Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Acoustics 66 (2005) 513–532
applied acoustics
www.elsevier.com/locate/apacoust

An approach to voice conversion using feature statistical mapping

M.M. Hasan *, A.M. Nasr, S. Sultana

Department of Mathematics, Faculty of Science, University Brunei Darussalam, Gadong, BE 1410, Brunei Darussalam

Received 16 June 2004; received in revised form 20 September 2004; accepted 20 September 2004
Available online 13 November 2004

Abstract

The voice conversion (VC) technique recently has emerged as a new branch of speech synthesis dealing with speaker identity. In this work, a linear prediction (LP) analysis is carried out on speech signals to obtain acoustical parameters related to speaker identity – the speech fundamental frequency, or pitch, voicing decision, signal energy, and vocal tract parameters. Once these parameters are established for two different speakers designated as *source* and *target* speakers, statistical mapping functions can then be applied to modify the established parameters. The mapping functions are derived from these parameters in such a way that the source parameters resemble those of the target. Finally, the modified parameters are used to produce the new speech signal. To illustrate the feasibility of the proposed approach, a simple to use voice conversion software has been developed. This VC technique has shown satisfactory results. The synthesized speech signal virtually matching that of the target speaker.

Keywords: Voice conversion; Linear prediction; Pitch contour modification; Speech synthesis

* Corresponding author. Fax: +673 2 461502.
E-mail address: mdmahmud@fos.ubd.edu.bn (M.M. Hasan).

0003-682X/$ - see front matter
doi:10.1016/j.apacoust.2004.09.005

1. Introduction

A voice conversion system works by transforming acoustic speech parameters relevant to a particular speaker while leaving speech message content intact. This task can be done by converting extracted speech parameters of one speaker (the source speaker) to those parameters of another speaker (the target speaker), as shown in Fig. 1.

Of all the acoustic parameters related to a speaker's individuality, pitch and formant frequencies are the two most important, ones consequently, any attempt at VC is usually done through the modification of these unique properties of the speech signal. However, there are few effective voice conversion systems. This has left scope for researchers to study and develop new techniques to solve this problem.

Over the years researchers have developed a variety of techniques and algorithms dealing with speech analysis and synthesis. Examples include digital filtering techniques, linear prediction (LP) analysis, short time Fourier transform (STFT), cepstral analysis, pitch determination algorithms (PDA), hidden Markov models (HMM), dynamic time warping (DTW), etc. Many speech parameters have been proven to be related to speaker identity. Examples include the speech fundamental frequency or pitch, formant frequencies and bandwidths, prosody and many more. However, the most important feature of speech signal is the pitch. Consequently, any attempt at VC is usually done through the modification of this unique property of the speech signal [1]. The speech determination algorithms (PDA) can be divided into three components; the generation of the excitation signal, the modulation of this signal by the vocal tract, and the radiation of the final speech signal (Fig. 2).

The excitation signal is generated when the airflow from the lungs, the main energy source, is forced through the larynx to the main cavities of the vocal tract. As the excitation signal moves through the vocal tract, its spectrum is shaped by the resonance and anti-resonance imposed by the physical shape of the tract. The signal so produced is then radiated from the oral and nasal cavities through the mouth and nose, respectively.

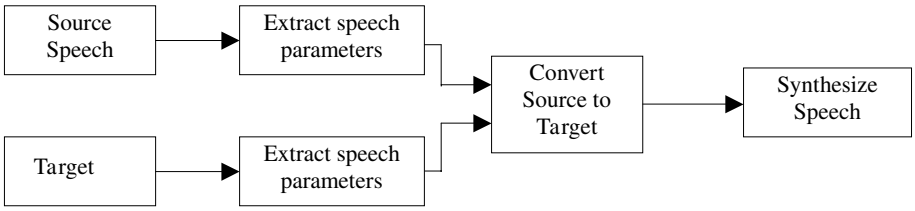

Fig. 1. Basic scheme of voice conversion.

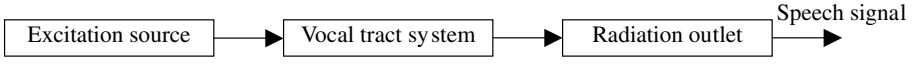

Fig. 2. Main components of speech production.

2. Acoustic features related to speaker identity

Speaker identity, also known as speaker individuality, is the property of speech that allows one speaker to be distinguished from another. Many factors contribute to voice individuality, which can be divided into two main types: static and dynamic features [2]. The static features are determined by the physiological and anatomical properties of the speech organs, such as the overall dimension of the vocal tract, the relative proportions between the various cavities in the tract, and the properties of the vocal cords. These features are the main contributors to the timbre of the voice or vocal quality [3]. Static features also can be measured more reliably than dynamic features, since the speaker has relatively little control over them. Dynamic features, also known as prosody or speaking style, convey information about the long-term variations of pitch, intensity, and timing. Dynamic features are currently difficult to measure, model and manipulate in speech synthesis [4]. For this reason static features are considered to be more useful for voice conversion applications.

2.1. Speech analysis and feature extraction techniques

Speech analysis and synthesis is the technology of efficiently extracting important features from speech signals and precisely reproducing original or modified speech sounds using these features. To perform short time Fourier transform (STFT) analysis, the speech signal is multiplied by a suitable window function (such as triangular, Hamming, Hanning, etc.) and the discrete Fourier transform (DFT) is computed using a fast Fourier transform (FFT) algorithm [5]. The STFT analysis technique has been used for many speech processing applications such as channel coding, transform coding, speech enhancement, short time Fourier synthesis, and spectrogram displays. However, this technique is limited as it can only compute the speech spectrum which gives mixed information of both the source and filter characteristics. It cannot estimate these characteristics separately.

Linear prediction (LP) analysis of speech is one of the most powerful analysis techniques in the field of speech signal processing. It is a highly efficient and computationally fast technique. LP has become the predominant technique for estimating the basic speech parameters such as pitch and vocal tract spectral parameters [6]. The quality of the synthesized speech is greatly influenced by quality of the estimation of pitch [7].

2.2. Voice conversion approaches

A study of voice conversion based on LP analysis synthesis, with the addition of a glottal source waveform model was proposed by Childers et al. [8]. Various acoustic parameters were obtained from identical sentences uttered by both the source and target speaker. These included vocal tract length factors, formant bandwidths, spectral shape, energy contours, pitch contours and glottal information, using both the speech signal and a signal from an electroglottograph (EGG). This second channel provides information on the glottal characteristics, such as instants of Glottal

closure identification (GCI), which provide reliable pitch and glottal waveform estimates. The short-term speech analysis was done using a fixed frame size and frame rate, but with a scalable window function to realize an effective frame size, which was adapted according to the type of speech encountered in each frame (voiced or unvoiced). This allows transient (unvoiced) sounds to be analyzed with short frames, while the voiced regions are handled by longer frames for improved frequency resolution.

A pitch synchronous analysis-synthesis system capable of independently modifying pitch, formant frequencies and formant bandwidths was developed by Kuwabara and Takagi [1]. The pitch period is extracted by estimating the instant of glottal closure (GCI). Although this determination of pitch is not very precise, it does preserve the pitch fluctuations and provides the best instance to estimate the vocal tract spectral envelope. Glottal excited linear prediction (GELP) capable of voice conversion was done [9]. Because acoustic parameter interpolation and extrapolation are especially important in non- parametric voice conversion, to smooth the feature mapping, LP parameters are generally not used directly. This is because it is not possible to guarantee stability when LP parameters are linearly interpolated. A set of derived spectral parameters, such as reflection coefficients (RC), log area ratios (LAR), line spectrum pairs (LSP) and cepstral parameters, are preferred because of their good interpolation properties. It is usually also necessary to time-align the source and target speaker data before spectral relationships can be developed using algorithms such as dynamic time warping (DTW).

Voice conversion could provide a simple alternative to the above approaches by creating entirely new voices with a fraction of the effort and the computer storage space required [10]. Other possible applications are in the entertainment industry. VC technology could be used to dub movies more effectively by allowing the dubbing actor to speak with the voice of the original actor, but in a different language.

3. Proposed VC model

The proposed framework of speech signal processing is based on the LP analysis technique. This approach is preferred over non-parametric methods for the following reasons:

- LP analysis is well documented in the speech processing literature, which provides the basic knowledge and understanding needed to carry out this task.
- It is simple to implement digitally as a set of numerical algorithms.
- It is a highly efficient and computationally fast technique.
- LP relies on a small amount of speech data compared to non-parametric approaches.
- It provides a convenient way to modify the acoustic features individually.

The main speech processing approach within this framework is illustrated in Fig. 3. It involves the LP analysis of both source and target speech signals, in order

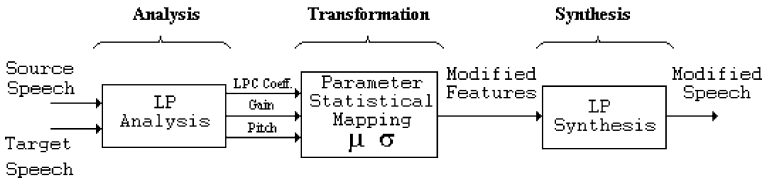

Fig. 3. Voice conversion framework.

to extract the acoustic features which need to be transformed. The extracted features are then statistically mapped to approximate those of the target speaker. Finally, the transformed features are used to synthesize the new speech signal.

Although speech is a continuously time-varying and almost random waveform, the underlying assumption in most speech processing techniques is that the properties of speech signal change relatively slowly with time. This assumption leads to the analysis of speech over short intervals of about 20–30 ms.

Fig. 4 shows the analysis frames implemented in this research. A Frame size of 240 samples is used with consecutive frames spaced 160 samples apart giving an overlap of 80 samples. Since the speech signals were sampled at 8000 samples per second, this yields frames of 30 ms duration with a frame overlap of 10 ms.

3.1. Computation of LP coefficients

Fig. 5 shows the block diagram of the LP analysis algorithms. In this section, the speech signal is first passed through a pre-emphasis filter in order to reduce the

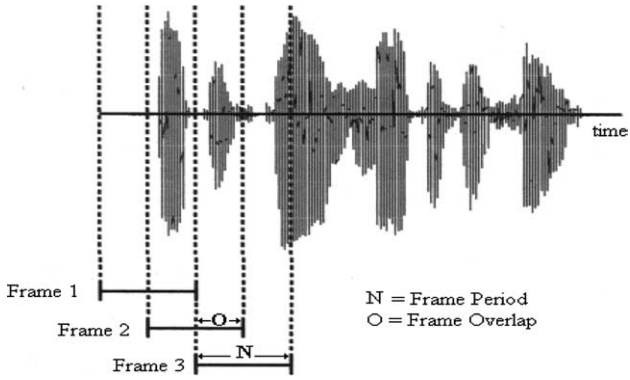

Fig. 4. Analysis frames.

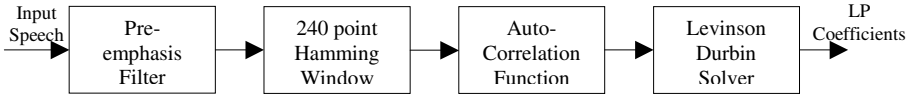

Fig. 5. Block diagram of LP coefficient computation.

dynamic range of speech spectra. The pre-emphasized speech is then segmented into analysis frames of short duration using 240 point Hamming window. After the window is applied, auto-correlation analysis is performed on these finite-length frames. The output of the auto-correlation analysis is a set of equations that can be solved recursively using the Levinson–Durbin algorithm.

3.2. Pre-emphasis filtering

The speech signal normally experiences some spectral roll-off of about 6-dB per octave. This means that the amplitude is halved for each doubling of frequency. This phenomenon occurs due to the radiation effects of the sound from the mouth. As a result, the majority of the spectral energy is concentrated in the lower frequencies, which results in an inaccurate estimation of the higher formants. However, the information in the high frequencies is just as important in understanding the speech as the low frequencies. To reduce this effect, the speech signal is filtered prior to LP analysis. This is done with a first order finite impulse response (FIR) filter, called the pre-emphasis filter. The filter has the form:

$$H_{\text{pre}}(z) = 1 - \lambda z^{-1}, \tag{1}$$

where $H_{\text{pre}}(z)$ is a mild highpass filter with a single zero at λ, and λ is a constant that controls the degree of pre-emphasis.

The value of λ is generally in the range $0.9 \leqslant \lambda \leqslant 1.0$, although the precise value does not affect the performance of the analysis. Fig. 6 shows the frequency response of the pre-emphasis filter for different values of λ.

3.3. Windowing

In the auto-correlation analysis of speech, a moving window is applied to divide the speech signal into frames of finite duration. The window function, $w(n)$, determines the portion of the speech signal that is to be analyzed by zeroing out the signal

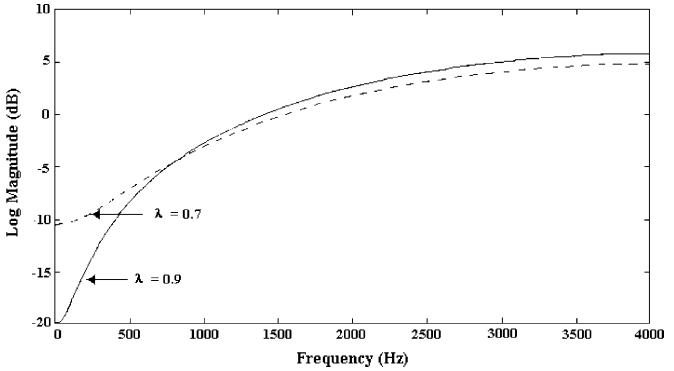

Fig. 6. Frequency response of the pre-emphasis filter.

outside the region of interest. A Hamming window is used in this work due to its tapered frequency response as shown in Fig. 7. This window has the effect of softening the signal discontinuities at the beginning and end of each analysis frame. The Hamming window function is given by

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N-1}\right) 0 \leqslant n \leqslant N-1, \tag{2}$$

where N is the window duration.

Window length is one of the important considerations in implementation of LP analysis. Basically, long windows give good frequency resolution while short windows give good time resolution. Therefore, a compromise is made by setting a fixed window length of 15–30 ms. In practice, the length of the window is usually chosen to cover several pitch periods for voiced speech segments.

3.4. Auto-correlation function (ACF)

After the Hamming window function is applied to the input speech signal, a finite interval signal is obtained. This signal is assumed to have zero values outside the window interval N. The finite interval speech signal is given by

$$s_n(m) = s(m+n)w(m), \tag{3}$$

where $w(m)$ is a Hamming window function of length N and $0 \leqslant m \leqslant N-1$.

Recall the following equations

$$\phi(i,k) = E\{s(n-i)s(n-k)\}, \tag{4}$$

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0), \quad i = 1,\ldots,p. \tag{5}$$

Introducing the short-time signal as given by Eq. (3), (4) can be re-written as

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k), \quad 1 \leqslant i \leqslant p,\ 0 \leqslant k \leqslant p. \tag{6}$$

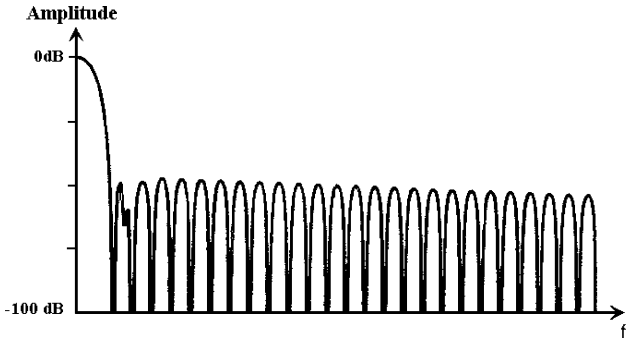

Fig. 7. Frequency response of the Hamming window.

Rearranging Eq. (6) we obtain the following

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad 1 \leqslant i \leqslant p,\ 0 \leqslant k \leqslant p. \tag{7}$$

Eq. (7) is the short-time auto-correlation function that can be expressed as

$$\phi_n(i,k) = R_n(i-k), \tag{8}$$

where

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k). \tag{9}$$

Therefore, Eq. (5) can be written as

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i), \quad 1 \leqslant i \leqslant p \tag{10}$$

and the predictor error is given by

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k). \tag{11}$$

This is a set of p equations that can be expressed in matrix form as

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ R_n(p) \end{bmatrix}. \tag{12}$$

The matrix given by Eq. (12) is a $p \times p$ Toeplitz matrix, which means that it is symmetrical and all the elements along a given diagonal are equal. This matrix can be solved for the predictor coefficients α_k using Levinson–Durbin algorithm.

3.5. Levinson–Durbin algorithm

The Levinson–Durbin algorithm takes advantage of the Toeplitz structure of the auto-correlation matrix. It computes the predictor coefficients in a recursive process. The following equations give the details of this process:

$$E^{(0)} = R(0), \tag{13}$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right] / E^{(i-1)}, \quad 1 \leqslant i \leqslant p, \tag{14}$$

$$\alpha_i^{(i)} = k_i, \tag{15}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \quad 1 \leqslant j \leqslant i-1, \tag{16}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}, \tag{17}$$

This process is repeated for all values of $i = 1,2,\ldots,p$ and the result is given by

$$\alpha_j = \alpha_j^p, \quad 1 \leqslant j \leqslant p. \tag{18}$$

Eq. (18) gives the predictor coefficients of the LP analysis. These coefficients are statistically modified and then used to build the speech synthesis filter.

3.6. Gain computation

The gain parameter, G, in LP analysis is used to produce a synthetic speech signal that has the same energy as the original speech signal. This can be achieved by matching the energy of the LP filter output to the energy of original signal.

$$s(n) = Gu(n) + \sum_{k=1}^{p} a_k s(n-k), \tag{19}$$

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k). \tag{20}$$

The gain parameter can be related to the excitation signal and the predictor error signal in the following manner:

The excitation signal can be written as

$$Gu(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{21}$$

and the predictor error signal is given by

$$e(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k). \tag{22}$$

To match the energy of the speech production model to the energy of the LP predictor, assume that $a_k = \alpha_k$, which indicates that the predictor coefficients and the model coefficients are identical. This assumption leads to the following

$$e(n) = Gu(n) \tag{23}$$

and for the short-time analysis, Eq. (23) can be written as

$$G^2 \sum_{m=0}^{N-1} u^2(m) = \sum_{m=0}^{N-1} e^2(m) = E_n. \tag{24}$$

Substituting for E_n using Eq. (11), the gain parameter is given by

$$G = \left[R(0) - \sum_{k=1}^{p} \alpha_k R(k) \right]^{1/2}. \tag{25}$$

3.7. Pitch period determination

Fig. 8 shows the basic steps of many pitch determination algorithms. In the pre-processing stage, the speech signal is preprocessed on a global basis to enhance the fundamental frequency F_0. Then the preprocessed speech is passed to the pitch estimator to calculate an estimate of F_0, or the pitch period $1/F_0$. The resultant F_0 estimates are cleaned by post-processing techniques to obtain the final pitch contour.

Although several pitch determination algorithms (PDAs) have been proposed over the past few decades [7], a parallel processing approach developed by Gold and Rabiner [11] was chosen for the following reasons:

- It has been used with great success in a variety of applications.
- It is a simple and fast algorithm working in time-domain.
- It can be implemented easily on a general-purpose computer.

The block diagram of this pitch detector is illustrated in Fig. 9. In this approach, six individual pitch estimators were used.

3.8. Pre-processing

In the pre-processing stage, the speech signal is passed through a low pass filter with a cutoff frequency at 900 Hz. This filter produces a relatively smooth waveform

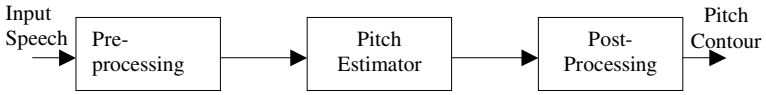

Fig. 8. Main steps in pitch determination.

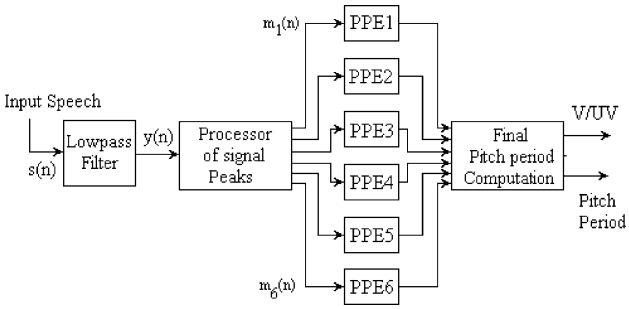

Fig. 9. Gold–Rabiner [11] parallel processing pitch detector.

and suppresses the effects of high frequency components of the input signal. A 79-tab finite impulse response (FIR) low pass filter is used for this purpose. The block diagram of this filter is depicted in Fig. 10 and the filter output is given by

$$y(n) = \sum_{i=0}^{R} h_i s(n-i), \tag{26}$$

where $s(n)$ is the input signal, h_i is the filter coefficients, and R is the filter order.

3.9. Estimation

After the speech signal is low pass filtered the peaks and valleys of the filtered signal are determined. Six impulse trains are then generated from the amplitudes and locations of these peaks and valleys as shown in Fig. 11. These pulses are defined as:

- $m_1(n)$ is an impulse equal to the peak amplitude occurring at the location of each peak.
- $m_2(n)$ is an impulse equal to the difference between the peak amplitude and the preceding valley amplitude occurring at each peak.
- $m_3(n)$ is an impulse equal to the difference between peak amplitude and the preceding peak amplitude occurring at each peak.

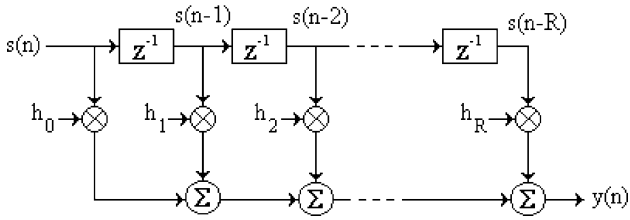

Fig. 10. The block diagram of the FIR filter.

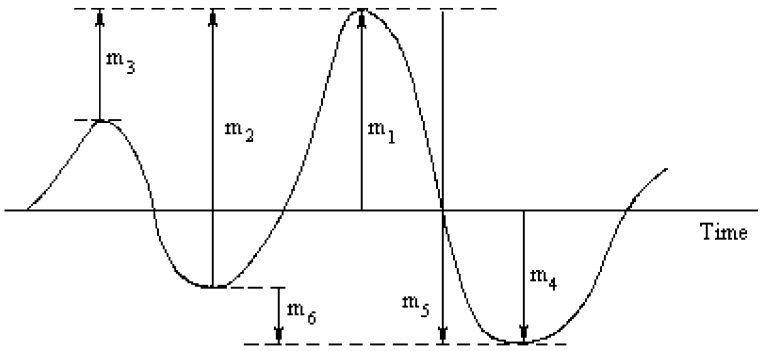

Fig. 11. Impulses generated from the peaks and valleys.

- $m_4(n)$ is an impulse equal to the negative of the amplitude at a valley occurring at each valley.
- $m_5(n)$ is an impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding peak occurring at each valley.
- $m_6(n)$ is an impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding local minimum occurring at each valley.

The generated impulse trains are then passed to the pitch period estimators (PPEs). The basic operation of each estimator is shown in Fig. 12. If an impulse is detected at the input, the output is set to the amplitude of that impulse and is held for a blanking period τ, during which no pulse can be detected. Then at the end of this period, the output starts to decay exponentially and the detection process starts again. If an impulse with sufficient amplitude exceeds the level of the decaying output, the process is repeated. The length of each pulse is considered as an estimate of the pitch period.

This procedure is applied to each of the six PPEs to obtain an estimate from each PPE. Finally these estimates are compared and the value with the most occurrence is chosen as the pitch period [12].

3.10. Post-processing

The initial estimates of pitch period are often inaccurate due to speech amplitude variability, vocal tract interference, and high noise margins. This may cause undesirable pitch doubling or halving. Consequently, post-processing is used to improve the naturalness of pitch contour. Fig. 13 shows an example of an estimated pitch-contour with some undesirable values.

To overcome this problem, the following criterion was used to remove the unwanted pitch values:

- If an unvoiced frame occurs between two voiced frames, the pitch period of the frame is interpolated.
- If a voiced frame occurs between two unvoiced frames, the frame is considered unvoiced and the pitch value is set to zero.
- Further improvement is achieved by performing median filtering within a series of voiced frames.

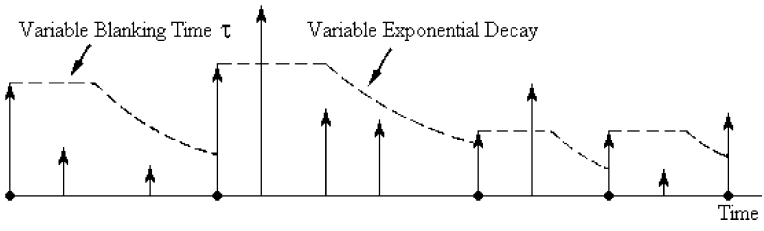

Fig. 12. The operation of each pitch period estimator.

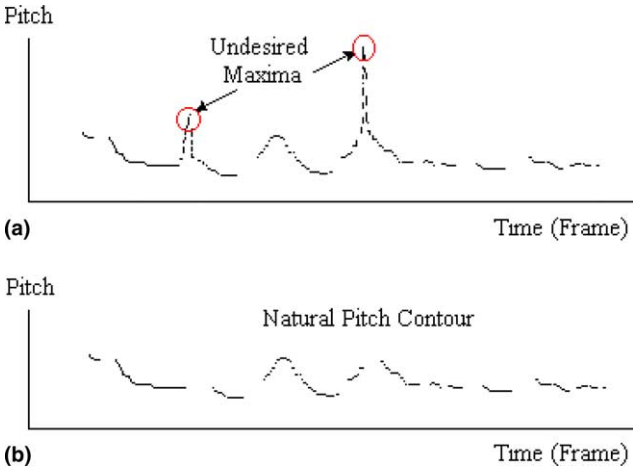

Fig. 13. Pitch contour. (a) Initial; (b) post-processed.

3.11. Mapping algorithm

The purpose of LP analysis within this framework is to extract the speech parameters related to speaker identity. Both vocal tract and excitation characteristics are extracted as discussed in the previous sections. The vocal tract information is represented by the LP coefficients, while the pitch period and the gain parameters provide the excitation characteristics. The analysis is carried out to extract these parameters from both the source and target speech signals. However, in order to achieve the goal of voice conversion the extracted parameters of the source speech signal have to be modified to match that of the target speaker.

3.11.1. Parameters statistical analysis

For each parameter extracted from the LP analysis a set of statistical values is obtained: mean, variance and standard deviation.

The mean μ, also known as the average value, estimates the value around which central clustering occurs. For a set of random variables x_j, the mean is given by

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j, \tag{27}$$

where x_j is the jth random variable and N is the total number of variables.

Variance describes the width or variability around the central value. It is defined as

$$\mathrm{var}(x) = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \mu)^2, \tag{28}$$

where μ is the mean as described above.

The standard deviation describes how far the variable x_j is from the mean. It is given by

$$\sigma(x) = \sqrt{\mathrm{var}(x)}, \tag{29}$$

where var(x) is the variance.

3.11.2. Pitch contour modification

The pitch contour modification involves matching both the pitch mean value and range. The modified pitch, p_{mod}, is obtained by modifying the source speaker pitch by applying the following mapping function:

$$P_{\mathrm{mod}} = AP_{\mathrm{s}} + B, \tag{30}$$

where P_{s} is the source pitch period of the current frame.

In Eq. (30), A and B are mapping parameters given by

$$A = \left(\frac{\mathrm{var_t}}{\mathrm{var_s}}\right)^{1/2}, \tag{31}$$

where $\mathrm{var_s}$ is the pitch variance of the source and $\mathrm{var_t}$ is the pitch variance of the target.

$$B = \mu_{\mathrm{t}} - (A\mu_{\mathrm{s}}), \tag{32}$$

where μ_{s} is the pitch mean of the source and μ_{t} is the pitch mean of the target.

The mapping parameter, A, is used to match the pitch range of the source speaker with the pitch range of the target. On the other hand, the value of B is set to achieve the same matching in the sense of mean value. This mapping technique guaranteed that the modified pitch is following the pitch envelope of the source speaker while having the average and range values of the target. Some satisfactory results were obtained by applying this mapping technique.

3.11.3. Gain contour modification

The gain parameter is modified in the same manner as the pitch period. The following mapping is used to match the gain of the source to that of the target.

$$G_{\mathrm{mod}} = (G_{\mathrm{s}} - \mu_{\mathrm{s}}) + \mu_{\mathrm{t}}, \tag{33}$$

where G_{s} is the source gain of the current frame, μ_{s} and μ_{t} are the average gain of the source and target, respectively.

3.11.4. LP coefficients modification

The mapping of the LP coefficients is carried out on the basis of the voicing decision. If the current speech frame is voiced, the mapping function is applied otherwise the modified coefficients are set equal to the source LP coefficients. The modified LP coefficients, K_{mod}, are obtained by applying the following equation:

$$K_{\text{mod}} = \begin{cases} a_t \mu_s & \text{if current frame is voiced,} \\ a_s & \text{otherwise,} \end{cases} \tag{34}$$

where μ_s is the source LP coefficients mean value, a_s and a_t are the LP coefficients of the source and target, respectively.

3.12. Speech synthesis

The modified parameters are used to build the synthesis filter, the output of which is the modified speech. Fig. 14 shows the lattice implementation of the LP synthesis filter.

In Fig. 14, the excitation signal $e(n)$ is generated from the modified parameter. These parameters are modified pitch, modified gain and voicing decision. Based on the voicing decision, the excitation signal is generated as a pulse train for voiced frames of speech, or a random noise signal for the unvoiced speech. An impulse train generator is used to generate a pulse of unit amplitude at the beginning of each pitch period. For unvoiced excitation, a random noise generator produces a uniformly distributed random signal. The amplitude of the final excitation signal is scaled by the gain parameter. The generated output speech is then passed to the de-emphasis filter to remove the effect of the pre-emphasis filter, which was applied prior to the LP analysis as shown in Fig. 15.

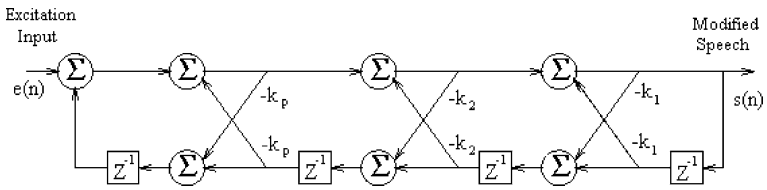

Fig. 14. The lattice implementation of the LP synthesis filter.

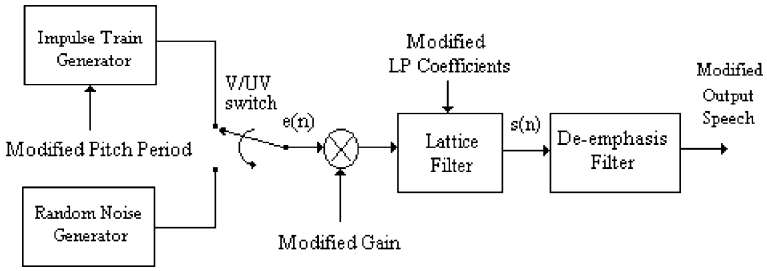

Fig. 15. Block diagram of the final synthesis process.

4. Implementation of VC

To illustrate the feasibility of the proposed methods and algorithms explained in the previous sections, MATLAB 5.3 and Simulink 3.0 were devised to accomplish this task. Fig. 16 shows the LP analysis simulation environment under Simulink. This configuration was used as a helping tool to simulate the LP analysis algorithms. The results of each block were stored separately for later use.

The *From Wave File* block takes a pre-stored speech in the wave format and allows the user to set the analysis frame length. The *Pre-emphasis Filter* block is used to apply the pre-emphasis filter to the input speech signal where the pre-emphasis filter parameters are set to the desired values. The *Window Function* block is used to apply Hamming window function to the input speech signal. The*Auto-correlation Function* block computes the auto-correlation matrix from the pre-emphasized speech. The maximum positive lag in the parameters dialog box is set according to the LP analysis order. The *Levinson–Durbin* block is used to perform the Levinson–Durbin algorithm. The input to this block is the auto-correlation matrix and the output is a set of LP coefficients. The *Time-varying Filter* block is used to construct the time-varying filter from the computed LP coefficients. For the analysis, an all zero configuration is used, while an all pole filter configuration is used for the synthesis part. The output of the analysis filter is the prediction error signal, which is also known as the prediction residual. The synthesis filter, on the other hand, produces a synthetic speech signal equivalent to the original speech.

4.1. VC software

The program starts by initializing all the variables to their initial values. The speech signal for the respective speaker is then loaded and the LP analysis routines

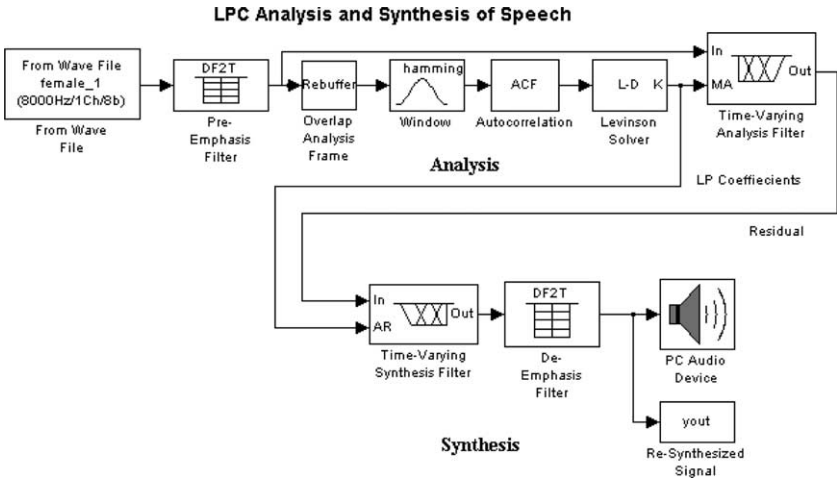

Fig. 16. Block diagram of LP analysis simulation using Simulink.

are performed. The results from the analysis are saved in the respective speaker analysis arrays. The modification procedure is then applied to the analysis results producing a set of modified parameters. These parameters are passed to the synthesis function to produce the final speech signal. The speech loading function returns the length of the speech signal and loads the speech data to the respective data variable. The LP analysis part produces the LP coefficients and the gain of the speech signal. The Gold and Rabiner algorithm is used to estimate the pitch period of the speech signal. The initial pitch estimates are post-processed to remove any inaccurate values and the final results are saved. The program provides a graphical user interface (GUI) for displaying the speech signals and the analysis results. It also allows the user to playback the original speech files and record new speech signals.

5. Results and discussion

The proposed algorithm of voice conversion was implemented using C++. Fig. 17 shows a screen shot of the program main window.

The initial analysis results are displayed on the main screen using *view analysis results*, as shown in Fig. 18.

Display speech signal function allows the user to display the speech waveform of both the source and target speakers. The LP analysis results also are displayed using this function. These results include the pitch contour, the gain contour, and the voicing decision. In Fig. 19, some LP analysis results for the same speech waveforms of Fig. 20 are shown.

One way of evaluating speech synthesis applications is through informal listening tests. In these tests, the listeners hear two successive of the original speech and the synthesized speech. In many cases, the listeners can be told the content of the

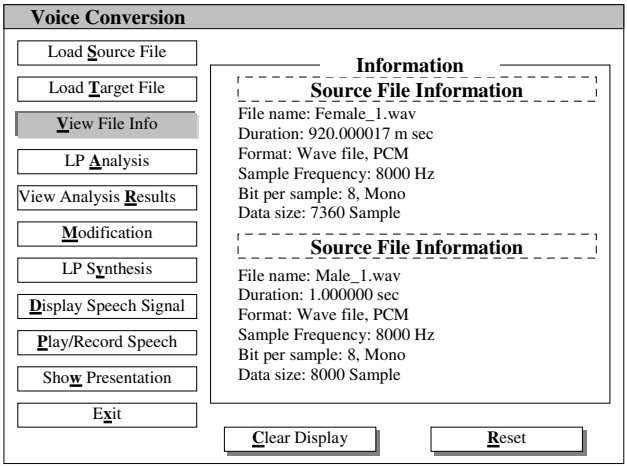

Fig. 17. Program main window.

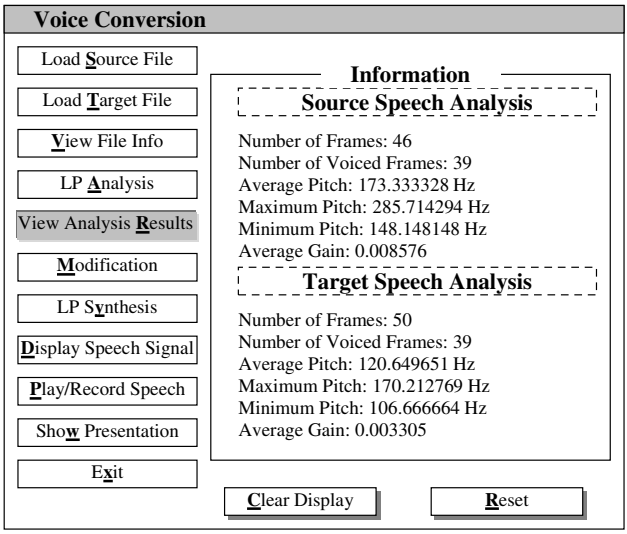

Fig. 18. LP analysis results.

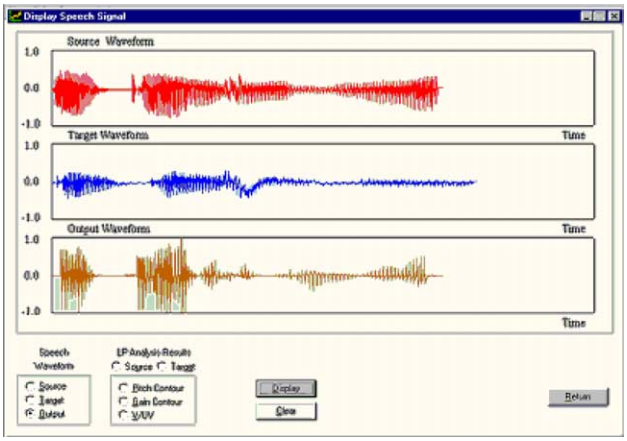

Fig. 19. Speech waveform display.

sentences that they will hear. After hearing both the speech utterances, listeners are asked certain questions about the quality and intelligibility of the modified speech [4]. These tests and the experimental results have shown that certain voices sounded better than others regardless of information content of speech signal. In general, the system works better with female voices as target speakers as female voices tend to have higher pitch than male voices. The experimental work also has shown that fe-

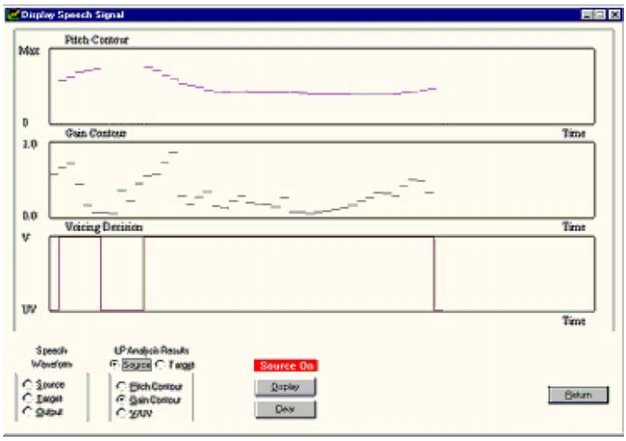

Fig. 20. LP analysis results display.

male voices tended to be much clearer and smoother, and it was often very easy to hear the distinguishing characteristics of female voices as compared with male voices.

However, in all cases it was obvious that the output voice was not the original, and even when the output speech message was clear, the voice was close to the target speaker. In some cases, due to noisy signal, the voicing decision and hence the pitch period determination were affected. This resulted in degradation of the output so that the output speech became unintelligible.

6. Conclusion

The numerical and the visual results from the program show that inter-conversion of the source speaker and the target speaker was achieved. The modified pitch contour in all cases follows the source pitch contour while maintaining the average value and the range of the target. The same results were obtained for the gain contour. The fundamental frequency or pitch period of speech signal is an important parameter. It is well known in the speech signal processing literature that pitch determination is a difficult task. This problem led researchers to develop a number of pitch determination algorithms (PDAs). However, no single PDA has given reliable results in all situations. Since the quality of the modified speech relies greatly on the accurate determination of pitch, any advances in PDAs will enhance voice conversion processes. To improve the VC throughput, good quality speech recordings are needed. Low quality speech files will affect the entire transformation process. Using higher quality speech databases, more reliable and satisfactory results can be achieved. In the case of real time analysis of speech signal, a powerful DSP hardware processor is required. Voice conversion is still an immature field and many new methods can be expected to appear in the literature over the next few years.

References

[1] Kuwabara H, Takagi T. Acoustic parameters of voice individuality and voice quality control by analysis-synthesis method. Speech Commun 1991;10(5):491–5.
[2] Kuwabara H, Sagisaka Y. Acoustic characteristics of speaker individuality: control and conversion. Speech Commun 1995;16(2):165–73.
[3] Childers DG, Lee CK. Vocal quality factors: analysis, synthesis and perception. J Acoust Soc Am 1991;90:2394–410.
[4] Childers DG. Speech processing and synthesis toolboxes. New York: Wiley; 2000.
[5] Porat B. A course in digital signal processing. New York: Wiley; 1997. p. 554–61.
[6] Atal BS, Hanauer S. Speech analysis and synthesis by linear prediction of the speech wave. Acoust Soc Am 1971;50(2):637–55.
[7] Hess W. Pitch determination of speech signals, algorithms and devices. Berlin, Heidelberg: Springer; 1983.
[8] Childers DG, Wu K, Hicks DM, Yegnanarayana B. Voice conversion. Speech Commun 1989;8(2):147–58.
[9] Childers DG. Glottal source modeling for voice conversion. Speech Commun 1995;16(2):127–38.
[10] Kain A, Macon M. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In: Proceedings of ICASSP, May 2001.
[11] Gold A, Rabiner LR. Parallel processing techniques for estimating pitch periods of speech in the time domain. J Acoust Soc Am 1969;46:442–8.
[12] Rabiner LR, Schafer RW. Digital processing of speech signals. Englewood Cliffs (NJ): Prentice-Hall; 1978.