameters corresponding to speaker identity are extracted from source as well as target speech. The parameters are necessarily the parameters related to speaker identity like pitch (F0), Formant Frequencies and prosody etcetera.
- 2 **Mapping:** In this phase the extracted parameters of source speaker are mapped such that they approach the target speech parameters. These parameters might be the extracted parameters from the target speech or provided directly or indirectly (in terms of age, gender) by the system user. This phase is controlled by a conversion rule obtained by a training phase.

- 3 **Synthesis:** The modified parameters are used to synthesize or reconstruct the new speech which shall have the target voice and if system offers, target prosody too.

Chapter 2

Speech Signal Analysis

- 2.1 Human Speech Production System
- 2.2 Modeling the Speech Signal
- 2.3 Source-Filter Decomposition
- 2.4 Mapping parameters for voice conversion

2.1 Human Speech Production System

To understand the voice conversion process, it is mandatory to digest the human speech production process and understand the parameters which are responsible for voice distinction in different humans. The anatomy of human speech production system is shown in figure 2.1.1.

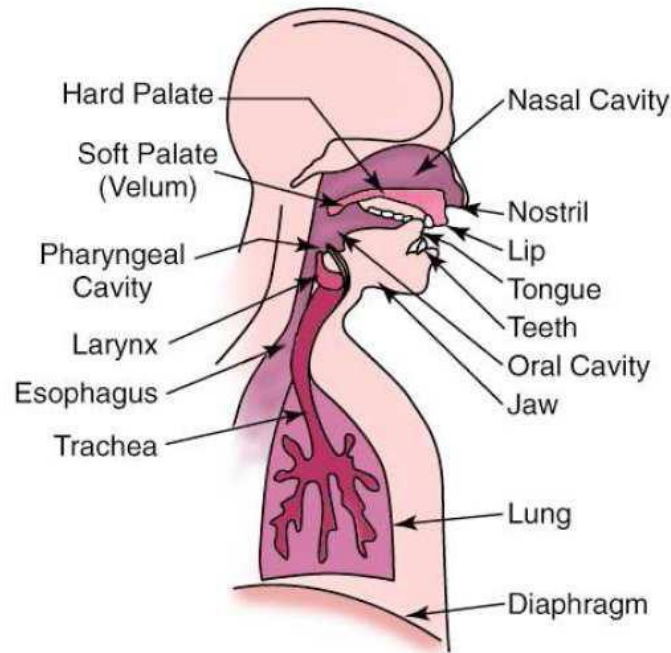


Figure 2.1.1: Human Speech Production System^[12]

The human speech production system begins with the lungs and end with mouth and nasal cavity with neural signals from human brain being the driving or controlling element in the whole speech production process. During speech production in humans, air flow from the lungs passes through the vocal cords first. When the vocal cords are tensed, the airflow causes them to vibrate and hence output of this stage is a periodic signal, such signals are called voiced components of the speech. When the vocal cords are relaxed, air flows more freely through them resulting in a turbulent flow of air and hence output airflow after this stage is very disordered. Such components are called unvoiced components. Amount of tension on vocal cords is driven by neural signal depending on what is to be spoken. A common observation suggests that all the vowels are voiced (/a, /e, /i, /o, /u); consonants whose production involves throat are also voiced (/m, /n). Rest of the consonants whose production is

caused by oral [lips, tongue and mouth] and nasal cavities are unvoiced (/s, /f, /b). So speech production model depends on whether a vowel is being spoken or a consonant. To be more precise whether the speech produced is voiced or unvoiced. Another observation suggests that in any speech voiced vowels possess maximum energy, voiced consonants comes the next whereas unvoiced components possess the least energy. So, for voiced components, periodic airflow coming out of the vocal cords can be described by a periodic pulse train with its period T , hence $F_0 [= 1/T]$ is called the pitch of speech signal. Hence, pitch is a factor directly related to resonance of vocal cords. After passing through the vocal cords the thus produced airflow enters the mouth creating some amount of acoustic disturbances and exits through lips and some times through nasal cavity. The mouth, tongue, teeth, lips [oral cavity] and nasal cavity are named together as vocal tract. The cross-section of this vocal tract tube varies along its length because of varying positions of teeth, tongue and lips. These positions are determined by neural signals depending on which speech component is to be produced. For example, the production of sound “ee” involves spreading the lips and bringing the teeth nearer. These variations result in a Linear Time Invariant System which has a frequency response as shown in Figure 2.1.2.

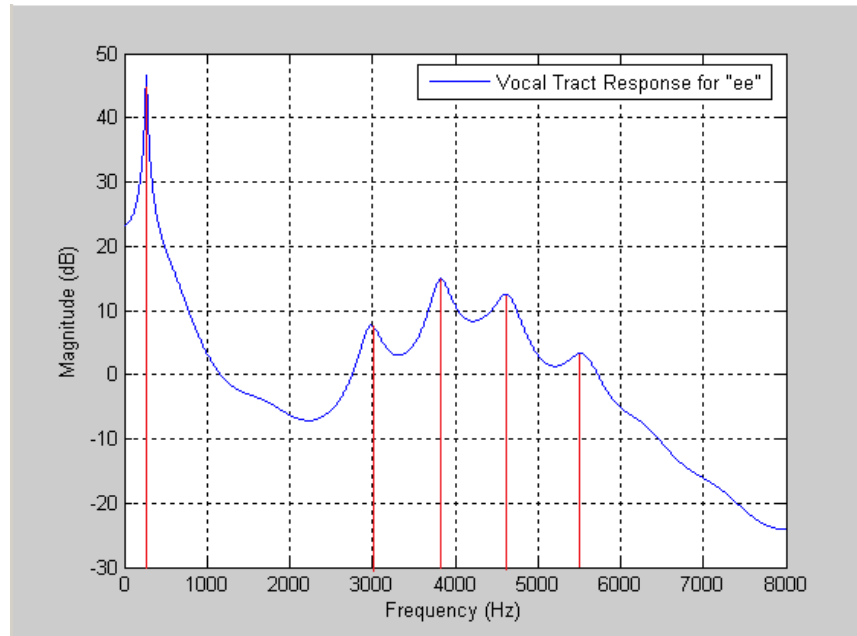


Figure 2.1.2: Frequency Response of Vocal Tract

The peaks on the frequency response curve shown by red lines are referred to as formants, so rather than naming it a filter, it is termed a shaping function which shapes the spectrum of airflow through vocal cords.

2.2 Modeling the Speech Signal

The speech signals can be modeled as unstructured signals generated by a source (lungs) and passed through interconnection of systems which structures the signal to yield speech. The system can be modeled either as a linear or a nonlinear model. Though a linear model does not mimic the exact behavior, it is preferred as it provides a fair amount of accuracy with ease of implementation.

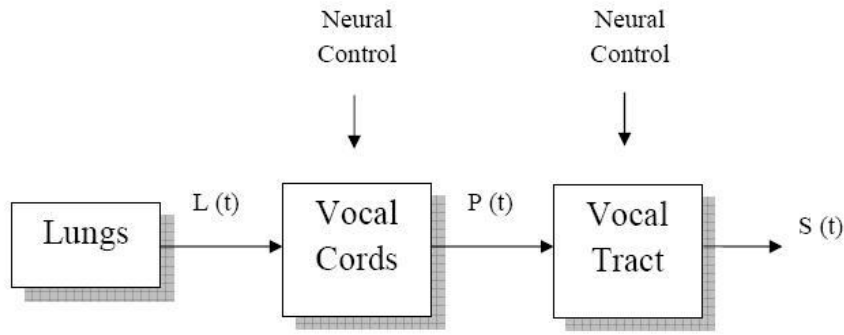


Figure 2.2.3 Block Diagram of the Speech Signal Production Process

The most common approach for modeling the speech signal is Source-Filter model. The Source-Filter model is a model of speech where the spoken word is comprised of a source component originating from the vocal cords which is then shaped by a filter imitating the effect of the vocal tract. The output of the vocal cords, $p(t)$ is the input to the filter and is called the excitation signal since it excites the vocal tract. The vocal tract is a Linear Time-Invariant system with impulse response $h(t)$. This is sometimes called the shaping function of speech since it shapes the spectrum of excitation signal. The output of this shaping function is spoken speech $s(t)$ which has to be modeled. Hence,

$$s(t) = p(t) \star h(t)$$

$$S(Z) = P(Z) \cdot H(Z)$$

As the signal to be analyzed is the speech signal $s(t)$, the aim of the source filter modeling is to obtain the two quantities at Right hand side, namely Excitation Component $p(t)$ and Filter or Vocal Tract Shaping function $h(t)$.

2.3 Source-Filter Decomposition

Various Models have been suggested for source-filter modeling like Linear Prediction analysis, Cepstrum Analysis, Line Spectral Frequencies, and Sinusoidal Modeling etc [11]. Here, LP Analysis is chosen for source filter modeling [8]. The reason is this method is well documented in speech literature and computationally efficient than other methods. The drawback of this method is that it does not provide stability check.

2.3.1 Pre-Emphasis and Framing

Prior to the analysis, the speech signal is passed through a pre-emphasis filter in order to reduce the dynamic range of speech spectra. The result of pre-emphasis is shown in Figure 2.3.1

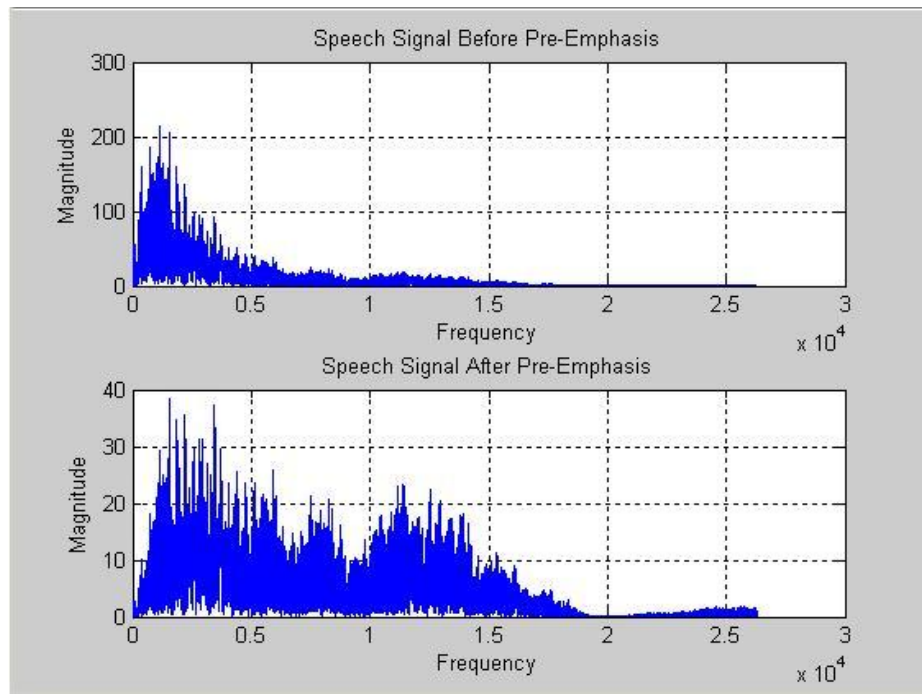


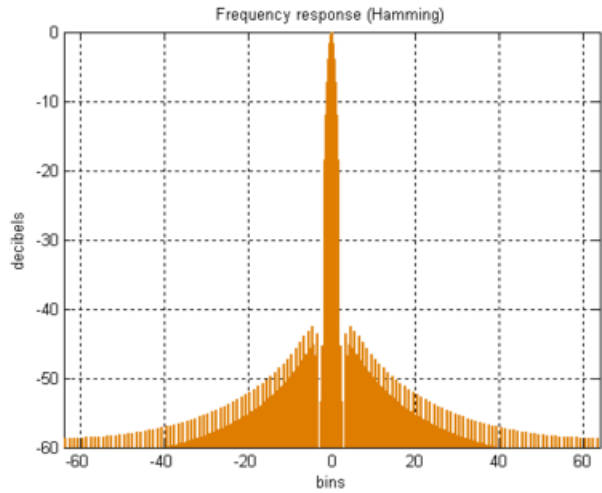
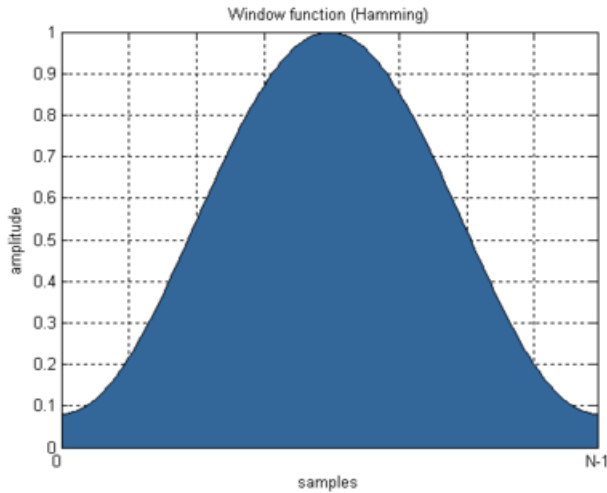
Figure 2.3.1: Spectrum of a Speech Segment prior to Pre-Emphasis and after Pre-Emphasis

The pre-emphasized speech is then segmented into short-term analysis frames using a Hamming window. A common observation is that the human pitch does not go below 50Hz, which corresponds to 20ms duration. So, here frame size of 30ms duration is chosen, to cover at least 2 pitch periods. Hamming window is used due to its tapered frequency response so that it reduces the effect of discontinuities at the beginning and end of each analysis frame. Hamming window is described by the following equation,

$$w(n) = 0.54 - 0.46 \cos\left(2\pi \frac{n}{N-1}\right)$$

$$0 \leq n \leq N-1$$

Hamming window and its frequency response are shown in Figure 2.3.2. An important point to note is that the window affects the temporal gain characteristic of segment, and hence the next window is applied such that it has some amount of overlap with its previous window[10].



Fi

Figure 2.3.2. Hamming Window (left) & Frequency Response (Right)

2.3.2 Modeling Source and Filter using LP Analysis

A generalized observation of human speech production states that a speech waveform is output of a time varying all-pole filter driven by a source component (source can be a periodic signal, noise or mixture of this two). So, transfer function of vocal tract can be approximated as,

$$V(z) = \frac{Gz^{-p/2}}{1 - \sum_{i=1}^p a_i z^{-i}}$$

$$V(z) = \frac{Gz^{-p/2}}{A(z)}$$

$$V(z) \approx \frac{G}{A(z)} (\text{delay} - z^{-p/2} \text{neglected})$$

The task of filter modeling is to approximate a_j coefficients such that the filter frequency response tracks the speech spectrum.

Linear prediction is based on the fact that in slowly varying signals it is possible to predict the future sample based on the values of a few past samples, the number of samples used to predict the next value is called prediction order p . The Linear Prediction Equation is given by,

$$s(n) = Gu'(n) + \sum_{i=1}^p a_i s(n-i)$$

$$\therefore \hat{s}(n) = \sum_{i=1}^p a_i s(n-i)$$

a_i s are called LP Coefficients. The most common choice in optimization of parameters a_i is the root mean square criterion which is also called the autocorrelation criterion. In this method the expected value of the squared error $E[e^2(n)]$ is minimized. This leads to a set of equations in terms of a_i s known as Yule-Walker equations which can be solved using Levinson-Durbin algorithm.

Here, the Prediction Error $e(n)$ is,

$$\therefore e(n) = s(n) - \hat{s}(n)$$

$$\therefore e(n) = s(n) - \sum_{j=1}^p a_j s(n-j)$$

Taking Z-Transform,

$$E(z) = S(z) - \sum_{j=1}^p a_j S(z) z^{-j}$$

$$\therefore E(z) = S(z) \left[1 - \sum_{j=1}^p a_j z^{-j} \right]$$

$$\therefore E(z) = S(z) \cdot A(z)$$

$$\therefore S(z) = \frac{E(z)}{A(z)} = Gu'(z) \cdot \frac{1}{A(z)}$$

1. $\frac{1}{A(Z)}$ Represents Vocal Tract Transfer Function, which is characterized by coefficients a_i .
2. $E(Z)$ Represents the Excitation Component, which can be obtained by inverse-filtering (filtering $S(z)$ by all-zero filter $A(z)$).
- 3.

Figure 2.3.3 shows the result obtained by implementing the vocal tract filter whose coefficients are obtained from a segment of speech by using a LP analysis of prediction order 20.

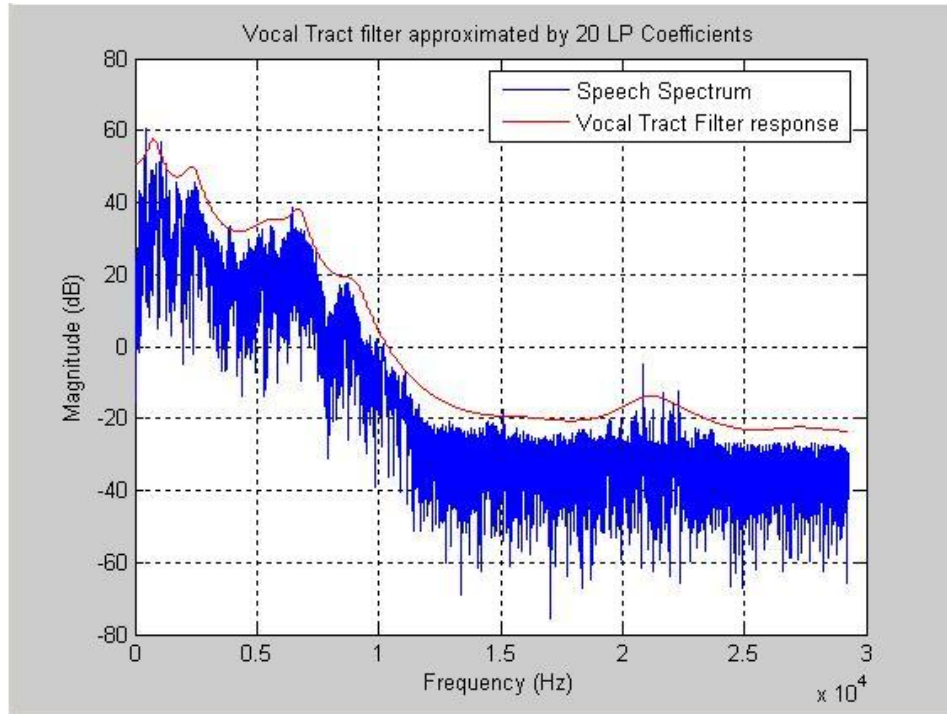


Figure 2.3.3 Response of filter characterized by LP coefficients

Figure 2.3.4 shows the excitation component obtained by inverse filtering. The output is enlarged to convey some important characteristics of residual/excitation signal.

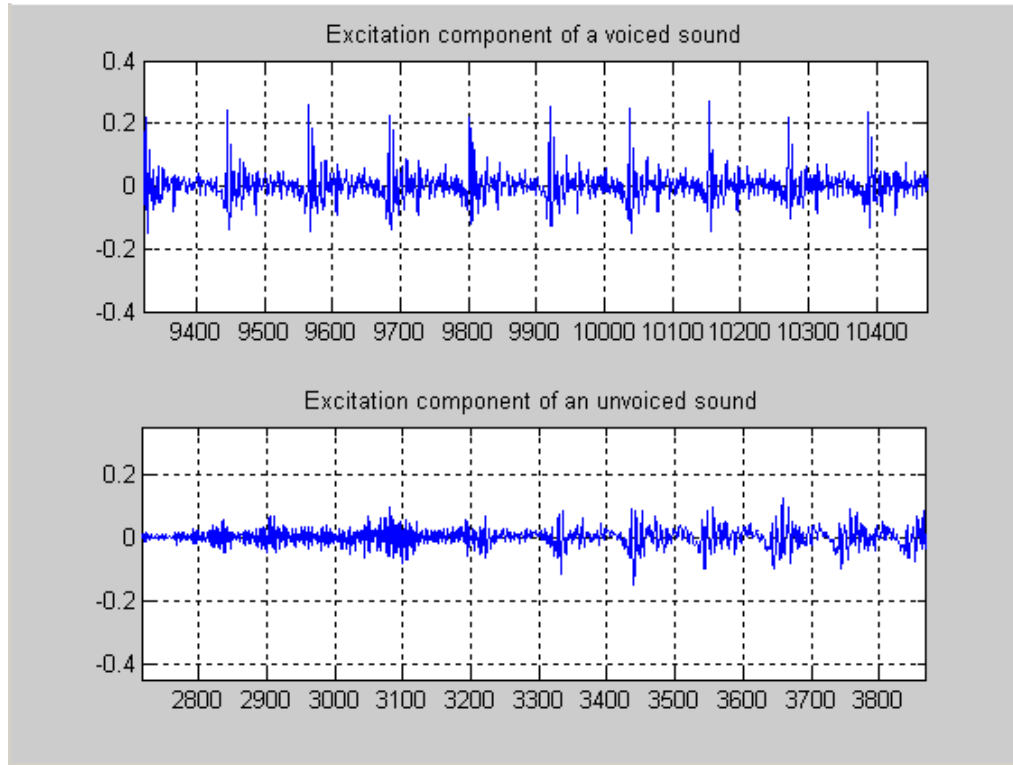


Figure 2.3.4 Excitation Components of voiced and unvoiced segments of a speech

It is apparent that the residual of a voiced component is made up of regularly spaced pulse-train; where as the residual of the unvoiced component is much noise-like. Modeling the excitation component deals with determining the “Spacing” that is the pitch period for voiced components. This is described in detail in chapters 3 and 4.

2.4 Mapping Parameters for Voice Conversion

Once both excitation (pitch also) and filter components are extracted, they should be cross-mapped to achieve voice conversion. Ideally, a robust and sophisticated Voice Conversion System is expected to have a Training phase. During this phase the source speaker mimics the target speaker. These few utterances are used to train the system for source and target voices and then the trained system

controls the mapping or conversion behavior for all unknown and unaligned utterances of source speaker. The training can be HMM, ANN or VQ based [4]. Training is not yet implemented and hence not discussed here. We assume availability of time aligned utterances of source and target and present the very basic idea of voice conversion.

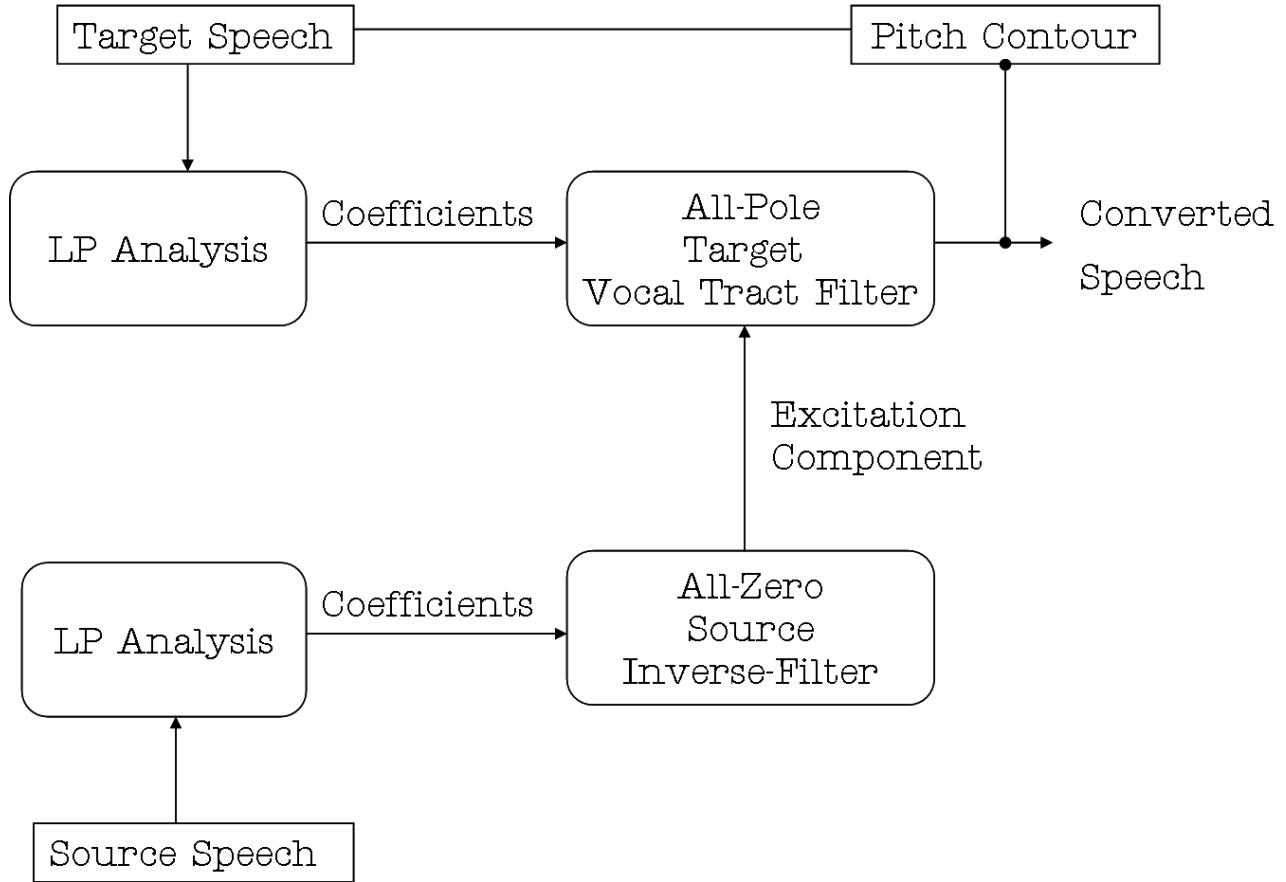


Figure 2.4.1 Block Diagram of Pitch and Vocal Tract Mapping Process

2.4.1 Mapping Vocal Tract Coefficients (Formant Frequencies)

After the analysis phase LP Coefficients for both source and Target have been extracted. These parameters are used to model a Vocal Tract Filter of Target Speech and an Inverse-Filter to extract excitation component of source speech. Now this excitation component is applied to the All Pole Vocal Tract Filter (Shaping Function), which ‘*Shapes*’ the spectrum of source excitation. Thus, by

this filtering operation, spectral shaping of source signal is achieved. The modified spectrum has formants at the target formant frequencies. This is demonstrated in figure 2.4.2.

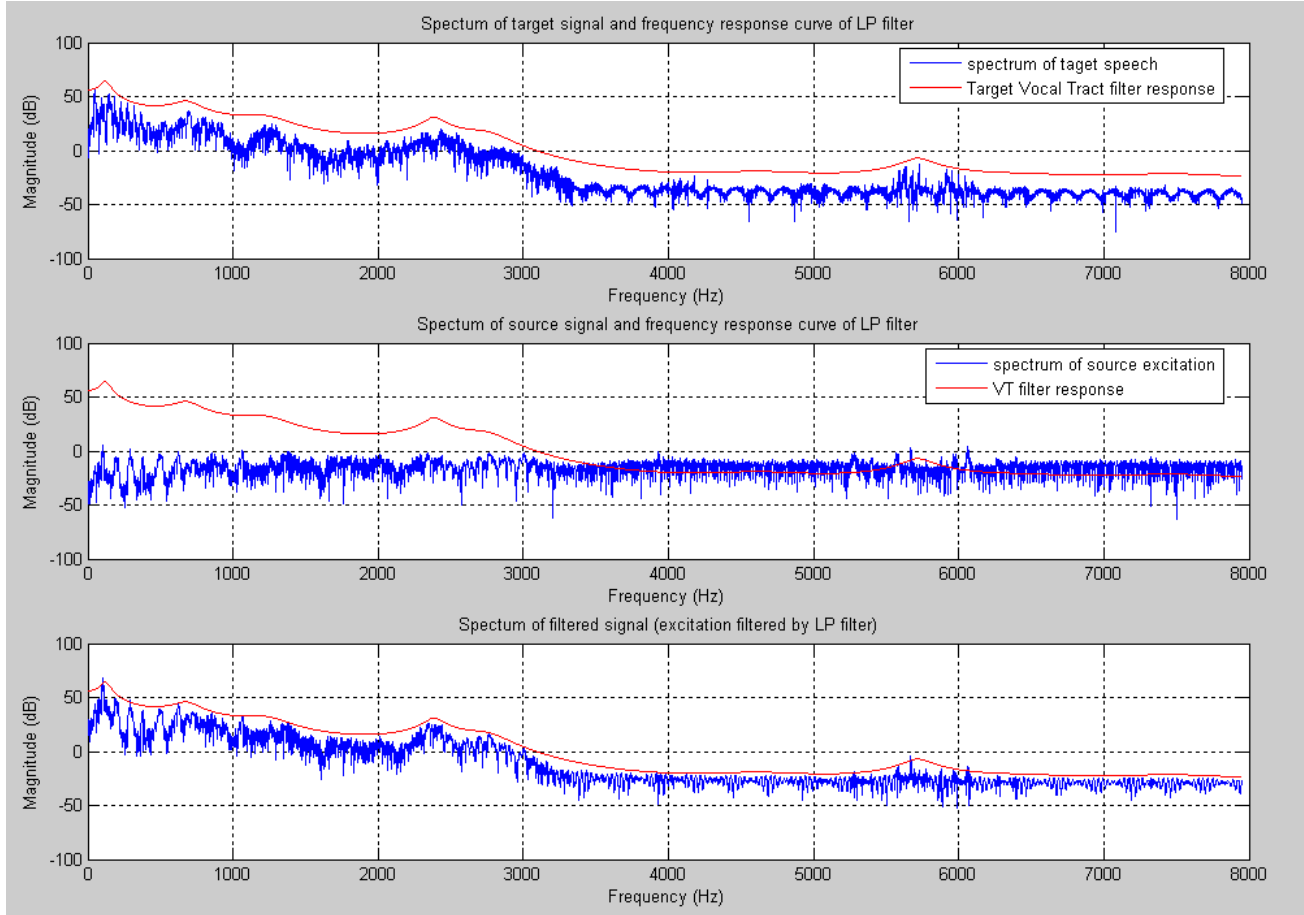


Figure 2.4.2: Result of showing spectra of target, source and filtered signal along with filter response

1. Subplot 1 shows spectrum of the target signal and magnitude response of the filter, modeled by LP Coefficients of Target Speech.
2. Subplot 2 shows spectrum of the source signal and magnitude response of the target Vocal Tract modeled by target LP coefficients.
3. Subplot 3 shows the spectrum of source after filtering process that is the spectrum of filtered signal. It is worth noting that the source spectrum is shaped by the target vocal tract filter and have formants (spectral peaks) at target formant frequencies.

2.4.2 Mapping or Modifying Excitation Component

Modification in excitation component refers to pitch modification. Here, two different approaches are used for voice conversion. Both use the same approach for Vocal Tract Modification, but they differ in their strategy to modify the excitation component (pitch modifications).

The first approach is known as TD-PSOLA (Time Domain Pitch Synchronous Overlap-Add) which modifies the source component so that the pitch of the signal approaches target pitch without any change in the time-scale. This is explained in detail in chapter 3.

The second approach is based on modeling the excitation component based on voiced/unvoiced detection. For each window voicing detection flag is set. For voiced components pitch value is determined and pulse train of determined pitch period is generated, whereas for unvoiced components white Gaussian noise is generated.

The key issue of the approach is finding precise pitch values for voiced windows. Pitch detection has always been a complex issue in speech processing. With many pitch detection algorithms proposed in many years [5], a general observation states they are context specific algorithm and works well on specific content only. The basic time domain methods for pitch detection are zero-crossing rate, autocorrelation method and covariance method. Here, an autocorrelation based approach is used along with voicing detection for pitch computation.

Chapter 3

PSOLA Based Approach

3.1 Aim of the approach

3.2 Time Domain Pitch Synchronous Overlap-Add

3.1 Aim of the approach

The aim of this approach is to modify the source pitch to match the target pitch. This cannot be done by simply increasing the pitch value / decreasing the pitch period as this will lead to compression or expansion of time scale and the speech will no longer remain intelligible. The goal of pitch modification is to modify up or down the pitch of a speech signal without losing its information. If done correctly the new audio signal will be of the same length, sound like the original signal, but at a desired target pitch. Out of various methods suggested over years for pitch modifications such as delay-line modulation, phase vocoders, variable speed relays and various SOLA methods, PSOLA has proven to give the best results. [9] Out of the two variants of PSOLA that is TD-PSOLA and FD-PSOLA, Time Domain Pitch Synchronous Overlap-Add is used because of lesser complexity involved.

3.2 TD-PSOLA

TD-PSOLA stands for Time-Domain Pitch Synchronous Overlap-Add. It is a simple and effective algorithm for both Time and Pitch scale modifications. The idea is to process the speech signal on a short-time basis where the segments are obtained pitch synchronously. These segments are concatenated in an appropriate manner to obtain the desired modifications. The main steps of the algorithm are explained here.

3.2.1 Pre-Processing

PSOLA works by extracting the short term analysis segments at a pitch synchronous rate. So, the start and end instants of pitch periods over the voiced regions are determined by pitch marking. Various algorithms have been suggested for pitch marking. Pitch detection methods are not suitable for this purpose as the exact instants where the pitch period starts and ends are required. The quality of whole modification process depends upon how robust and effective the pitch marking algorithm is. Here, a pitch marking program developed by “Mekhmoukh Abdenour” from MATLAB Central Exchange is used, which has proven to be very robust. Figure 3.2.1 shows the result of pitch marking on a speech file

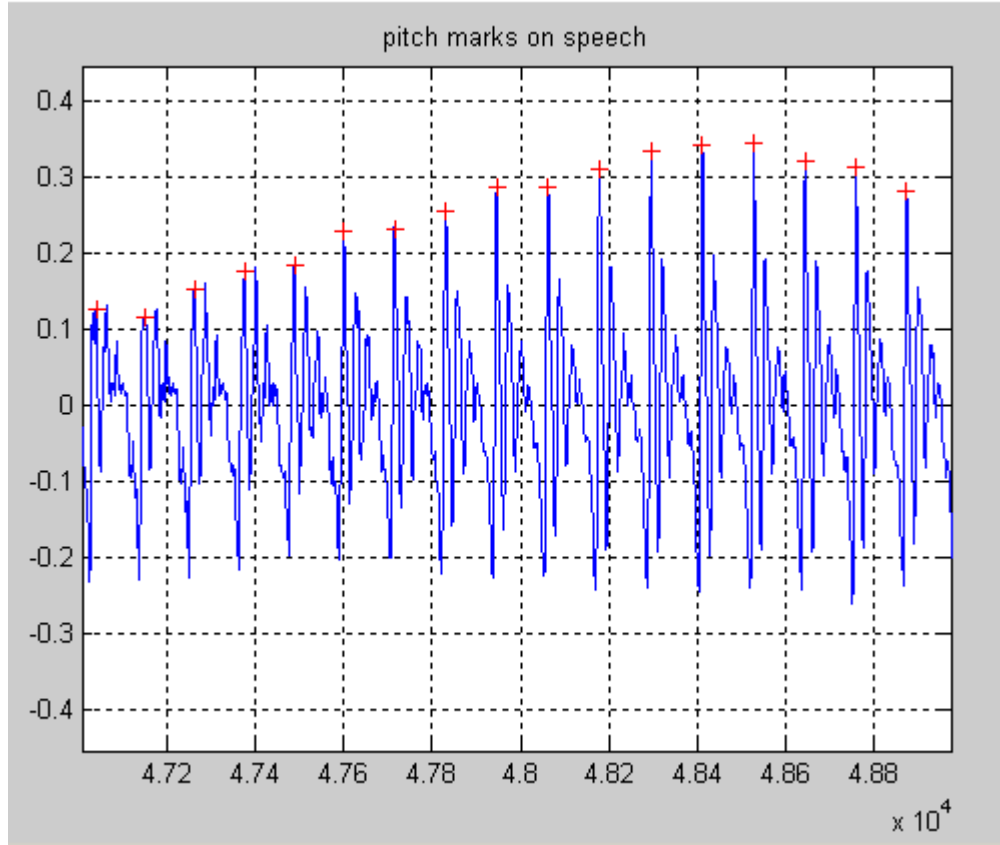


Figure 3.2.1: A pitch marked speech segment

3.2.2 Analysis

First, the input speech signal is divided into short term analysis signals by applying a hamming window with window size being two pitch periods, this is described by the following equation,

$$x_m(n) = h_m(t_m - n)x(n)$$

$$x_m(n) = \text{SHORT TERM SIGNAL (STS)}$$

$$h_m(n) = \text{WINDOW (TYPICALLY HANN / HAMMING)}$$

$$x(n) = \text{SPEECH SIGNAL}$$

$$t_m = \text{PITCH MARKS}$$

3.2.3 Modification and Synthesis

1. Time Scale Modifications

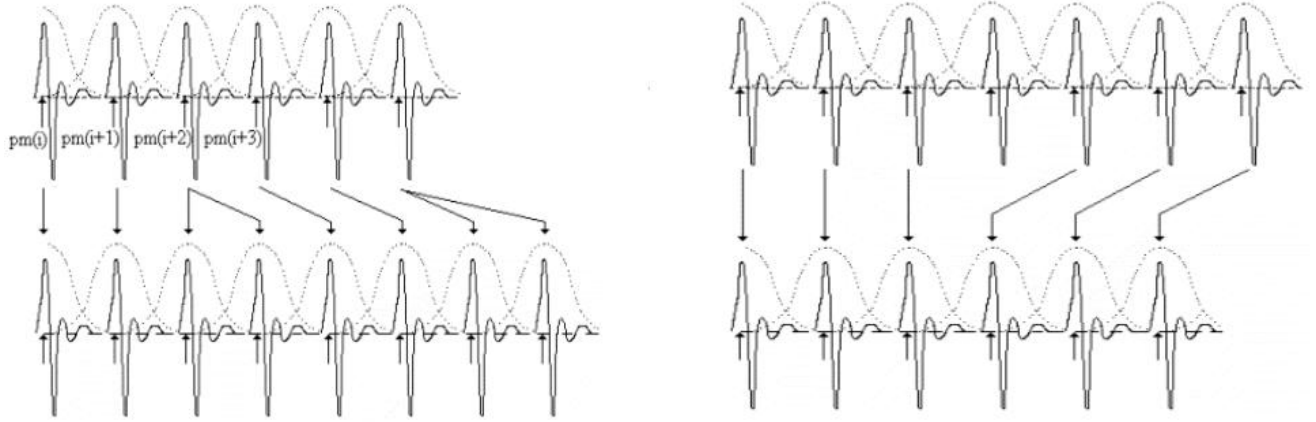


Figure 3.2.2: Time Scale Expansion(left) and Compression(Right)

TD-PSOLA modifies the temporal content by repeating or removing integer number of speech segments. Segment repetition produces a signal that is expanded in the time domain while the output using deletion is a time-compressed version of the original signal as shown in Figure 3.2.2. Repetition/deletion of integer number of frames does not modify the short-time spectral content and distort the relationship between the pitch harmonics.

2. Pitch Scale Modifications

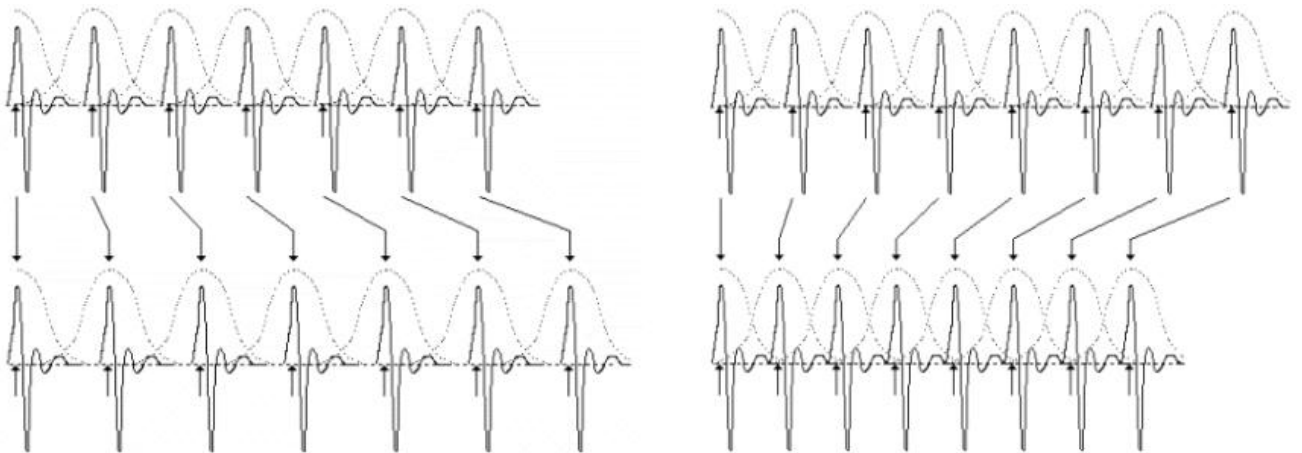


Figure 3.2.3: Pitch scale Compression (Left) and Expansion (Right)

To modify the pitch, TD-PSOLA modifies the amount of overlap between successive pitch-synchronous segments as demonstrated in Figure 3.2.3. It is also clear that pitch scale modification results in the modification of the time-scale. Since this is not desired, compensating time-scale modification must be employed.

3. Synthesis

In the final step, the output signal is constructed using overlap-add method with windowing. All the procedure described above determines the new locations and overlap ratios of the frames. The frames are then concatenated using the overlap ratios obtained in the pitch-scale modification step. The main advantage of TD-PSOLA is its simplicity and yet high efficiency. However, when severe amounts of time and pitch scaling are applied, the output quality degrades. So, it is preferable to do simultaneous modification of both the scales then scaling them one by one.

Chapter 4

Speech Synthesizer Approach

4.1 Aim of the approach

4.2 Pitch Detection using AUTO-CORRELATION

4.3 Excitation Generation and Synthesis

4.4 Results

4.1 Aim of the Approach

This approach is based on modeling the excitation component based on voiced/unvoiced detection. For each window voicing detection flag is set. For voiced components pitch value is determined and pulse train of determined pitch period is generated, where as for unvoiced components white Gaussian noise is generated.

Voicing detection is made using a simple fact that the unvoiced components are much noise-like and have very less energy as compared to the voiced components. So, mean energy for each frame is compared with a threshold value and if the mean value is greater than the threshold then the voicing flag for the frame is set to 'True' otherwise it is set to 'False'. The next task is to determine the pitch for voiced frames. The frame size is kept such that it at least covers two pitch periods. Human speech does not go below 50Hz in general. This corresponds to duration greater than 20ms. So window size of 30ms duration is considered. The next window has 10ms overlap with the previous window for the reason explained in 2.3.1.

4.2 Pitch Detection using Auto-Correlation

Auto-correlation is a method to measure the signals' correlation with its own shifted version. At zero shift, any signal has maximum correlation, as the shift increases correlation tends to zero and for periodic signals it again tends to increase as the shift approaches the period. Due to this effect, autocorrelation is a well-known tool for determining unknown periodicity. Speech signals being Quasi-Periodic in nature, autocorrelation can be used to find out its pitch period

4.2.1 Auto-Correlation based approach

In this method first using the voiced frame of signal, we generate the autocorrelation function $r(s)$ defined as the sum of the point wise absolute difference between the two signals over some interval. Figure 4.2.1 shows how the signals begin to align with each other as the shift amount nears the fundamental period.

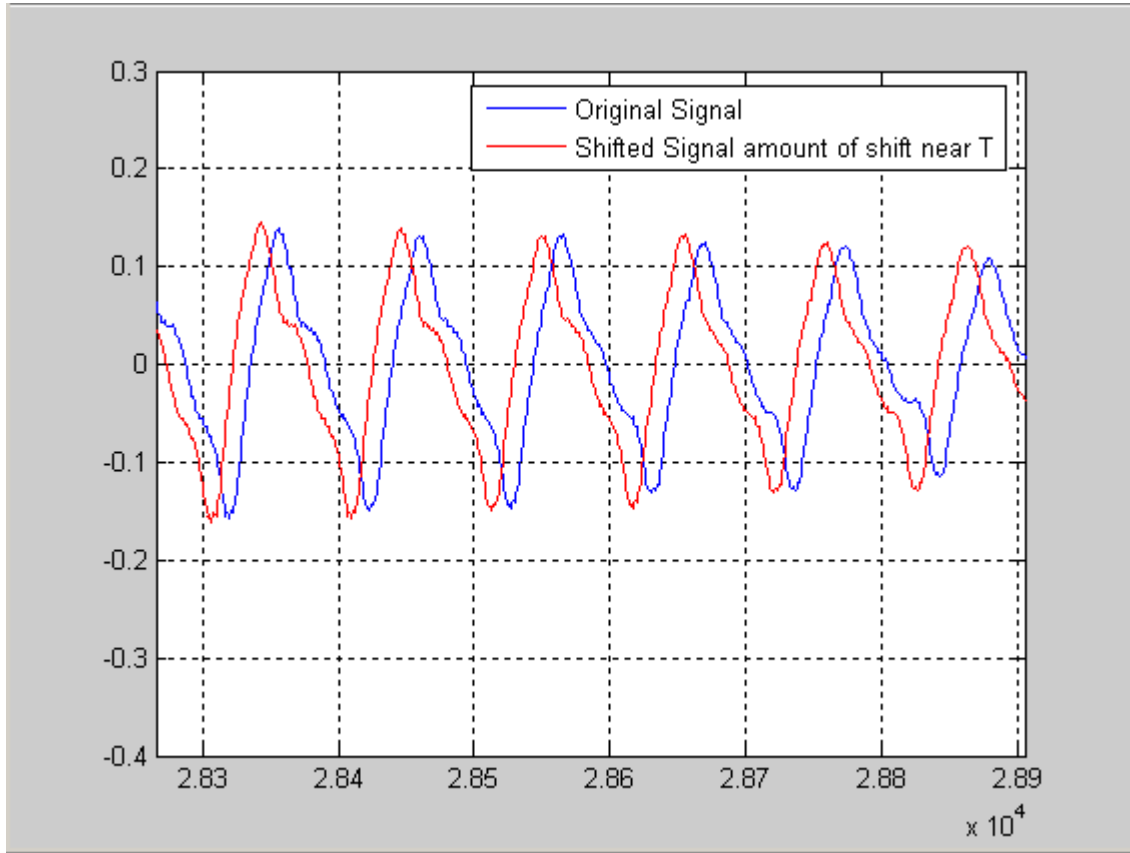


Figure 4.2.1: Representation of Autocorrelation at a particular shift

Intuitively, it should make sense that as the shift value s begins to reach the fundamental period of the signal T , the difference between the shifted signal and the original signal will begin to decrease. Indeed, this can be seen in the Figure 4.2.2 in which the autocorrelation function rapidly approaches zero at the fundamental period.

This value can be detected by differentiating the autocorrelation function and then looking for a change of sign, which yields critical points. Here, look at the direction of the sign change across points (positive difference to negative), to take only the minima. Then search for the first minimum below some threshold, i.e. the minimum corresponding to the smallest s . The location of this minimum gives us the fundamental period of the windowed portion of signal, from which we can easily determine the frequency.

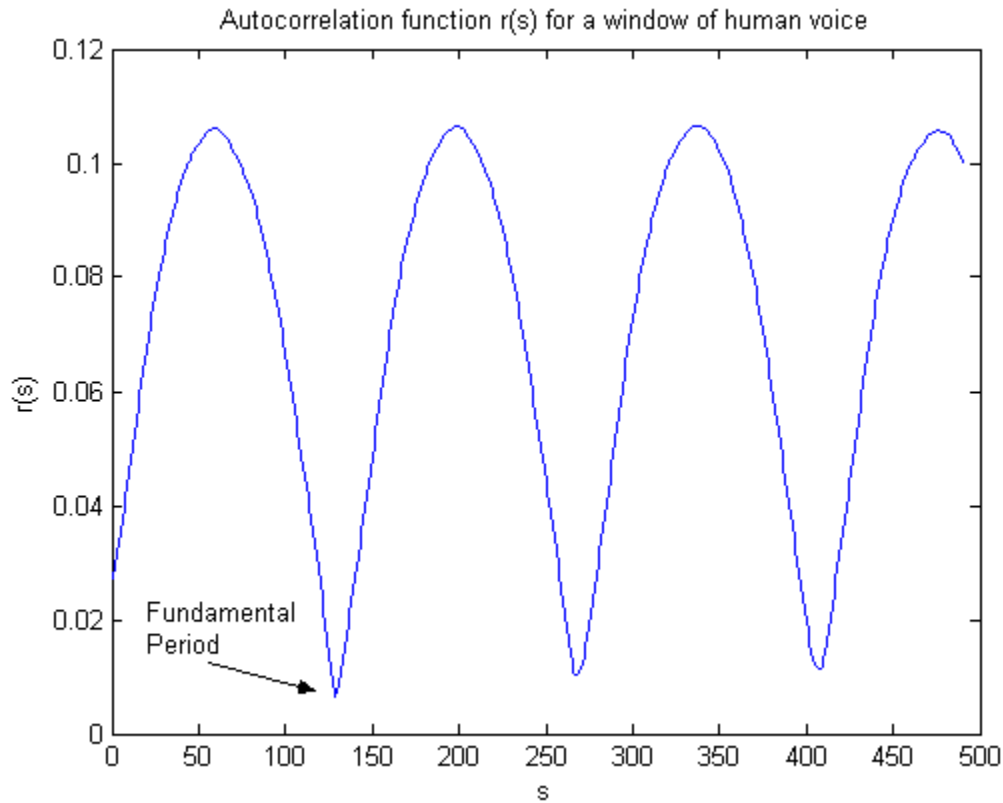


Figure 4.2.2: Autocorrelation function of a speech signal

4.2.2 Fast Auto-Correlation

Clearly, this algorithm requires a great deal of computation. First, it is required to generate the autocorrelation function $r(s)$ for some positive range of s . For each value of s , the total difference between the shifted signals has to be calculated. Range of s is chosen to be 0 to 599, so the same routine is repeated 600 times for each window. Next, this signal has to be differentiated to search the minima. This process is repeated for all the windows.

In effort to improve the efficiency of this algorithm, an alternative called Fast Autocorrelation is used [13], which has yielded speed improvements in excess of 70%.

Here, the nature of the signal is exploited, specifically the fact that if the signal was generated using a high sampling rate and if the windows are narrow enough, it can be assumed that the pitch will not vary drastically from window to window.

1 Improvement 1

Begin calculating the $r(s)$ function using values of s that correspond to areas near the previous minimum. This means that, if the previous window had a fundamental period of 156 samples, begin calculating $r(s)$ for $s = 136$. If a minimal s could not be found in this area, calculate further and further from the previous s until a minimum is encountered

2 Improvement 2

Also, the first minimum (valued below the threshold) is always going to correspond to the fundamental frequency. Thus, the difference equation $dr(s)/ds$ can be calculated as $r(s)$ is generated. Then, as the first minimum below threshold is found, stop calculating altogether and move on to the next window.

If we use only the second improvement, we usually cut down the range of s from 600 points to around 200. If we then couple in the first improvement, we wind up calculating $r(s)$ for only about 20 values of s , which is a savings of $(580) * (1200) = 700000$ calculations per window. When the signal may consist of hundreds of windows, this improvement is substantial indeed.

4.3 Excitation Generation and Synthesis

The key factor driving whole generation process is voiced/unvoiced decision for each window. After the pitch values for all the windows are found the excitation component is generated for each window of size equal to window size. A sinusoidal pulse train of pitch period is used to generate excitation for voiced windows. For unvoiced excitation, a random noise generator produces a uniformly distributed random signal. The amplitude of the generated excitation signal is scaled by gain value and then passed through a filter characterized by LP coefficients of target speech as explained in chapter 2. This process results in output speech windows, which are added with same amount of overlap used at the time of analysis. The whole synthesis process is shown in the figure 4.3.1,

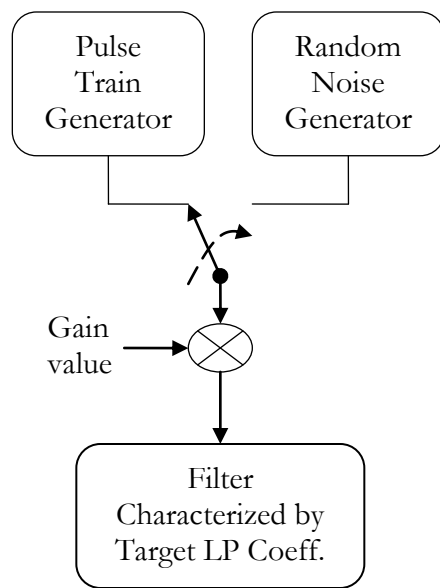


Figure 4.3.1 Synthesis Process controlled by voicing detection

Chapter 5

Voice Conversion Demo

5.1 Demo 1

5.2 Demo 2

5.3 Demo 3

5.1 Demo 1

The first Demo is a GUI based program made in MATLAB 7.0, which gives a basic insight into how the basic parameter modifications enable the voice to be changed effectively, Figure 5.1.1 is a snapshot of the main screen of the GUI.

In this the target voice is not a pre-recorded sound but it is specified by certain speech parameters. The amount of modification or change is specified by the user by changing the slider values. Various other operations are also supported by a set of push buttons whose functional description is given in Table 5.1.1

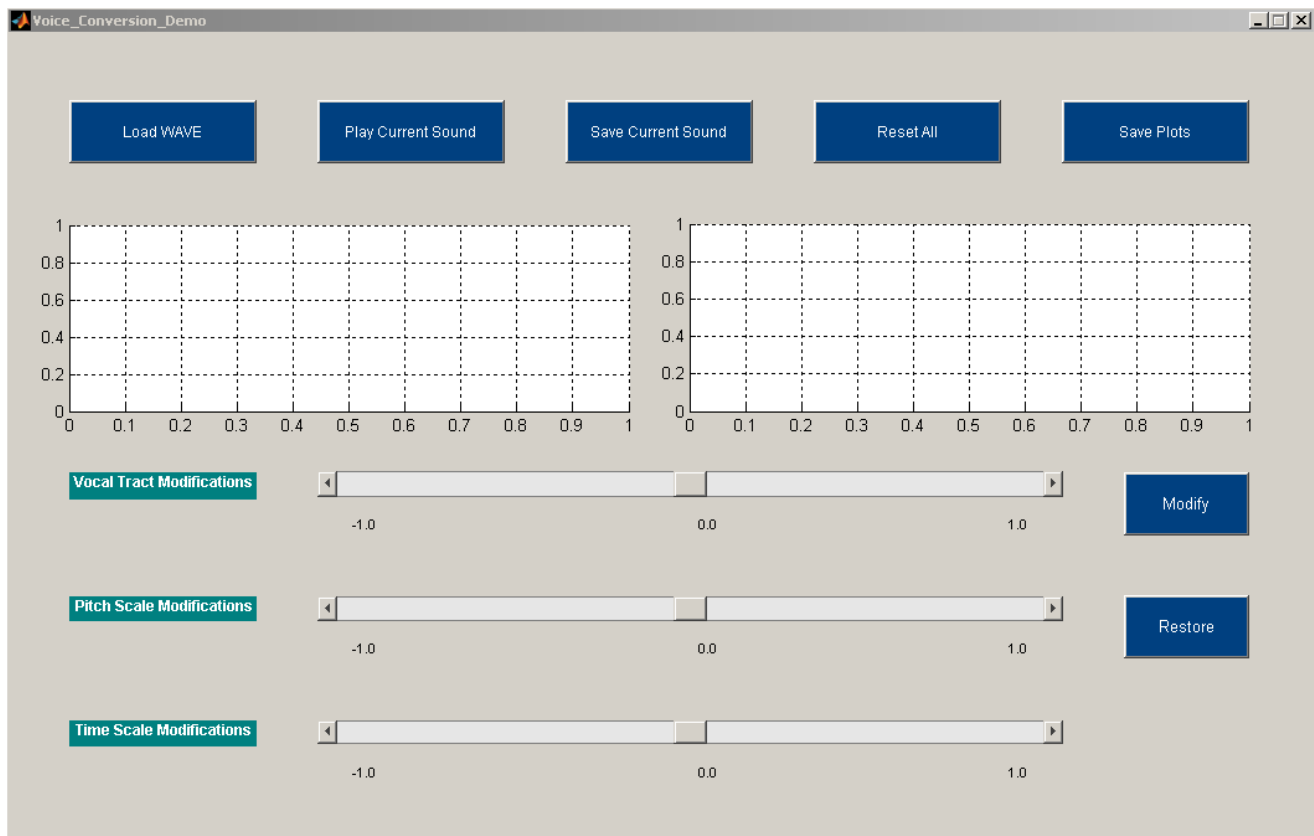


Figure 5.1.1: Screenshot of Demo 1 GUI program

The main components of the GUI are,

Component	Description
Load Wave Push button	When this button is pressed it prompts the user to select a sound file. The sound file has to be in .wav format. The selected sound is graphically plotted in plot area 1 and its spectrum is plotted in plot area 2.
Save Current Sound Push button	This button prompts the user to select a location to store the modified sound file and saves it at the location specified by the user.
Reset All Push Button	This push button reset the GUI to its initial state and clears all the variables.
Save Plots	This push button enables the user to save the plots in supported formats.
Vocal Tract Modification Slider	This slider enables the user to set the amount of change in formant frequencies.
Pitch Scale Modification Slider	This slider enables the user to set the amount of change in pitch scale of current speech. The range -1 to 1 is mapped into the range 2^{-1} to 2^1 , so the new pitch can attain the value from half the current value to the double of current value.
Time Scale Modification Slider	Similar to the above slider, the value set here reflects the time scale modification that is the speed of speech can be changed.
Modify Push Button	After setting all the desired values, pressing this button allows the modifications to take place. It also plots the modified speech and its spectrum in plot area 1 and plot area 2. This modified sound can be listened to or saved by the corresponding Buttons.
Restore Push Button	This Push Button reverses all the modifications and restores the current sound and all the graphs to their original states.

Table 5.1.1 Demo 1 GUI Components and their functions

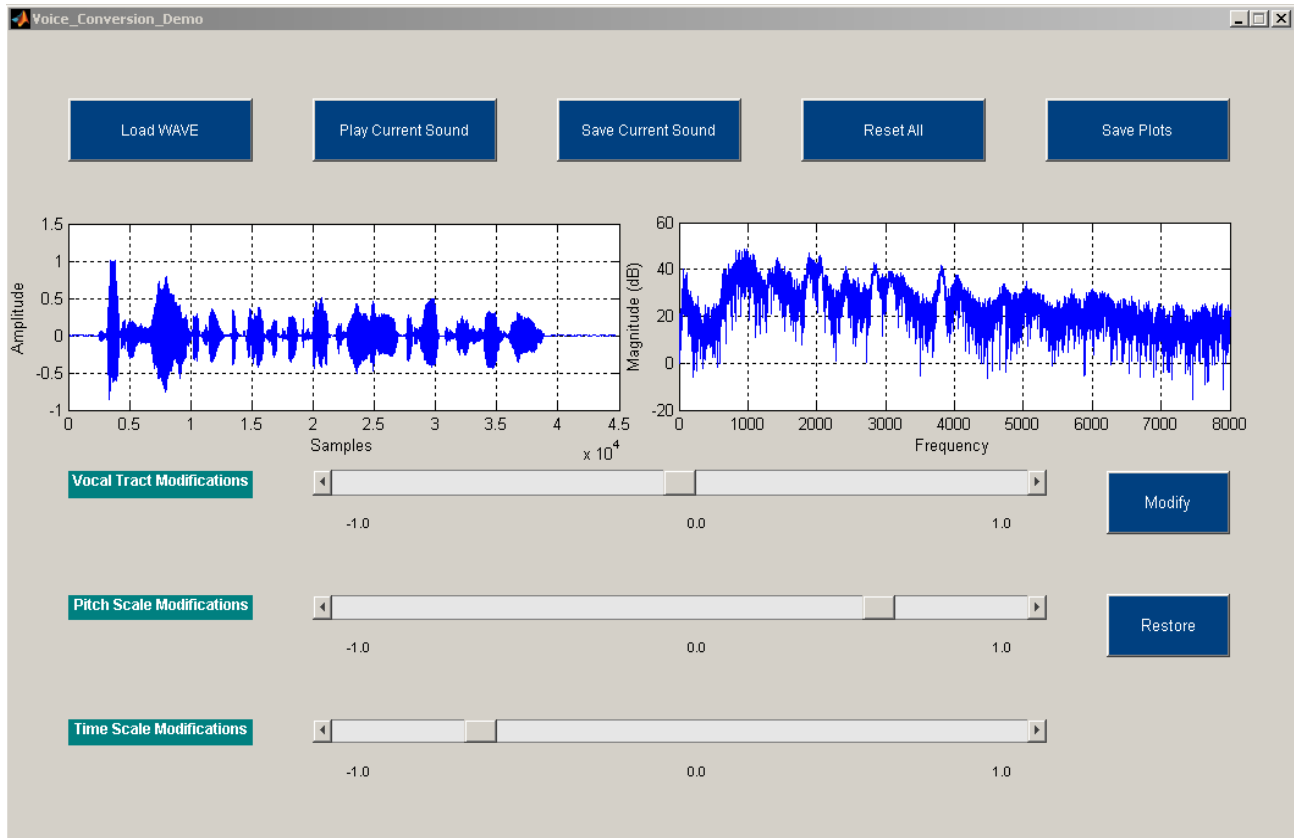


Figure 5.1.2: Screenshot of Demo 1 GUI program in operation

5.2 Demo 2

This is also a MATLAB GUI program, which presents the conversion process described in chapter 3, PSOLA based approach. This GUI actually takes as inputs recorded source and target speeches and extract the parameters and carry out the conversion as suggested.

It is worth mentioning that this is not a generic program, the samples used in the database are specific samples having the same prosody and are recorded in the soundproof recording environment. The quality of whole conversion process is deteriorated.

The screenshot of the demo2 GUI in operation is shown in figure 5.2.1

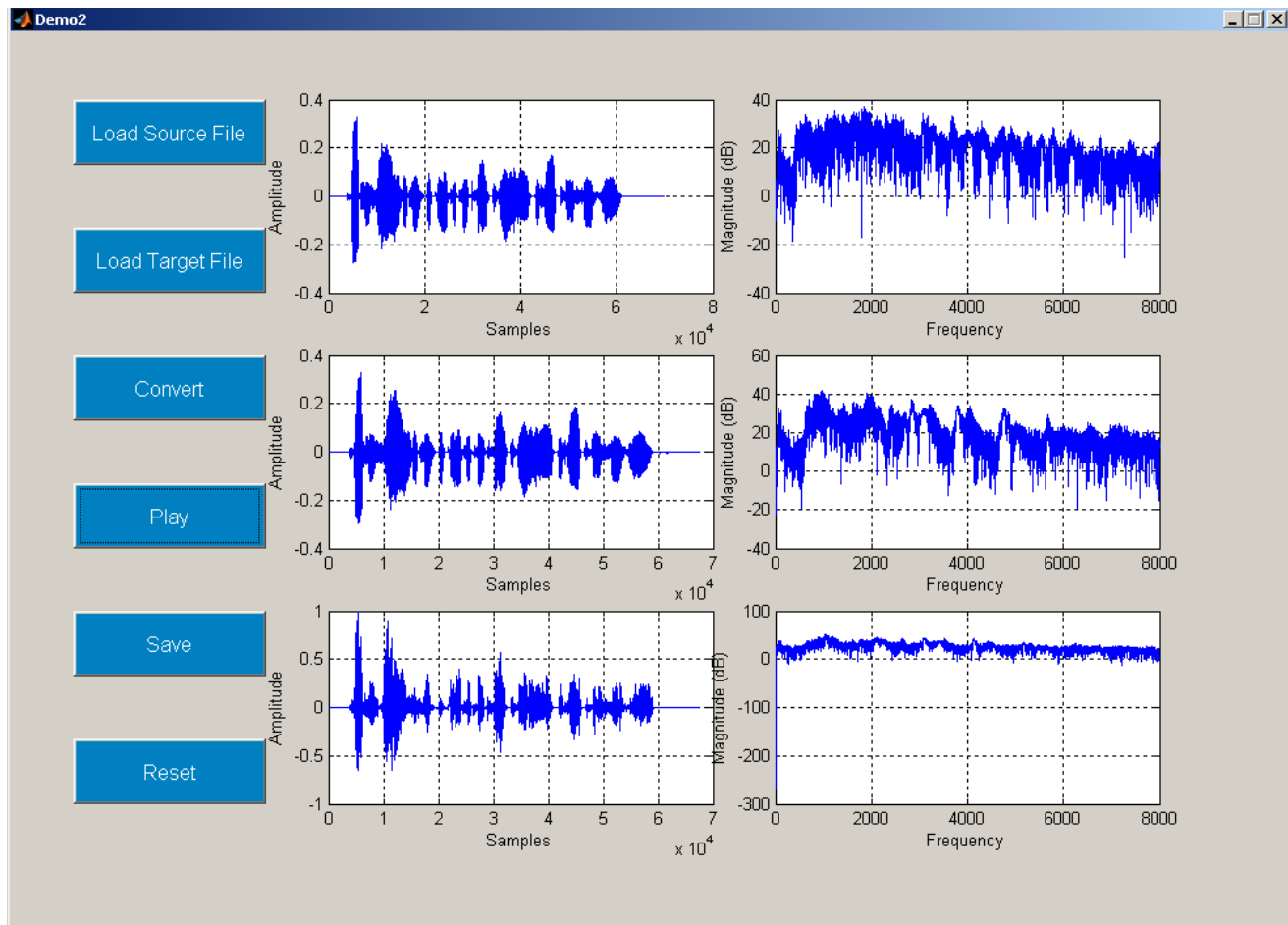


Figure 5.2.1 Screenshot of demo 2 in operation

The functional descriptions of all the pushbuttons are similar to those of Demo 1. Only the Play Pushbutton here pops up a menu as shown in the figure 5.2.2

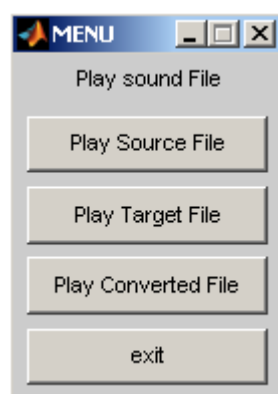


Figure 5.2.2 Screenshot of Play Menu

The functions of all buttons on menu are self explanatory.

5.3 Demo 3

The demo 3 is implementation of speech synthesizer approach using a MATLAB program. The results obtained are not very encouraging, yet they are displayed here for two test files. The converted voice definitely has resemblance with target sound, but it is too “noisy” to be used any further.

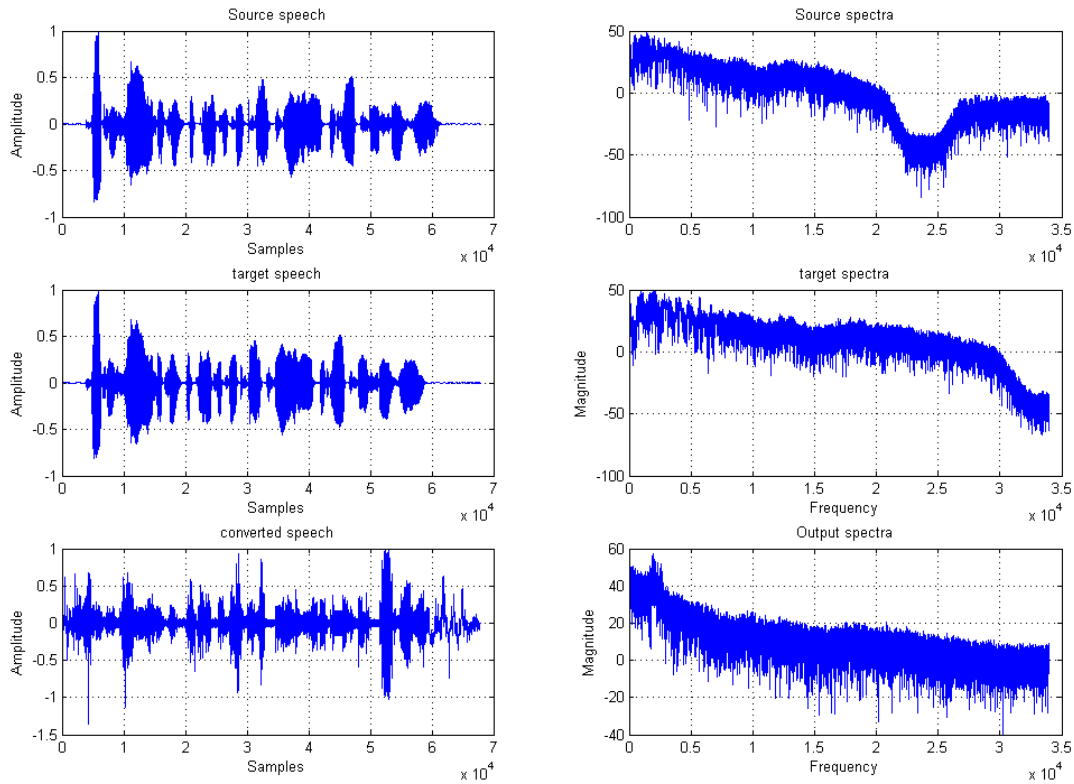


Figure 5.3.1 Result of Speech Synthesizer based approach for two sample speeches

The performance of both of the approaches is compared in chapter 6 “Results and Discussions”

Chapter 6

Results and Discussion

6.1 Quality Measurement of Conversion Process

6.2 Improvement and Future Work

6.1 Quality Measurement of Conversion Process

As seen earlier, two different approaches were used for conversion process. Performance of both the approaches is evaluated in Qualitative terms. Here, a survey was conducted among 8 persons, asking them to give percentage measure for Resemblance of converted speech with target voice and Quality of conversion. Table 6.1.1 shows the results obtained for PSOLA based Approach and Table 6.1.2 shows the results obtained for Speech Synthesizer based Approach. In the tables, R represents the Resemblance factor and Q represents the Quality factor

	Target											
	Female1		Female2		Female3		Male1		Male2		Male3	
Source	R	Q	R	Q	R	Q	R	Q	R	Q	R	Q
Female1	-	-	94%	90%	85%	87%	87%	82%	72%	87%	96%	81%
Female2	93%	89%	-	-	87%	93%	91%	95%	91%	89%	95%	90%
Female3	86%	94%	89%	92%	-	-	92%	90%	85%	87%	91%	92%
Male1	95%	93%	93%	94%	85%	90%	-	-	91%	87%	82%	91%
Male2	87%	94%	92%	91%	89%	87%	85%	85%	-	-	91%	80%
Male3	93%	85%	87%	90%	87%	86%	90%	82%	85%	87%	-	-

Table 6.1.1 Performance Measure for PSOLA based Approach

	Target											
	Female1		Female2		Female3		Male1		Male2		Male3	
Source	R	Q	R	Q	R	Q	R	Q	R	Q	R	Q
Female1	-	-	35%	19%	56%	16%	72%	14%	60%	20%	50%	25%
Female2	55%	15%	-	-	51%	18%	67%	21%	63%	21%	52%	28%
Female3	40%	18%	51%	14%	-	-	56%	18%	56%	16%	65%	14%
Male1	66%	24%	37%	21%	41%	14%	-	-	44%	19%	55%	21%
Male2	62%	21%	44%	18%	62%	21%	53%	21%	-	-	61%	22%
Male3	52%	12%	57%	24%	67%	22%	51%	16%	46%	15%	-	-

Table 6.1.2 Performance Measure for Speech Synthesizer based Approach

6.2 Improvements and Future Work

The results obtained suggests that PSOLA based approach outperforms Synthesizer based approach; a noise cancellation strategy can further improve the Quality. Also, the system can be made more efficient by including prosodic modifications and time alignment. The system becomes very robust, efficient and generic in nature if training is implemented. Being a new field in the speech technology, it has a lot of scope for implementing improvements in the system.

CONCLUSION

Here, two different approaches are developed to achieve voice conversion. Also the MATLAB Demos developed here give a primitive insight into the field of voice conversion. The system discussed here processes on pre-time aligned speech samples. Numerous efforts are required to implement many modifications to make the present system more robust, efficient and generic. One of such modification is training. An ideal voice conversion system should include a training phase so that the system can be trained with target speech and can be used to convert any arbitrary speech uttered by source speaker which could not be done because of restriction of time and inadequate knowledge at the current stage. High quality transformations can be obtained with more complex and computationally expensive techniques. Also real-time voice conversion can be achieved with powerful Digital Signal Processors or similar Hardware. Voice conversion is yet an unexplored field in speech technology and expects a lot of contribution from speech researchers in the future years.

REFERENCES

- [1] Oytun Türk and Levent M. Arslan, “Robust processing techniques for voice conversion”, *Computer Speech and Language*, vol.20, 2006.
- [2] Masanobu ABE, Kiyohiro SHIKANO, Hisao KUWABARA, “Cross language voice conversion”, *IEEE Transactions on Acoustics, Speech, And Signal Processing*, 1990.
- [3] M.M.Hasan, A.M.Nasr and S.Sultana, “An approach to voice conversion using feature statistical mapping”, *Applied Acoustics*, vol. 66, 2006.
- [4] Levent Arslan, “Speaker Transformation Algorithm using Segmental Codebooks (STASC)”, *Speech Communication*, vol.28, 1999.
- [5] Lawrence R. Rabiner, Michael J. Cheng, Aaron E. Rosenberg And Carol A. McGonegal, “A Comparative Performance Study of Several Pitch Detection Algorithms”, *IEEE Transactions On Acoustics, Speech, And Signal Processing*, vol.assp-24, no. 5, 1976.
- [6] David Sunderman, “Text-Independent Voice Conversion”, PHD Thesis, Busim University, Munich, Germany
- [7] Hui Ye and Steve Young, “Perceptually weighted linear transformation for voice morphing”, Thesis, Cambridge University
- [8] Oytun Türk, “New Methods for voice conversion”, MS Thesis, Bogazaci University, Turkey
- [9] Joshua Patton, Pitch Synchronous Overlap-Add, Project Report, University of Victoria, BC, Canada
- [10] L.R.Rabiner and R.W.Schafer, “Digital processing of speech signal” Pearson education, 1992.
- [11] Ben gold and Nelson Morgan, “Speech and audio signal processing”. Wiley India, 2000.
- [12] Don Johnson, “Modeling the Speech Signal”, Unpublished.
- [13] Gareth Middleton, “Pitch Detection Algorithms”, Unpublished.

ABBREVIATIONS

TTS.....	Text-to-Speech
HMM.....	Hidden Markov Model
ANN.....	Artificial Neural Networks
VQ.....	Vector Quantization
LP.....	Linear Prediction
PSOLA.....	Pitch Synchronous Overlap-Add
LSF.....	Line Spectral Frequencies
GUI.....	Graphical User Interface