IJICT International Journal of Information & Communication Technology

ITRC

Volume 1- Number 1- March 2009 (pp. 13-20)

Wavelet-based Robust Voice conversion systems

M.Farhid
Department of Electrical and computer Engineering
University of Tabriz
Tabriz, Iran
morfid@gmail.com

M.A.Tinati
Department of Electrical and computer Engineering
University of Tabriz
Tabriz, Iran
tinati@tabrizu.ac.ir

Abstract— **Voice conversion is a method used to transform one speaker's voice into another speaker's voice. New modification approach for voice conversion is proposed in this paper. We take Mel-frequency Discrete Wavelet coefficients (MFDWC) as the basic feature. This feature copes well with small training sets of high dimension, which is a problem often encountered in voice conversion. The proposed voice conversion system consists of both off-line (training) and on-line (transformation-synthesis) procedures and assumes parallel training data from source and target speakers and uses the theory of wavelets in order to extract speaker feature information. The satisfactory performance of the voice conversion system can be confirmed through ABX listening test and MOS grade.**

Keywords- voice conversion; wavelet; vector quantization; Formants; ABX Test;

I. INTRODUCTION

Speech signal conveys mainly two kinds of information, namely the speech message part and the speaker identity section or voice individuality part. Isolating the personality of speech and speaker from the signal is a challenging problem in speech research. Extracting the message part of the information is the focus of research in the area of speech recognition. Voice conversion, on the other hand, is a technique that aims to transform the input (source) speech signal such that the output (transformed) signal will be perceived as produced by another (target) speaker. It transforms how something is said without changing it. Analysis of speaker dependent characteristics is also necessary for developing speech synthesis systems. Improving synthesized speech quality by adding natural characteristics of voice individuality and converting synthesized voice individuality from one speaker to another is one of the important problems in speech technology. Two types of acoustic characteristics, the voice source characteristics and the vocal tract resonance characteristics, act together to influence voice individuality. The applications of such technology are several such as text-to-speech adaptation where the voice conversion system can be trained on relatively small amounts of data and allows new voices to be created at a much lower cost than the currently existing systems. Other applications can be found in broadcasting, voice editing, karaoke applications, internet voice applications as well as computer and video games. Voice conversion can be used for looping and dubbing applications. Looping is defined as replacing the undesired utterances in a speech recording by desired ones. This technique can be used for processing movies for TV broadcast. In order to obtain transparent quality such that the listeners will not be able to distinguish the replacement necessitates the use of voice conversion. E-mail readers supply as an important tool in Interactive Voice Response (IVR) systems. With these systems, people can listen to their e-mail messages on the phone. Personification of e-mail readers using voice conversion will offer the possibility to attach a voice font to each personality and the messages can be read by the sender's voice or any voice the user may prefer. All these applications

require high quality output. It is also important that the methods to be used must facilitate fast and believable voice conversion.

Voice conversion is performed in two steps. In the training stage, acoustic parameters of the speech signals uttered by both the source and target speakers are computed and suitable rules mapping the acoustic space of the source speaker into that of the target speaker are obtained. In the transformation stage, the acoustic features of the source signal are transformed using the mapping rules such that the synthesized speech sounds like the target speaker [1].There has been a considerable amount of research directed at the problem of voice transformation [2, 3], using the general approach described above. The first approaches were based around linear predictive coding (LPC) [1]. This approach was improved by using residual-excited LPC (RELP), where the residual error was measured and used to produce the excitation signal [8]. Most authors developed methods based on the interpolation of speech parameters and modeling the speech signals using formant frequencies [1], Linear Prediction Coding (LPC) cepstrum coefficients, and Line Spectral Frequencies (LSFs) [7], and harmonic-plus-noise model parameters [9] or based on mixed time- and frequency- domain methods to alter the pitch, duration, and spectral features. These methods are forms of single-scale morphing. Although the above methods provide good approximation to the source-filter model of human vocal tract and they encode good quality speech, they face two problems [10]: artifacts and hiding detailed information during the extraction of formant coefficients and the excitation signal. Mel-Frequency Cepstrum Coefficient (MFCC) has become the default set of features for speech recognition right from the time they were discovered. However, with more research it was found that MFCC perform rather poorly when speech input is noisy. At the same time wavelets have gained a lot of importance recently in the field of signal processing. Wavelet spaces are a series of function spaces that are highly decorrelated from each other and are particularly suitable for the representation of signals and operators at different resolution scales that exhibit speech and speaker feature behaviors. Our proposed model uses the theory of Wavelets as a means of extracting the speech features, Mel-Frequency Discrete Wavelet Coefficients (MFDWC), as followed by Data grouping for modeling the spectral conversion.

II. PROPOSED ALGORITHM

Recently, wavelet transforms have found widespread use in various fields of speech processing. Among the many applications, wavelets have been used in automatic speech recognition, pitch detection, speech coding and compression, and speech denoising and enhancement. The lack of detail in the converted speech produced by the existing methods leads to the conclusion a multi-scale voice conversion method should be tested that performs the conversion in different levels of analysis (subbands) and capture in more detail the range of frequencies of speech signals. It is generally believed that abrupt stimulus changes, which in speech may be time-varying frequency edges associated with consonants, transitions between consonants and vowels and transitions within vowels are critical to the perception of speech by humans and for speech recognition by machines. Noise affects speech transitions more than it affects quasi-steady-state speech. We believe that identifying and selectively amplifying speech transitions may enhance the intelligibility of speech in noisy conditions. Wavelet analysis, because of its multiresolution properties, can detect voiced stops, since stops have a sudden burst of high frequency.

A. Feature Extraction

Efficient representation of speech signal in terms of slowly varying parameters is a problem of considerable importance in speech research. Most methods for analyzing speech start by transforming the acoustic data into spectral form by performing a short-time Fourier analysis of speech wave. Although spectral analysis is a well-known technique for studying signals, its application to speech signals suffers from a number of serious limitations arising from the non stationary as well as the quasi-periodic properties of speech wave. In addition, methods based on spectral analysis often do not provide a sufficient accurate description of speech articulation. For instance, the traditional Fourier analysis methods require a relatively long speech segment to provide adequate spectral resolution. As a result, rapidly changing speech events can not be accurately followed. Furthermore because of periodic nature of voiced speech, little information about the spectrum between harmonic is available.

In modern signal processing techniques, the procedures for analyzing a signal make use of all the information that can be obtained in advance about the structure of that signal. Based on such information, the first step in signal analysis is thus to make a model of the signal. Subband decomposition is implemented using the Discrete Wavelet Transform (DWT). The use of wavelets in signal processing applications is continually increasing. This use is partly due to the ability of wavelet transforms to present a time-frequency (or time-scale) representation of signals that is better than that offered by Short-time Fourier transform (STFT). Unlike the STFT, the wavelet transform uses a variable-width window (wide at low frequencies and narrow at high frequencies) which enables it to "zoom in" on very short duration high frequency phenomena like transients in signals Wavelets are a class of functions that possess compact support and form a basis for all finite energy signals. The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. They are able to capture the non-stationary spectral characteristics of a signal by decomposing it over a set of atoms which are localized in both time and frequency. The DWT uses the set of dyadic scales and translates of the mother wavelet to form an orthonormal basis for signal analysis. In wavelet decomposition of a signal, the signal is split using high-pass and low-pass filters into an approximation and details. The approximation is then itself split again into an approximation and a

detail. This process is repeated until no further splitting is possible or until a specified level is reached. The DWT provides a good signal processing tool as it guarantees perfect reconstruction and prevents aliasing when appropriate filter pairs are used. Decomposing a signal into k levels of decomposition, Therefore results in k+1 set of coefficients at different frequency resolutions, k levels of detail and 1 level of approximation coefficients. The original signal *S* is split into an approximation ca1 and a detail cd1. The approximation is then itself split into an approximation and a detail and so on.

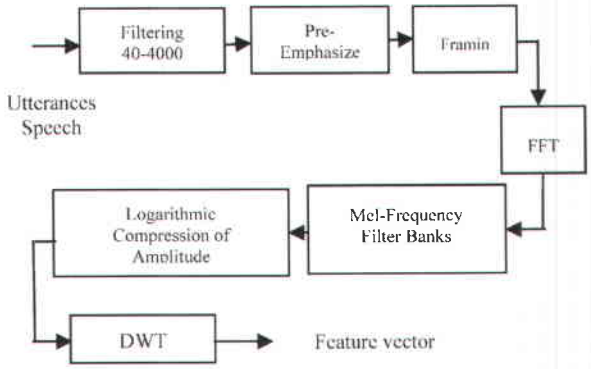


Fig. 1 Schematic of MFDWC computation

Speaker characteristics, also known as speaker personality, are 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 [12]. 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 [13]. 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. For this reason static features are considered to be more useful for voice conversion applications. 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. The aim of signal analysis is to extract relevant information from a signal by transforming it. Some methods make a priori assumptions on the signal to be analyzed; this may yield sharp results if these assumptions are valid, But is obviously not of general applicability. Selecting a useful and relevant subset of features from a larger set is crucial to improve the performance of voice conversion systems. Figure 1 gives the flowchart of the process of obtaining MFDWC coefficients from speech. This feature first time was used for Speech Recognition in [4]. The sampled speech signal is filtered by a 48 order Butterworth filter with pass-band from 40-4000 Hz. Then is pre-emphasis which basically makes one sample in speech influence the next sample by a certain weight. Next the speech signal is divided in to frames. After this Hamming window is applied to each sample in the frame have a smooth transition between samples of a frame. The ear has better frequency resolution at lower frequencies. A Mel is a unit of measure of perceived pitch or frequency of a tone. The Mel-scale is therefore a mapping between the real frequency scale (Hz) and the perceived frequency scale (mels). The mapping is virtually linear below 1 KHz and logarithmic above. Next total energy of each frame is calculated. The difference between MFCC and MFDWC is in the final step. In MFDWC instead of DCT, discrete wavelet transform (DWT) is performed in the final step of obtaining feature vectors. Essentially, DCT step in the calculation of MFCC features decorrelates the filterbank energies. It has been shown that the wavelet transform is a better decorrelator in coding applications.

We use the orthogonal filters corresponding to Daubechies's orthogonal wavelets [5] in the wavelet transform. So energy is preserved in the transformation. Our voice conversion algorithm is implemented, has shown in Figure 2, using the following steps:

- The raw source and target sentences are time-aligned using Dynamic Time Warping (DTW) algorithm, so that they have the same length.
- Divide the speech signal into frames of equal size. In this paper different frame sizes are tested to see how the frame size will affect the performance of the reconstructed signal. Three different frame sizes are examined since wavelet analysis is not affected by the stationary problem. Increasing the frame length will speed up the processing time which reduces the processing delay.

Apply the Mel-frequency discrete wavelet transform to each one of these frames separately at the five decomposition levels. Choosing the right decomposition level in the DWT is important for many reasons. For processing speech signals no advantage is gained in going beyond scale 5. This level is chosen since the best performance of the reconstructed signal is obtained at this level. The wavelet used by me for performing DWT is known as Daubechies D4 wavelets. Daubechies did not propose a particular function as wavelets but simply specified a set of desired features for the wavelet coefficients (wavelet coefficients are evaluations of a wavelet function for different values of scaling and translation parameters). As a result of this she proposed a set of equations required to be satisfied by the coefficients. Once one chooses the number of coefficients (which needs to be an even number), and can solve the equations and obtain the coefficients. In general optimum wavelets can be selected based on the energy conservation properties in the approximation part of the wavelet coefficients. Wavelets with more

vanishing moments provide better reconstruction quality, as they introduce less distortion into the processed speech and concentrate more signal energy in a few neighboring coefficients. However the computational complexity of the DWT increases with the number of vanishing moments and hence for real time applications it is not practical to use wavelets with an arbitrarily high number of vanishing moments. The wavelets thus obtained are called Daubechies DX wavelets, where X stands for the number of coefficients.

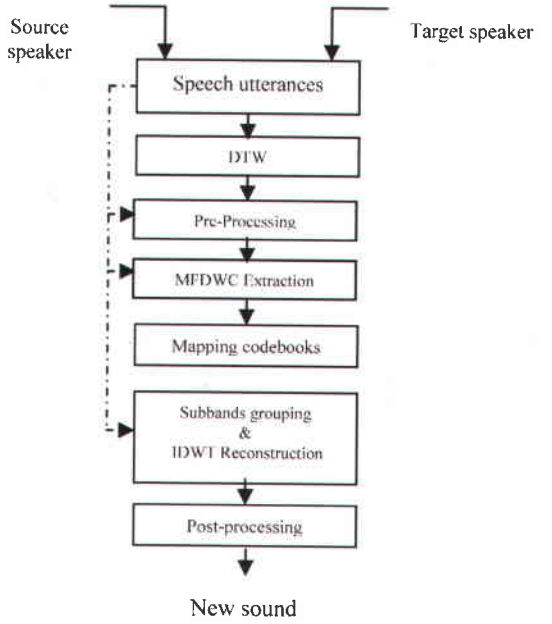


Fig. 2 Proposed Model

- The level 1 and level 2 detail coefficients are set to zero [10]. For each level of decomposition, Data grouping method used for training using the Source and target training sample wavelet coefficients.
- The transformed coefficients are used in order to reconstruct the signal.

We use vector quantization to extract codebook for frames of the source speaker. Once the training procedure is accomplished, we can acquire the mapping codebook. The structure of the mapping codebook is one by one. Therefore, the synthesizer only compares features among the mapping codebook and discovers the most similar features to decode. The figure 3 demonstrates the flowchart of the mapping codebook generation.

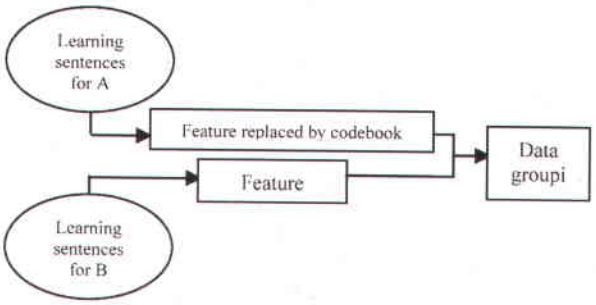
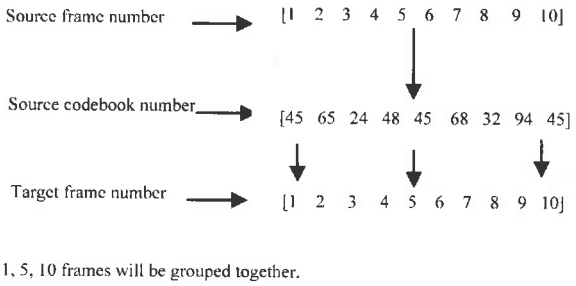


Fig. 3 Mapping codebook

B. Data Assemblage for Target Speaker

Once we have the codebooks for the source speaker, we can locate the data grouping (for the target speaker) induced by the alignment procedure of dynamic time warping [6]. Once the data grouping is done, we can find the centroid of each group and the mapping from target centroids to source centroids can be established intuitively. The principle of the induced grouping can be explained via the following example.



1, 5, 10 frames will be grouped together.

Fig. 4 An example of data grouping

In figure 4, the first and second rows show the mapping from the frame indices of the source speech to the code vector indices. The first and the third rows show the alignment between the frame indices of the source sentence and those of the target one. As a result, we can establish the mapping from the code vector indices to the frame indices of the target sentence. This mapping induces a grouping on the target frames. For instance, frames 1, 5, and 10 of the target sentence should be in the same group since they have the same code vector index 45 in the second row. After alignment and grouping, the frames in the target sentence are partition into several groups. In general, the number of group in the target sentence should be the same as the number of codebook size of the source sentence. To establish the mapping from the source frames to the target ones, we still need to compute the "centroid" of each group in the target frames. Usually the centroid of a group is the average of all vectors in the group. However, in our case, we would like to avoid the conversion from MFDWC to a speech frame. As a result, the centroid should be one of the data point in the group. Thus the centroid is obtained as the data point that has a minimal total distance to all the other data points in the same group.

Once the centroid of each target group is determined, we can perform frame to frame voice conversion via the following steps:

- GET A SOURCE FRAME AND FIND ITS CLOSEST CODE VECTOR.
- FIND THE CORRESPONDING GROUP IN THE TARGET FRAME.
- RETURN THE CENTROID OF THE IDENTIFIED TARGET GROUP

III. EXPRIMENTAL RESULTS

We take 4660 frames of a female source speaker as training data and compare them with 4406 frames of a male target speaker to extract our mapping codebooks. The codebooks have128 vectors for frames. We take 20 sentences for training the codebooks. In the first experiment we modified a utility provided to generate MFCC to make it generate MFDWC coefficients. The vector of each feature was now 16 instead of 13. The number of melfilters was changed from 24 to 28. This was necessary performing DWT. Figures 5, 6 show results of our converting on speech from male-to-female and female to-male pairs of speakers with MFDWC.

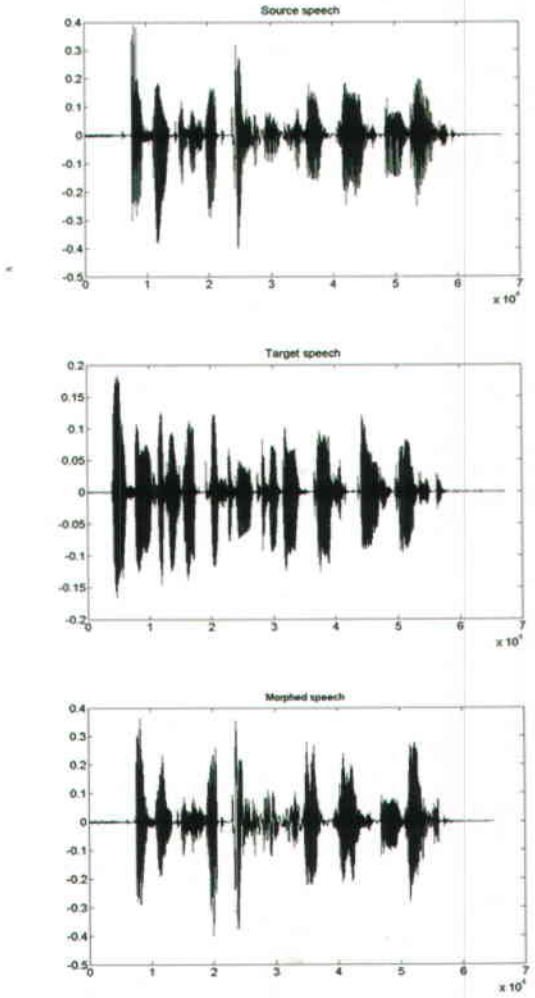


Fig. 5 Source, Target and Converted sentence waveforms for male-to-female speaker speech transformation of utterance.

To subjectively evaluate the performance of our system, two forced-choice (ABX) experiments and an MOS (mean opinion score) tests were performed. We take on two ABX experiments to evaluate the improvement of the proposed method. (A and B are the source and target speech utterances, respectively, and X is the result of converting source speaker's Utterance to *target speaker's ones).* ABX listening tests were performed to evaluate the subjective performance of the proposed method. We have used different combinations of 4 male and 4 female speakers as the source and the target. Four types of voice conversion is performed as far as the gender of the source and the target is concerned (female-to-female, female-to-male, male-to-female, and male-to male). First, training and test utterances were recorded at 11.025 KHz. Full-band and subband based codebooks are generated by two separate training sessions for each conversion. The subjects were provided with three recordings each time they were asked to make a decision: (A) Full-band based conversion output, (B) Subband based conversion output, and (X) Target recording. The conversion outputs are presented in random order and the listener is asked to judge whether (A) or (B) sounds more like the target speech (X). The order in which the full-band and the subband based output are presented is also changed randomly. And the MOS experiment was carried out here to estimate the listening quality, using a 5-point scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent.The following tables show the experimental results.

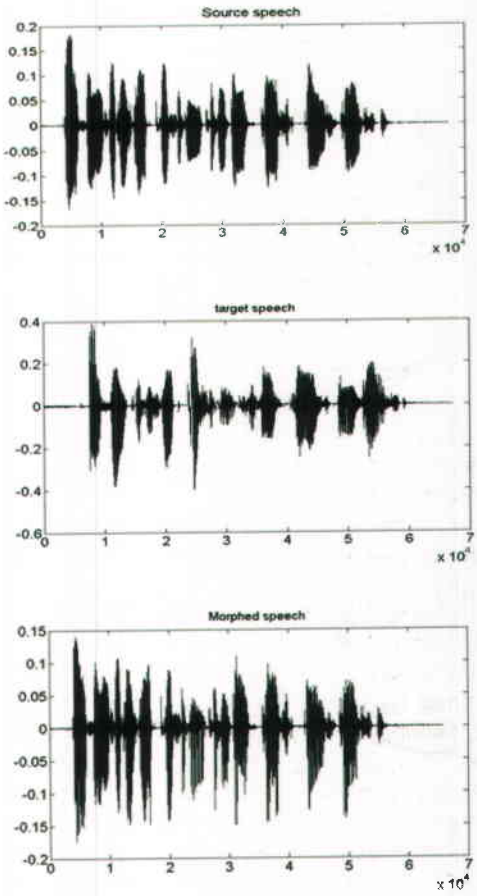


Fig. 6 Source, Target and converted sentence waveforms for female-to-male

From the results of tables I and II, it is obvious that the proposed feature can effectively increase the degree of naturalness of the converted speech. In [11] the human factor cepstral coefficient (HFCC) was used for voice conversion systems. In general, the system works better with male voices as target speakers as female voices tend to have higher pitch than male voices. The experimental effort also has shown that female 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. The first three formants are extracted. Assuming the female speaker as the source speaker and male speaker is target. 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.

TABLE I. RESULTS OF PERCEPTUAL TESTS FOR FEMAIL TO MAIL. (TEST QUESTION: "X IS CLOSER TO A OR TO B?")

64 cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	69.0%	71.0%	78.0%	78.0%
MOS	3.0	3.0	3.1	3.4

128 cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	74.0%	76.0%	86.0%	91.0%
MOS	3.6	3.6	4	4.2

256 cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	78.0%	81.0%	88.0%	93.0%
MOS	3.7	3.8	4.1	4.5

In some cases, due to noisy signal, MFDWC works better than other three features that was explored in TableII. For future works it can be tested for other types of noise for example convolutive noise. There are many factors at the suprasegmental level which affect the speaker's voice quality. The prosodic features such as pitch, duration and intensity as a function of time are unique for a given speaker. In fact a speaker is identified more by these factors than by the vocal tract features, as is evident while listening to the linear prediction residual signal of speech. But these features are dependent on the context and also on the language of the speaker. As we know spectrograms convert a two-dimensional speech waveform (amplitude/time) into a three-dimensional pattern (amplitude/frequency/time). With time and frequency on the horizontal and vertical axes respectively, amplitude is noted by the darkness of the display and peaks in the spectrum (e.g., formant resonances) appear as dark horizontal bands. In this way, spectrograms furnish much useful information relevant to the detailed motion of formants.

TABLE II. RESULTS OF PERCEPTUAL NOISY TESTS FOR FEMAIL TO MAIL. (TEST QUESTION: "X IS CLOSER TO A OR TO B?")

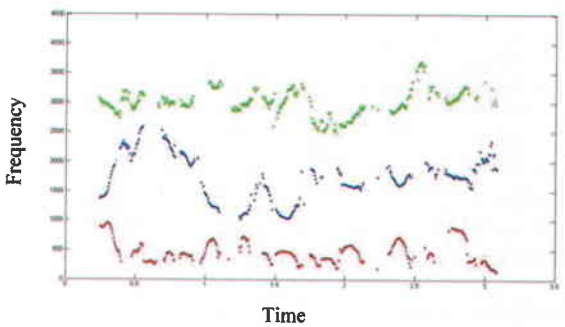
64cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	52.0%	68.0%	76.0%	84.0%
MOS	2.5	2.8	3.1	3.3

128 cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	60.0%	70.0%	81.0%	87.0%
MOS	3.0	3.2	3.6	3.8

256 cluster	MFCC	MFCC with cepstral mean subtraction	HFCC	MFDWC
ABX	74.0%	76.0%	86.0%	91.0%
MOS	3.6	3.6	4	4.2

The next example concerns the spectral transformation of English sentence from a female speaker to male speaker. The corresponding sound patterns are shown in the formant trajectory in Figure 7. Because it has been shown that first three formants are important and play a main role in individuality, we plot related formant frequency versus time and then compare them with each other.

From a pervious works [14], we know when all formants are shifted uniformly; voice individually is completely lost in case of shifts of less than eight percent towards both the high and the low frequency regions. In the formant frequency shift, voice personality is more sensitive to the lower three formants than to the higher formants.



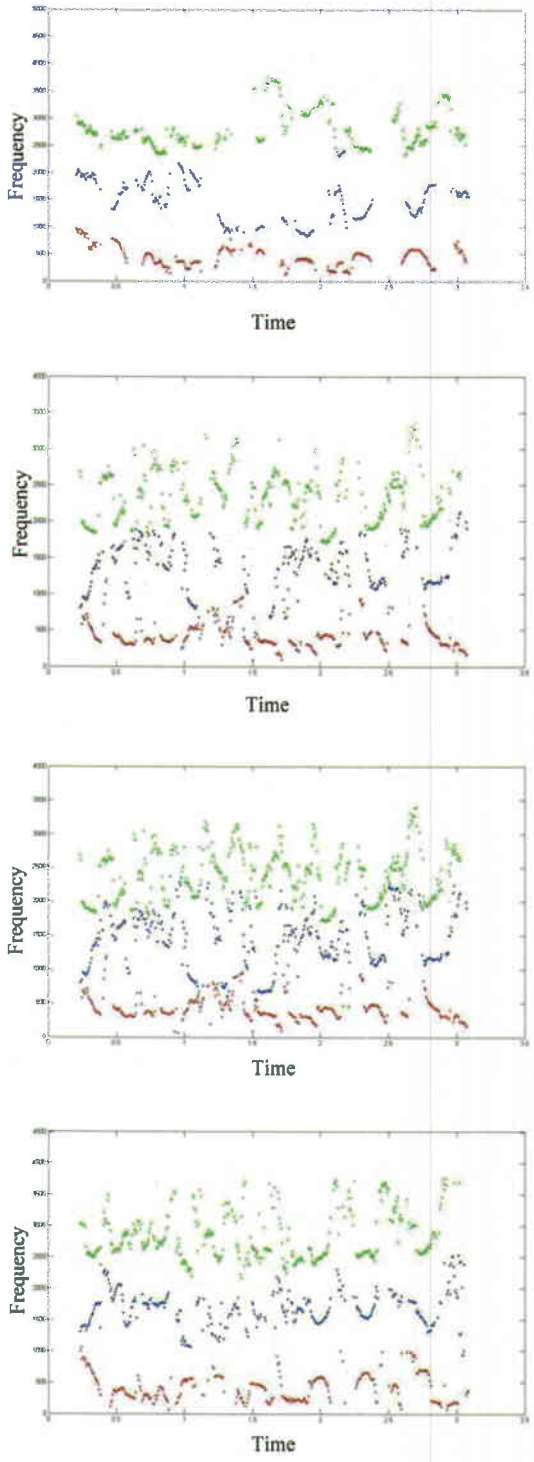


Fig. 7 Source, Target and coverted sentence first three formant trajectory for female-to-male speaker speech transformation of utterance with MFCC with cepstral mean subtraction and HFCC and MFDWC respectively.

The results for the shift of individual formants indicate that voice personality is lost rapidly when F2 shifts towards the high frequency region, or conversely, when F3 shifts towards the low frequency region. This strongly suggests that a significant amount of personality information is to be found at a frequency somewhere between F2 and F3.

According to Figure 7, it's clearly in MFDWC feature based, system works well. The results of this study on both spectral and wavelet domains voice conversion can be summarized in the following points:

- Feasibility of both domains conversion was proven by performing listening tests and examining the first three formants trace of the source, target, and transformed signal.
- Conversion between male and female speakers (especially female to male speaker) is easier than male-to male conversion or female-to-female conversion.
- Increasing the number of quantization class (cluster) from 64 to 256 results in a slight improvement of performance of the spectral transformation system.
- The models do not involve extensive computation and are easily implemented on real-time platform.

As training material, we used three kind of database: English, Persian and Germany one. These results are just for English language. The name of German database is SPINA, 1991 and for Persian one we made it in our university from college students.

IV. Conclusion

The aim of Voice conversion algorithms is to provide high level of similarity to the target voice with an acceptable level of quality. This study focuses on formulating robust technique for a codebook mapping based voice conversion algorithm. In this study, a new feature set based on the wavelet transform of speech signal and data grouping is proposed with application to voice conversion systems. Listening tests were performed to demonstrate the performance of the system. Furthermore it was observed that most distortion occurred at the unvoiced parts of signals. Consequently, we need to treat unvoiced and voiced parts separately in order to generate high-quality synthesized voice. The experimental results show that if we choose different wavelet functions for reconstruction section, we achieve interesting output voices

References

[1] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H. "Voice conversion through vector quantization ". ICASSP-88.,1988

[2] Kain, A., Macon, M.W. "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction", Acoustics, Speech, and Signal Processing, 2001,p813 -816

[3] Kain, A., Macon, M.W. "Spectral voice conversion for text-to-speech synthesis", Acoustics, Speech and Signal Processing, 1998 p285 – 289.

[4] Zahir Koradia "MelFrequency Discrete wavelet coefficient (MFDWC) for Isolated word Hindi Speech Recognition." 1May 2006

[5] I.Daubechies, "Orthonormal bases of compactly supported wavelets," Comm.Pure Apple.Math.,pp.909-996,1988.

[6] Cheng-Yuan Lin and J.-S. Roger Jang." New Refinements schemes for voice conversion" 2004,256- 260.

[7] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. "Spoken Language Processing." Prentice Hall, 2000, p424-426.

[8] L.Arsalan and D.Talkin:" voice conversion by codebook mapping of LSF and excitation spectrum". Proc. Eurospeech ,1997 ,1347-1350

[9] H.Valbert, E.Moulines and J.P.Tubach:"voice conversion using PSOLA techniques. Speech Communication 11,1992, 175-187

[10] Christana Orphanidou, Irene M.Moroz and Stephan J. Roberts "Multiscale Voice Morphing using Radial Basis function Analysis" 2004

[11] M.Farhid .M.A.Tinati "Improving quality of voice conversion systems" Advances in computer science and Engineering Springer-Verlag CSICC series 2008,880-884

[12] Kuwabara H, Sagisaka Y. Acoustic characteristics of speaker individuality: control and conversion. Speech Commun 1995; 16(2) 165-173

[13] Childers DG, Lee CK. Vocal quality factors: analysis, synthesis and perception. J Acoust Soc Am 1991

[14] M.Geravanchizadeh Spectral Voice conversion based on locally linear transformation rules for local tract characteristics ; phD thesis shaker verlag 2002

Mohammad Ali Tinati was born in 1953 in Iran. He received his B.S. degree (with high honor) in 1977, his M.S. degree in 1978 from Northeastern University, Boston, Mass, USA, and his Ph.D. degree from Adelaide University, Australia, in 1999. He had a long affiliation with the University of Tabriz, Iran. He served as an academic member of the Faculty of Electrical Engineering since 1979. His main research interests are biomedical signal processing and speech and image processing.

Morteza Farhid was born in 1982 in Iran. He received his B.S. degree in communication systems from the University of Tabriz, Tabriz, Iran in 2004 and his M.S. degree in signal processing from the same university in 2007. His main research interests include speech and image processing.