Volume 3, Issue 6, June 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Effect of LSF Based Features for Speech Signal Alignments

Jang Bahadur Singh , Parveen Lehana
DSP lab,
Department of Physics and Electronics & university of Jammu,
Jammu, India

Abstract— Speech synthesis, speech recognition, and speech transformation are the essential techniques used for human-machine communication. Feature extraction procedures give a compressed representation of the speech signals. The Harmonic plus noise model (HNM) module used for the analyses and synthesis provides high quality speech with less number of parameters. Dynamic time warping procedure is used for aligning two given multidimensional sequences. The improvement in the alignment is anticipated by the corresponding distances between the sequences. The objective of this research is to investigate the effect of LSF, HNM and dynamic time warping on phrases, words, and phonemes based alignments. The speech signals in the form of twenty five phrases have been recorded. The recorded speech is segmented manually and aligned at sentence, word, and phoneme level. The Mahalanobis distance (MD) is calculated between the aligned frames. The study has exposed better alignment in case of HNM parametric domain. It has been observed that effective speech alignment at phrase level.

Keywords— LSF, HNM, Mahalanobis distance, speech recognition, speech transformation and Dynamic time warping

I. INTRODUCTION

Signal processing involves signal transformation and signal representation. For speech signals a human speaker is the information source and the waveform to be measured is generally auditory in nature. The process involves various steps which are depicted by the flow diagram of Fig. 1. The first step involves the representation of the speech signal based on a given model and then the application of some higher level transformation in order to put the signal into a more convenient form so as it can be used properly. Finally the extraction of the message signal is done by the human listener or automatically by machines. An interesting example is the speaker recognition phenomenon done automatically by machines, which functions by involuntarily identifying the desired speaker from a given set of speakers. The system might use a time-dependent spectral representation of the speech signal. One possible signal transformation would be an average spectra across an entire sentence, compare the average spectrum to a stored averaged spectrum template for each possible narrator, and then based on a spectral similarity measurement decide the individuality of the speaker, representing his uniqueness [1].

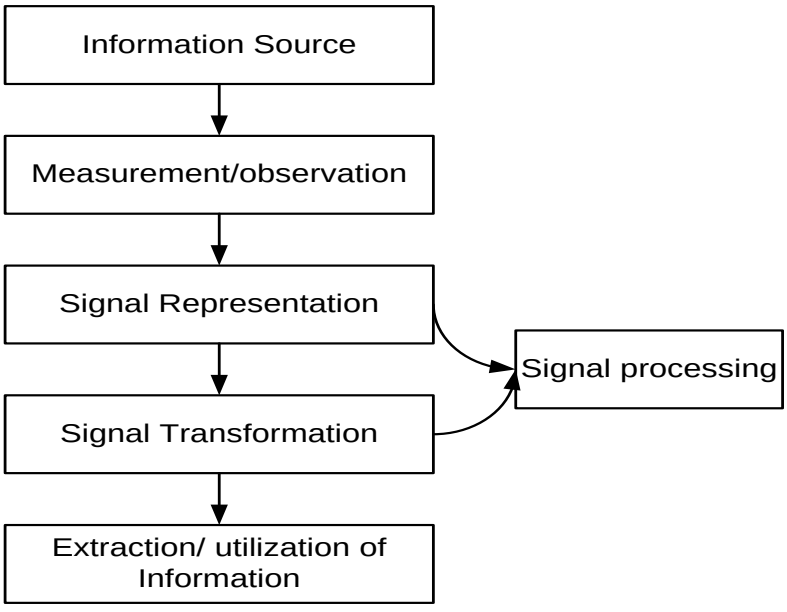

Fig. 1 Flowchart diagram for signal processing [1].

In a general picture we can define the processing of speech signal as the acquisition, management, storage, relocation and production of vocal utterances by a processor. The main applications are the recognition, synthesis and compression of human vocalizations. There has been an evolution in Human Computer Interaction (HCI) with the beginning of speech and language technologies facilitating humans to interact and carry out dialogue with computers [2]. Such inventions offer to a great level a healthy human machine interface making it possible for even a computer illiterate to handle a sophisticated environment, showing the intelligence of human mind and the extent to which it can reach in order to form an environment which is feasible for every person to work in. Such inventions and creation of models for speech synthesis for human computer interface has made work much easier and has also been able to reduce man labor to a great deal. Section II describes about speaker transformation, methodology of the investigations is explained in section III. The results and conclusions are presented in section IV and section V respectively.

II. SPEAKER TRANSFORMATION

It modifies the characteristics source speech signal to make it similar to that of target speaker [3]. In other words, transforming source speech parameters in such a way as if speech is pronounced by target speaker. Therefore establishing a transformation is the main thing [4]. Also the parameters modified/transformed for source and target are roughly classified as static and dynamic. Those natural speech parameters in which the speaker has least control such as vocal tract structure, natural pitch and are modeled correctly are called Static Parameters. Those speech parameters which are controlled by the speaker itself such as speaking style, emotions etc are called Dynamic Parameters. They are also termed as prosody of the speaker [5]. Fig. 2 illustrates the main steps involved in several speech conversion systems. The source speech must be a recorded speech while the target speech can be either recorded speech or set of parameters like pitch, formant frequencies, prosody etc. Thus in the analysis phase speaker individual parameters are extracted from source speech along with target speech. Comparatively the vocal tract parameters are considered to be more important as compare to the excitation source parameters as it helps in identifying the speaker individuality [6]. The pitch contour is supportive in the hint of speaker individuality [7]. In the mapping phase the parameters extracted of source speaker are mapped in such ways so that they come close to the target speech parameters. These parameters can be extracted from the target speech.

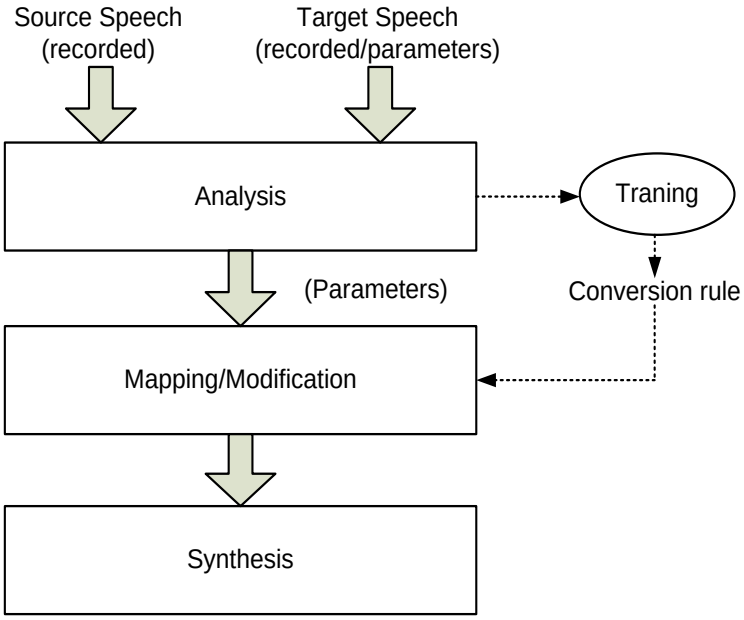

Fig. 2 General steps involved in speech transformation [5].

These parameters can also be made available directly or indirectly. This phase is also controlled by a conversion rule attained with the help of training phase. The parameters of the source spectrum are modified by this transformation function for obtaining the transformed speech. The speech is commonly represented using parameters such as formant frequencies [8], Cepstrum [9], Mel Frequency Cepstrum Coefficient (MFCC) [10], or Line Spectral Frequencies (LSF) [11]. Estimation of the transformation function from the source feature vectors to the corresponding target feature vectors are commonly based on Vector Quantization (VQ) [12], Artificial Neural Networks (ANN) [13], Gaussian Mixture Model (GMM) [14], Hidden Markov Model (HMM) [15]. In the synthesis block the modified parameters are used to synthesize or reconstruct the new target voice based speech. The emerging application of speech processing is speaker transformation or speech transformation. Applications of speaker transformation systems are applicable in numerous fields of automatic Speech-to-Speech translation, education, healthcare and entertainment [16].

III. METHODOLOGY

The comprehensive procedure for the analysis of the speech signals at phrase, word and phoneme has been divided into different steps. The analysis processes are carried out with the raw recordings of six speakers in Hindi language.

Speakers are of different age group, from different regions of Jammu. For recording purposes Sony recorder (ICD-UX513F) is used. It is a 4GB UX Digital Voice Recorder with expandable memory capabilities, and provides high quality voice recording. The study further includes the segmentation, labeling of the recorded speech, feature extraction and lastly alignment of source and target feature vectors using DTW technique and then calculating Mahalanobis distance between them. Fig. 3 shows sub-section of speech signal waveform and spectrogram labeled at word level. Feature extraction and lastly alignment of source and target feature vectors using DTW technique and then calculating Mahalanobis distance between them.

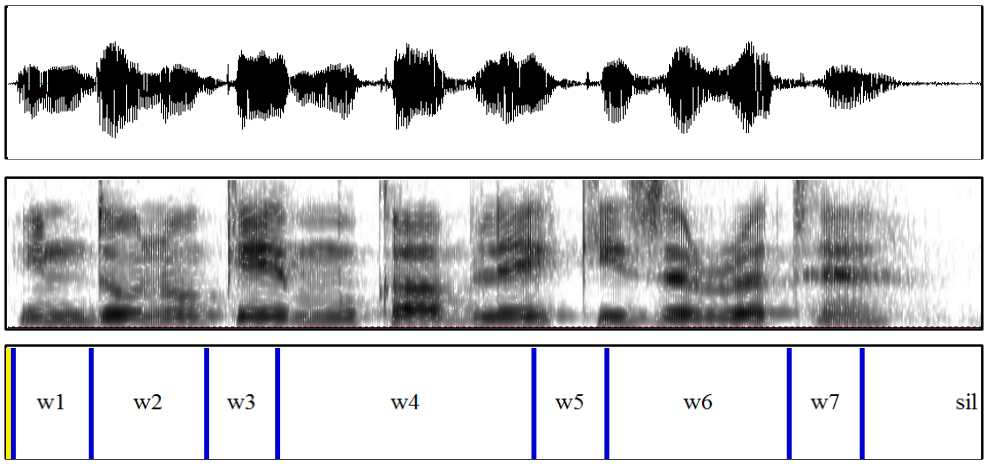

Fig. 3 Sub-section of speech signal waveform and spectrogram labeled at word level.

The main experiment part is separated into two main sections. In first section the analysis are performed without using HNM technique and in second section the analysis are performed using HNM technique. In the both steps features vectors are extracted using LSF technique and are aligned using DTW. Let source and target speech phrase are represented by *S1(a)* and *S2(b)* respectively.

In first section described in Fig. 4, at word level, segmented source and target speech, features vectors are extracted using LSF algorithms. The features extracted are aligned separately by means of DTW techniques and alignment error is calculated using Mahalanobis distance.

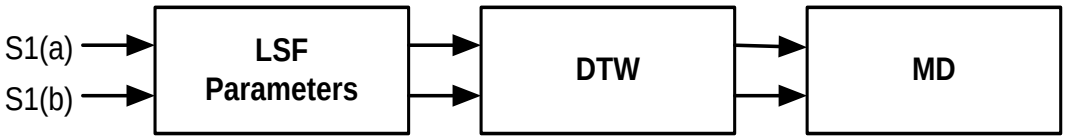

Fig. 4 Estimation of Mahalanobis distances.

In second section explained in Fig. 5, segmented source and target speech are analyzed first by HNM module in order to obtain the HNM parameters afterward LSF features are calculated. The extracted features are aligned by means of DTW techniques and alignment error is calculated using Mahalanobis distance.

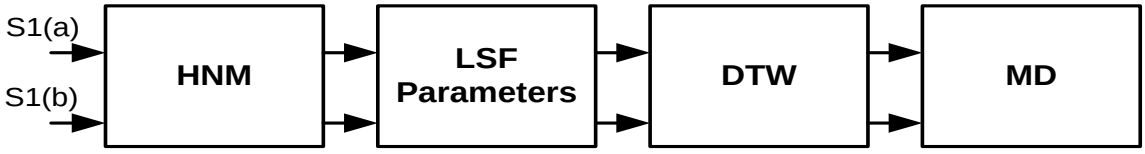

Fig. 5 Estimation of Mahalanobis distances using HNM parameters.

The above mentioned method is also implemented on the word and phoneme level segmentation of speech signal. The Mahalanobis distance and mean and standard deviation of alignment error are calculated in both the sections to interpret the results.

IV. RESULTS AND DISCUSSION

In order to examine the results based on the techniques namely LSF, HNM, and DTW for the alignment of segmented speech at phrase, word, and phoneme levels. These studies are carried out with male to male, female to male, and female to female speaker combinations. In order to compare alignment techniques, mean and standard deviation of Mahalanobis distances are calculated. Alignment error using Mahalanobis distances of various male-male combinations are

represented from Fig. 6 to Fig. 8. The various combinations of speaker pair's use as source and targets are written on the bottom of each plot, e.g. aman-w-nakul represents source target alignment without using HNM at word level while aman-wh-nakul represents source target alignment using HNM at word level. From the graphs it can be analysed that alignment error at all segmented level decreases by using HNM model, as Mahalanobis distances reduces. Hence using HNM model, alignment can more improved. Reduction in the Mahalanobis distances means improved alignment between the two sequences. On the whole comparisons of the speech alignment are shown in form of bar graphs in Fig. 9. It can be well anticipated that better speech alignment can be attained at phrase level rather than word level or phoneme level segmentation of speech.

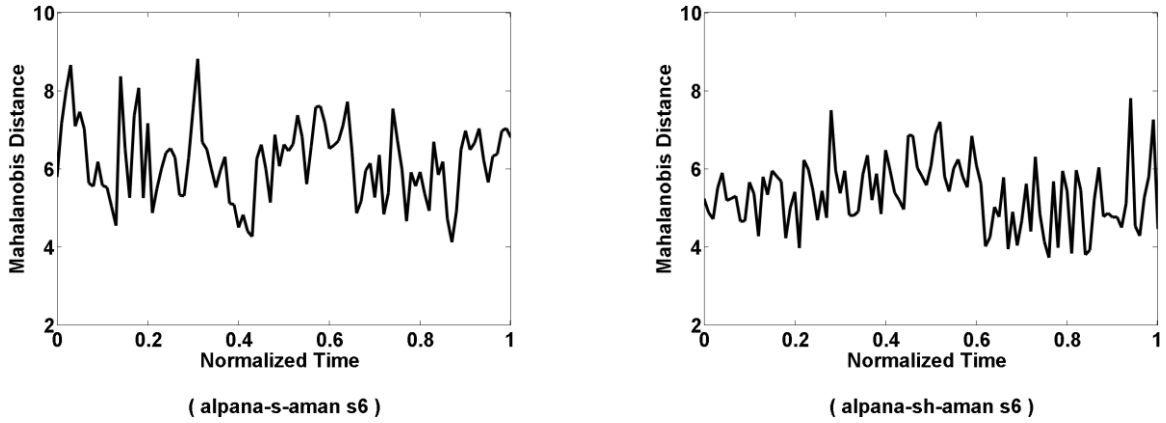

Fig. 6 LSF based speech alignment errors using Mahalanobis distance at sentence level with female to male combination.

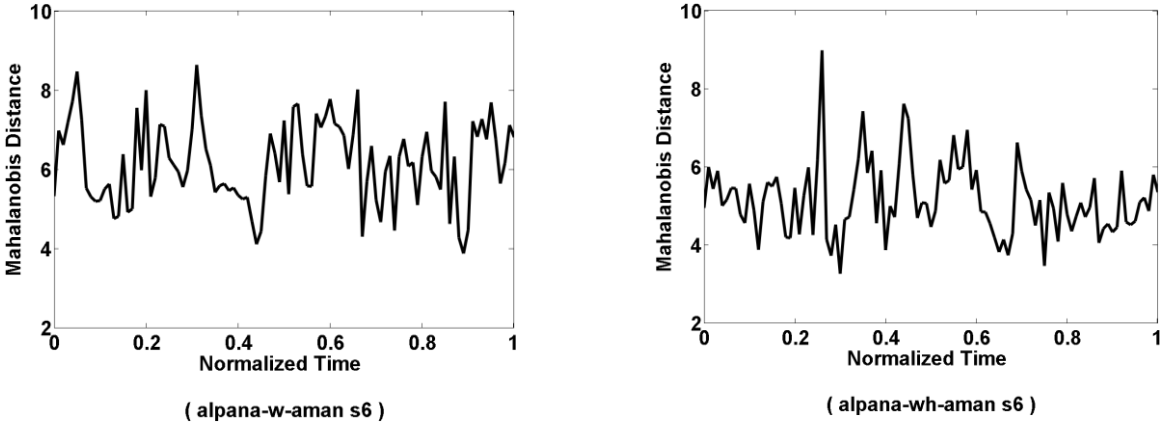

Fig. 7 LSF based speech alignment errors using Mahalanobis distance at word level with female to male combination.

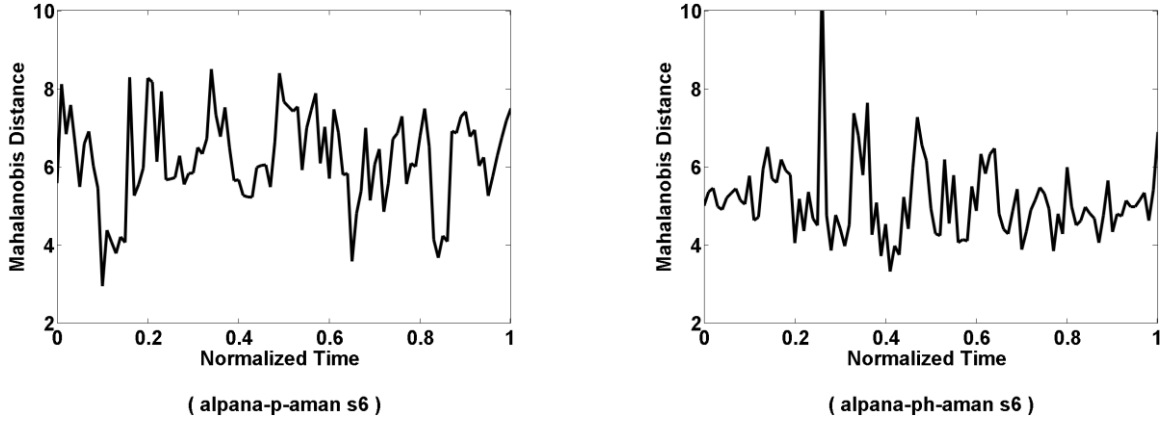

Fig. 8 LSF based speech alignment errors using Mahalanobis distance at phoneme level with female to male combination.

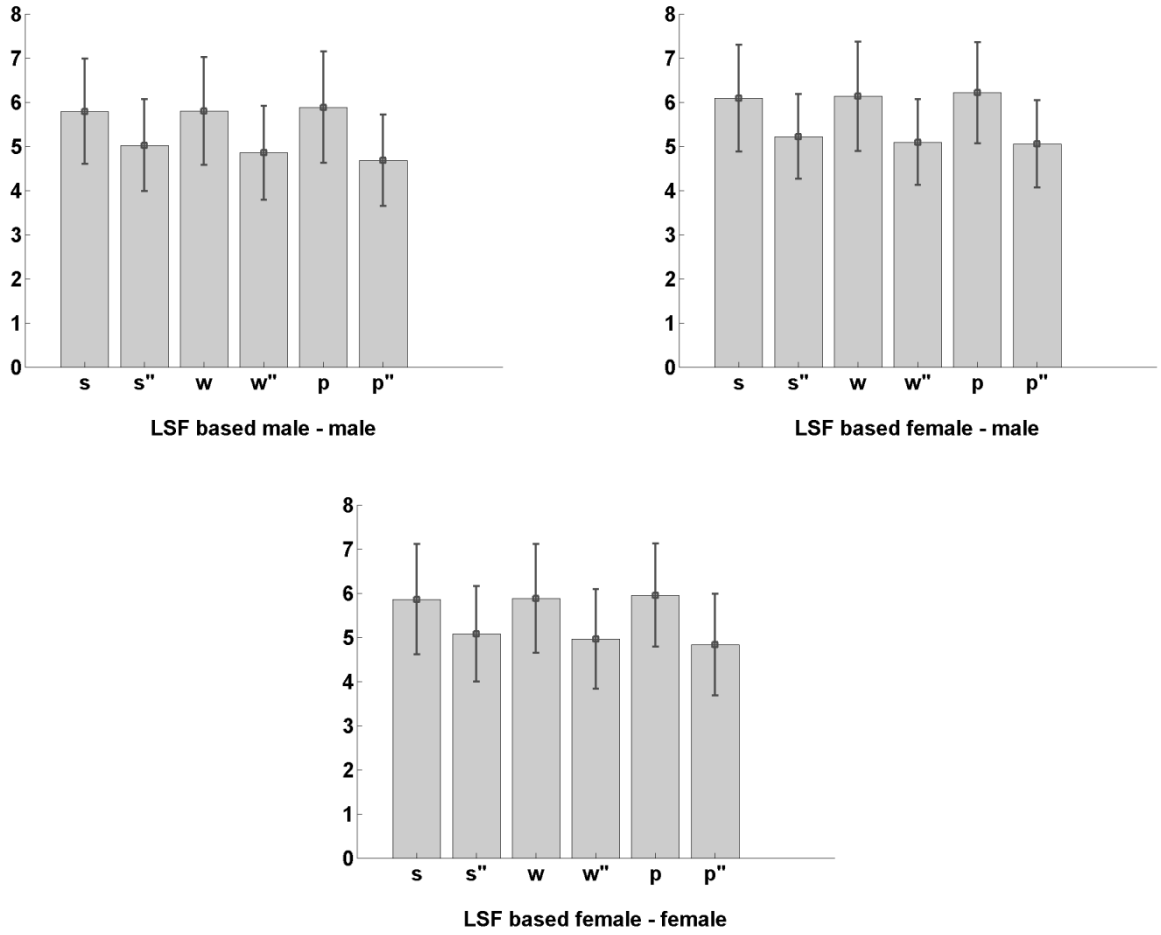

Fig. 9 Bar graph illustrate averaged LSF based mean along standard deviation speech alignment errors using Mahalanobis distance at sentence, word and phoneme level with all male to male, female to male and female to female combinations. The symbol s, w, p shows the results for alignment without the use of HNM based method and the symbol s", w", p" shows the results for HNM based method.

V. CONCLUSIONS

The recorded speech signals of six speakers are investigated with different combinations of speakers: male to male, female to male, and female to female. LSF and HNM techniques are used as feature vectors extraction of the recorded speech signals. Speech alignment error using Mahalanobis distances for labeled sentences at phrases, words, and phonemes levels are aligned by means of DTW are calculated. From the investigation of the results it can be concluded that alignment error using HNM model decreases at all the levels of labeled speech levels. Therefore implementing HNM model alignment error can be reduced in compression with the feature extraction method based on LSF only. Decrease in alignment error means better alignment between two speech signals.

Thus this research work is concluded as follows in brief:

1) HNM based alignment is more effective algorithm.
2) The accuracy of speech alignment cannot be significantly increased even labeling the speech signal at phoneme level. Valuable speech alignment can be achieved at phrase level, which saves our valuable time and avoidable step involved in the algorithms.
3) The alignment error in case of female-male is much larger than other combinations of male-male and female-female. Therefore such combination must be avoided in speech recognition and speaker transformation.

REFERENCES

[1] L. R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals*," Prentice Hall, Englewood Cliffs*, New Jersey, pp. 13, 1978.

[2] F. Alam, P. K. Nath and M. Khan, "Text To Speech for Bangla Language using Festial," *BRAC University*, Bangladesh, 2004-2007.

[3] B. Gillet, "Transforming Voice Quality and Intonation," M.S thesis, *Research university of Edinburgh*, 2003.

[4] H. D. Barrobes, "Voice Conversion applied to Text-to-Speech systems," Ph.D. Thesis, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, 2006.

[5] R. Shah and P. Vaya, "Development of a Voice Conversion System", Report, Department of Electrical Engineering, Electronics & Communication Engineering Branch, *Institute of Technology*, Ahmedabad, 2008.

[6] D. G. Childers, B. Yegnanarayana and W. Ke, "Voice conversion: factors responsible for quality," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florida, **1**, pp.748, 1985.

[7] H. Matsumoto, S. Hiki, T. Sone and T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," *IEEE transactions on audio and electroacoustics*, **21**, pp. 428, 1973.

[8] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectral tilt," *Speech Communication*, **16**, pp. 153, 1995.

[9] C. H. Wu, C. C. Hsia, T. H. Liu and J. F. Wang, "Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, **14**, pp. 1109, 2006.

[10] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, **6,** pp. 131, 1998.

[11] L. M. Arslan, "Speaker transformation algorithm using codebooks (STASC)," *Speech Communication.*, **28**, pp. 211, 1999.

[12] K. Shikano, K. Lee and R. Reddy, "Speaker adaptation through vector quantization," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Japan, **1**, pp. 2643, 1986.

[13] M. Narendranath, H. A. Murthy, S. Rajendran and B. Yegnanarayana "Transformation of formants for speaker transformation using artificial neural networks," *Speech Communication*, **16**, pp. 207, 1995.

[14] Y. Stylianou and O. Cappe, "A system for voice conversion based on probabilistic classification and a harmonic plus noise model," *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Seattle, **1**, pp. 281, 1998.

[15] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformation," *IEEE Trans. Audio, Speech, and Language Processing*, **14**, pp. 1301, 2006.

[16] H. D. Barrobes, "Voice Conversion applied to Text-to-Speech systems," Ph.D. Thesis, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona, 2006.

