Volume 3, Issue 7, July 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 Dynamic time Warping Based Alignment on the Accuracy of the Transformation Function for Voice Conversion

Radhika Khanna, Parveen Lehana
Department of Physics and Electronics,
University of Jammu, Jammu , India

Absract--- Voice conversion involves transformation of speaker characteristics in a speech uttered by a speaker called source speaker to generate a speech having voice characteristics of a desired speaker called the target speaker. Voice conversion is used in many applications namely dubbing, to enhance the quality of the speech, text-to-speech synthesizers, online games, multimedia, music, cross-language speaker conversion, restoration of old audio tapes, cellular applications, low bit-rate speech coding, etc. There are various models used for voice conversion such as Hidden Markov Model (HMM), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), and Vector Quantization (VQ). The quality and the identity conveyed by the transformed speech depend upon the accuracy of the transformation function derived from the given training data. The estimation of the transformation function requires properly alignedpassages spoken by source and target speakers. Exact alignmentof the corresponding speech units in the source and target passages is mandatory for the accurate estimation of the transformation function as the durations of speech units (i.e. phonemes or sub-phonemes) mayhave quite different distributions among speakers.Generally, DTW and VQ are used for this purpose. The objective of this paper is to compare the effectiveness of DTW and VQ based estimation of the transformation function. The analysis of the results shows that DTW provides about five percent more reduction in the transformed target distances of the speech. It means, DTW based technique is relatively better for the estimation of the transformation function.

Keywords--- HMM, ANN, DTW, VQ, Voice conversion.

I. INTRODUCTION

Speech is an important biomedical signal carrying two types of information. The first type of information is related to the message to be communicated and the second one carries the information about the identity of the speaker. The area of research dealing speech signals may be broadly divided into three sub fields: speech recognition, speech synthesis, and speech processing. The hot area of speech processing, i.e. voice conversion, is the modification of the speech of one speaker (called source speaker) into the speech of another speaker (called target speaker) [1]-[4]. Voice conversion is used in many applications namely dubbing, to enhance the quality of the speech, text-to-speech synthesizers, online games, multimedia, music, cross-language speaker conversion, restoration of old audio tapes, cellular applications, low bit-rate speech coding, etc.Voice conversion is carried out using a speech analysis-synthesis system, in which the parameters of the source speech are modified by a transformation function and resynthesis is carried out using the modified parameters. The transformation function is obtained by analyzing the aligned source and target speakers' utterances. Various techniques are used for voice conversion such as codebook based transformation [2], [3], dynamic frequency warping technique [5]-[7], speaker interpolation [8], ANN [9], Gaussian mixture models (GMMs) [10], [11], HMM [12], [13], and VQ [14].

Voice conversion technique involves five phases: alignment, feature extraction, source to target mapping (transformation function) estimation, source parameters transformation, and re-synthesis of speech from the transformed parameters. In alignment, the source and target passages are aligned in the same patterns of phonemes. The parameters related to vocal tract and excitation are estimated in the feature extraction phase. The transformation function is obtained from the parameters of the aligned passages and further used for transforming the source speech parameters. Finally, the transformed speech is resynthesised. The quality and the identity conveyed by the resynthesised speech depend upon the precise estimation of the transformation function, which is very difficult as there are many features of speech which are difficult to extract automatically, such as meaning of the passage and intention of the speaker.Although a good estimate of the transformation function may be obtained from the dynamics of the spectral envelopes of source and target speakers, the misalignment of the patterns of phonemes in the passages of the source and target may hamper the precise estimation[15], [16], [17]. The quality of the transformed speech depends upon various factors such as alignment of phonemes of the source and the target speaker, estimation of transition segments, estimation of transformation function, and resynthesis algorithm. The quality of the speech will be further deteriorated if the transition segments are not aligned properly.Exact alignmentof the corresponding speech

© 2013, IJARCSSE All Rights Reserved

units in the source and target passages is mandatory for the accurate estimation of the transformation function as the durations of speech units (i.e. phonemes or sub-phonemes) may have quite different distributions among speakers. Generally Dynamic time warping (DTW) is used for this purpose [18], [19]. DTW is pattern matching, dynamic programming based approach for finding an optimal distance between two given sequences wrapped in a non-linear fashion under certain restrictions; it is a well-established technique for time alignment and comparison of speech and image patterns [20]. For alignment, segmentation of the input speech in phonemes or diphones is carried out using HMM [21]-[24] and Viterbi algorithm [25] based methods. The state transition property in HMM-based methods presents a good approximation of the spectral envelope evolution along the time axis. The HMM is trained with the source and target speech data simultaneously. It models the probability distribution of the feature vector sequence according to its actual state sequence and the evolution of speech with transition probabilities between states.

Codebook mappingis also used to find the correspondence between acoustic classes of the source and target speakers [3] [26]. Tocreate smooth transition between neighboring frames, vector quantization may be used to extract codebooksfor aligned frames.In vector quantization, mapping functions are formed that represent correspondence between the acoustic space (vector space) of the source and the target speakers respectively [14]. These correspondences are formed by segmenting the spectral vectors of the source and target speakers into clusters using VQ based clustering [10]. For dividing the source feature vectors into m classes, m feature vectors are randomly selected as initial class centroid. Each source feature vector is assigned to one of the class based on the minimum distance from the class centroid. Mean of the vectors in each class is taken as the new class centroid and class membership of the feature vector is updated. The process repeated until the means stop changing. Grouping of the target feature vectors follows the grouping of the corresponding source vectors because the two are aligned. Generally, DTW and VQ are used for alignment as the third one, i.e. HMM requires a lot of data for training.The objective of this paper is to compare the effectiveness of DTW and VQ based estimation of the transformation function.The scope of the paper is limited to only objective evaluation of the closeness of the transformed speech to the target speech. For objective evaluation, Mahalanobis distance between the transformed and the target speech parameters have been used. Methodology of the investigations is described in the following section. The results and discussions are presented in Section III. The conclusion is given in Section IV.

II. MATERIAL AND METHODS

Methodology is divided into three phases: material used estimation of transformation function, transformation of the source speech and error estimation. These phases were described as follows.

A. Material Used

Speech data is required for both training and testing. Speech material was recorded from eight speakers (4 male and 4 female, ages: 20-23years). The male speakers are referred to as M1, M2, M3, and M4 and the females F1, F2, F3, and F4. The speakers in our experiment were university students of the same age group and had Hindi as their first language. It is desirable that the speakers belong to same group in terms of language to avoid accent related bias. The material was recorded in an acoustically treated room with 16 kHz sampling and 16-bit quantization rate.

B. Estimation of Transformation Function

The scheme for estimation of the transformation function is shown in Fig. 1. The recorded speech which is in the form of wave files is converted into mel frequency cepstral coefficient (MFCCs) speech vectors.MFCC representation is a beneficial approach for speech recognition [27]. The corresponding coefficients in source and target MFCCs have been reported to be correlated, and this property is very useful for using them in stochastic modeling [28], [29]. MFCCs are based on the knownvariation of the human ear's critical bandwidths withfrequency [30-32]. MFCC is perhaps the best known and most popular, and are more robust to background noise [33], [31]-[32], so we use MFCCs for our investigations.Then frame by frame alignment of the source and target speaker's parameters are done. The aligned source and target frames are used to estimate the transformation function.

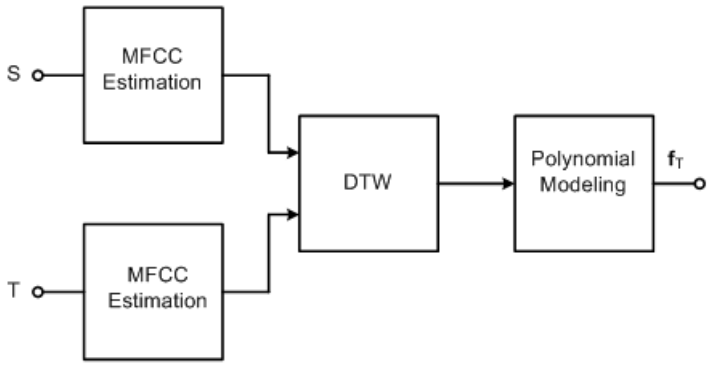


Fig. 1 Estimation of transformation function.

© 2013, IJARCSSE All Rights Reserved

For time alignment of the source and target frames, dynamic time warping (DTW) is used. Let the source and target feature vectors be represented by X of length n and Yof length m, respectively [34], where:

$$\mathbf{X} = [x_1, x_2, ...x_i, ..., x_n] \quad (1)$$

$$\mathbf{Y} = [y_1, y_2, ..., y_j, ...y_m] \quad (2)$$

A grid of an n x m matrix is constructed where the (i, j) element of the matrix contain the distance $d\left(p_i, q_j\right)$ between the two points x_i and y_j . At each grid point the absolute distance is calculated using Euclidean distance

$$d\left(x_i, y_j\right) = \left(x_i, y_j\right)^2$$

The value of i and j along the path define the time warping function between the source and target feature vectors. The optimal path in the (i, j) grid is searched using three constraints [35].

Boundary condition: starting and ending point of the warping path must be the first andthe last point of the aligned sequence.

Monotonicity condition: the time ordering of the points must be preserved.

Step size condition: this condition limits the warping path from long shifts in time.

From the aligned features vectors, the first MFCC coefficient representing the log energy is removed to avoid any biasing due to energy fluctuation. The feature vectors having vector distances less than a set threshold are removed from the corresponding positions of the aligned source and target feature vectors so that the number of feature vectors are reduced. For investigating the effect of DTW on the effectiveness of the transformation function, system is trained by 80% of aligned feature vectors and tested by remaining 20% of the aligned feature vectors. For investigating the effect of class based mapping (VQ) on the effectiveness of the transformation function, the feature vectors are divided into m discrete classes using K- means algorithm [36]. Although 21 distinct feature vectors are enough for deriving the mapping, we have investigated with m= 32, 64, 128, 256, and 512 to study the effect on the reduced distances between the transformed and actual target feature vectors. It may be noted that the input speech data has been taken limited only to estimate the minimal size of the required speech data. It will not have any significant effect on the boundaries of the acoustic classes as the transformation function will need only the centroids of the classes. The only constraint on the data may be that it should be phonetically balanced. Because of the use of linear mapping, the reduction of distance was not significant above 128 classes, whose centroids can be easily estimated by using about 20 phonetically balanced long sentences. Feature vectors nearest to the centroid of each class are taken and these limited feature vectors are used to train the system. For testing the system, the aligned feature vectors are used. We also investigate the effect of the number of coefficients of MFCCs on the transformation function by taking the different values of coefficients as 13, 21, and 50. Polynomial modeling is used for deriving the mapping from the acoustic space of the source speaker to that of the target speaker. Let the p-dimensional feature vectors of source and target are represented by $x = [x_1 x_2 x_3 .....x_p]$ and $y = [y_1 y_2 y_3 .....y_p]$, respectively. In our experiment, the transformation function is estimated using multivariate linear modeling (MLM). In MLM each element of the target feature vector is assumed to be linear function of all elements in the source feature vectors,

$$y_i = f_i[x_1, x_2, ...x_i, ..., x_p], \quad (3)$$

$$y_i = c_{0,i} + c_{1,i}x_1 + c_{2,i}x_2 + ... + c_{n,i}x_{p,} \quad (4)$$

If a multidimensional function g is known at q points, a multivariate polynomial surface f can be constructed such that it approximates the given function within some error at each point [4], [17], [37]-[41]

$$g({}^n w_1, {}^n w_2, \cdots, {}^n w_m) = f({}^n w_1, {}^n w_2, \cdots, {}^n w_m) + \varepsilon_n, \quad 0 \le n \le q-1 \quad (5)$$

The multivariate function can be written as

$$f(w_1, w_2, \cdots, w_m) = \sum_{k=0}^{p-1} c_k \phi_k(w_1, w_2, \cdots, w_m) \quad (6)$$

where p is the number of terms in the polynomial of m variables. By combining (5) and (6), we get a matrix equation

$$\mathbf{b = Az + \varepsilon} \quad (7)$$

where vectors b, z, and ε are given by

© 2013, IJARCSSE All Rights Reserved

$$\mathbf{b}^{\mathrm{T}} = [g_0 \quad g_1 \quad \cdots \quad g_{q-1}]$$

$$\mathbf{z}^{\mathrm{T}} = [c_0 \quad c_1 \quad \cdots \quad c_{p-1}]$$

$$\boldsymbol{\varepsilon}^{T} = [\varepsilon_0 \quad \varepsilon_1 \quad \cdots \quad \varepsilon_{q-1}]$$

Matrix A is a q × p matrix, with elements given as

$$a(n,k) = \phi_k({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m), \quad 0 \le n \le q-1 \text{ and } \quad 0 \le k \le p-1$$

If the number of data points is greater than the number of terms in the polynomial ($q \ge p$), then coefficients ck's can be determined for minimizing the sum of squared errors

$$E = \sum_{n=0}^{q-1} \left[g({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) - f({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) \right]^2 \tag{8}$$

and we get the solution

$$\mathbf{z} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \tag{9}$$

where $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is known as pseudo–inverse of A [17], [42].

We model the mapping between the acoustic space of source and target feature vectors using MLM as it is a linear function the transformation is very smooth with less fluctuation, less complexity and required less memory. Each component in the target feature vector is modeled as a multivariate linear function of all the component in the source vector and is represented as

$$y_i = f_i[x_1, x_2, \ldots. x_i, \ldots\ldots, x_{M-1}], \qquad 1 \le i \le M-1$$

C. Transformation of Source Speech Parameters and Error Estimation

The scheme for transformation of the source speech to the target speech is shown in Fig. 2.The source speech is converted into MFCCs feature vectors. The spectral parameters MFCCs are transformed using the transformation function, obtained in the transformation function estimation block, for the given speaker pair.

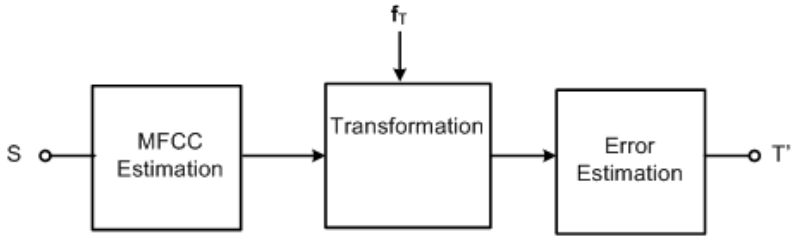


Fig. 2 Transformation of source speech parameters and error estimation

The error is estimated by finding the percentage of reduction in the spectral distance [4]. The reduction in the spectral distance is carried out by calculating the cepstral Mahalanobis distance [17], [43-48]. The distance between the target frames and the source frames is calculated as target-source distance ST. The distance between the target frames and the transformed frames is calculated as target transformed distance TT'. The distances were averaged across the frames in each of the test set of utterances. The relative decrease in the distance, i.e. (ST-TT'))/(ST) is taken as a measure of decrease in the distance between spectral envelopes.

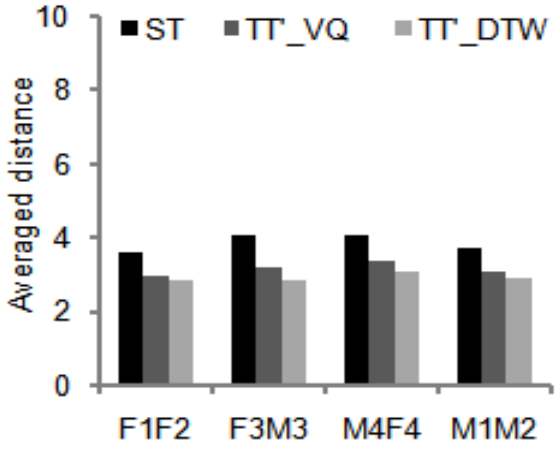
III. RESULTS AND DISCUSSION

Using the technique described in the methodology, eight transformation functions were estimated from the MFCCs derived from 20utterancesof aligned source and target speech material. The first four of these transformation functions were derived using DTW based alignment and the other four were estimated using VQ based algorithm. Four different pairs (F1-F2, F3-M3, M4-F4, and M1-M2) were taken for estimating the transformation functions. The accuracy of the transformation functions was assessed objectively using five utterances different from the 20 utterances used for training. The closeness of the transformed feature vectors to the actual target feature vectors was quantified using Mahalanobis distance. The mean distance between the source and target (ST), target to transformed speech TT' is shown in Fig. 3 as histograms.In Fig. 3(a), 3(b) and 3(c), histograms are shown for MFCCs having 13 coefficients, 21 coefficients and 50 coefficients respectively. For

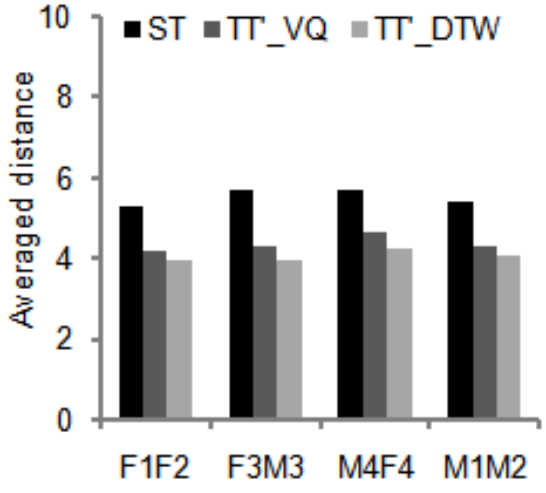
© 2013, IJARCSSE All Rights Reserved

example for MFCC having 21 coefficients, reduction in mean distance using DTW between ST and TT' is from 5.3 to 4.0 in case of F1-F2,4.1 to 2.9 in case of F3-M3, 4.1 to 3.1 in case of M4-F4, 3.7 to 2.9 in case of M1-M2 and using VQ for 512 classes the decrease in mean distance between ST and TT' is from 3.6 to 3.0 in case of F1-F2,4.1 to 3.2 in case of F3-M3, 4.1 to 3.4 in case of M4-F4, 3.7 to 3.1 in case of M1-M2.The DTW shows more reduction in the mean distance between the source and target ST, target to transformed speech TT' than VQ and reduction in mean is larger for cross gender conversion as compared to same gender conversion.

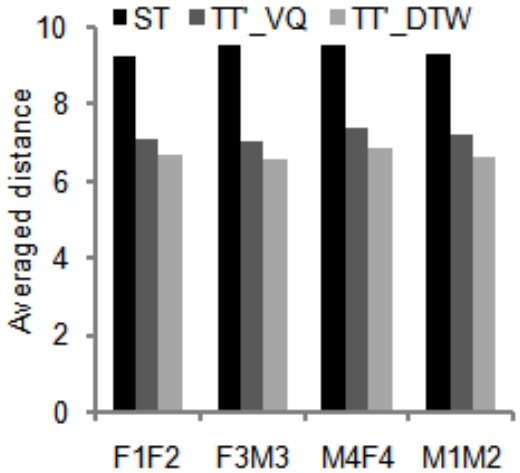
Table1 shows the percentage of reduction in distance between the source and target, target to transformed speech. Although the reduction in TT' increases with the number of MFFC coefficient, informal listening tests showed that there is no significant improvement in the quality or identity of the transformed speech beyond 21 coefficients. To investigate the effect of DTW based alignment on the transformation of different phonemes, Mahalanobis distance was estimated for each of the phonemes in a set of five sentences for each speaker pair. Figure 4 shows the distances for an utterance of a female speaker to be transformed into male speech. Analysis of the frame wise reduction showed that it is phoneme dependant. For example for sentence "धोबिन जब सोकर उठती ता`देखती कि चौका साफ पडा है और बर्तन मझे हुए है" phoneme wise reduction were carried out. It was observed that for the phoneme /ठ/, there is the minimum reduction of about 18% and for /ह/ phoneme there is maximum reduction i.e of 44.4%. It was found that only the phonemes /ठ/ and /प/ showed minimum closeness towards the target.



(a)



(b)



(c)

Fig. 3 The mean distance between ST and TT'(a) for 13 MFCCs coefficients (b) for 21 MFCCs coefficients (c) for 50 MFCCs coefficients

© 2013, IJARCSSE All Rights Reserved

TABLE I.:PERCENTAGE REDUCTION OF THE MEAN DISTANCE BETWEEN ST AND TT'

Technique	Speaker pair	Reduction (%) for order		
		13	21	50
DTW	F1F2	21.8	24.7	30.0
	F3M3	29.7	31.5	31.5
	M4F4	24.7	26.2	30.0
	M1M2	22.6	24.2	28.8
VQ	F1F2	17.6	21.3	23.3
	F3M3	22.1	24.5	26.1
	M4F4	17.2	19.0	22.1
	M1M2	18.3	20.2	23.0

IV. CONCLUSIONS

Investigations were carried out to study the effect of DTW and VQ based transformation function on the closeness of the transformed speech to the target speech using multivariate linear mapping between the acoustic spaces of the source and the target speakers. The analysis of the results shows that DTW provides about five percent more reduction in the transformed target distances of the speech. It means, DTW based technique is relatively better for the estimation of the transformation function. Subjective evaluation of the transformed speech using ABX test is on our future plan.

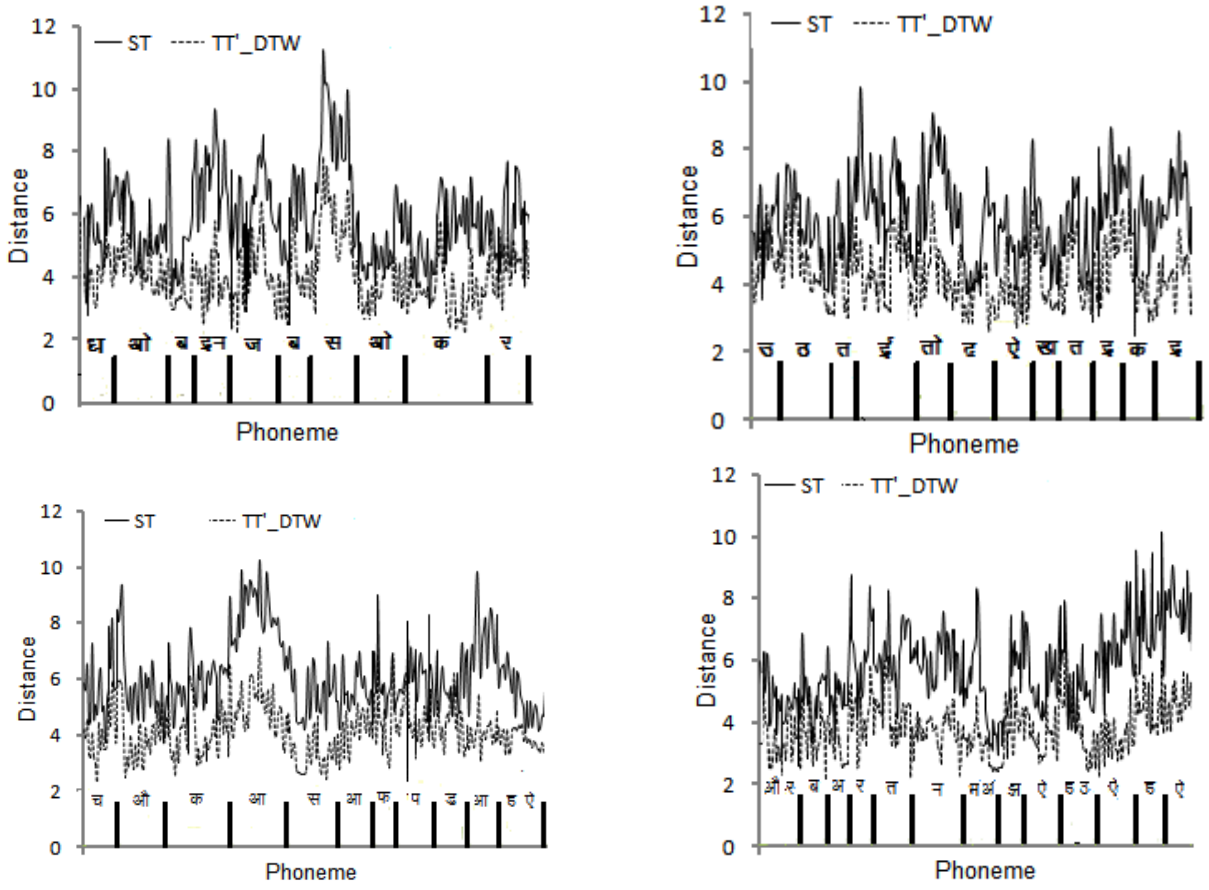


Fig. 4 Distances for an utterance of a female speaker to be transformed into male speech

REFERENCES

[1] E. Moulines and Y. Sagisaka, *Speaker Transformation State of the Art and Perspectives*. Netherlands: Elsevier, 1995.

[2] H. Mizuno and M. Abe, "Voice conversion based on piecewise linear conversion rules of formant frequency and spectrum tilt," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 1, pp. 469–472.

© 2013, IJARCSSE All Rights Reserved

[3] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Commun*., no. 28, 1999.

[4] P.K .Lehana, P.C. Pandey, "Transformation of short-term spectral envelope of speech signal using multivariate polynomial modeling", *in Proc. National Conference on Communication*, 2011 , pp. 1-5.

[5] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Commun*., vol. 1, pp. 145–148, 1992.

[6] D. Rentzos, S. Vaseghi, Q. Yan, and C. H. Ho, "Voice conversion through transformation of spectral and intonation features," *in Proc IEEE Int. Conf. Acoust., Speech, Signal Process*., 2004, vol. 1, pp. 21–24.

[7] Z. W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," *in Proc. Int.Conf. Spoken Lang. Process*., 2006.

[8] N. Iwahashi and Y. Sagisaka, "Speech spectrum transformation by speaker interpolation," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 1, pp. 461–464.

[9] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun*., vol. 16, no. 2, pp. 207–216, 1995.

[10] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Paris, France, 1996.

[11] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process*., 1998, vol. 6, pp.131–142.

[12] K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMS with dynamic features," *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process.*, 1996, pp.389–392.

[14] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., "Voice conversion through vector quantization," *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process,* 565-56o8.

[15] W. Endres, W. Bambach, and G. Fl¨osser, "Voice spectrograms as a function of age, voice disguise, and voice imitation," *J. Acoust. Soc. Amer*., vol. 49, pp. 1842–1848, 1971.

[16] M. R. Sambur,"Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, 1975.

[17] P. K. Lehana, " Spectral mapping using multivariate polynomial modeling fro voice conversion" , Ph. D. thesis, Department of Electrical Engineering, IIT Bombay, 2012.

[18] F. Itakura , "Line spectrum representation of linear predictor coefficients of speech signals", *J. Acoust. Soc. Amer*, vol. 57, S35 (A), 1975b.

[19] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", *Trans. IEEE Acoust., Speech, Signal Process*., vol. 26, no.1, pp. 43– 49,1978.

[20] M. A. Al-Manie, M. I. Alkanhal, and M.M. Al-Ghamdi, "Arabic speech segmentation:Automatic verses manual method and zero crossing measurements," *Indian Journal of Science and Technology*, vol. 3, no. 12, 2010.

[21] Arslan, L. M. and Talkin, D., "Speaker transformation using sentence HMM based alignments and detailed prosody modification", *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process,*, 1998, 1: 289-292.

[22] L. R. Rabiner, "A tutorial on Hidden markov models and selected applications in speech recognition", in *Proc. of the IEEE*, vol 77 (2), pp. 257-286, 1989.

[23] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in speaker transformation systems," *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process.*, 2004, pp. 5-8.

[24] A. Kain and M. Macon, "Spectral speaker transformation for text-to-speech synthesis," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 285-288.

[25] L. Rabiner, B. H. Juang, *Fundaments of speech recognition*. Englewood Cliffs: Prentice Hall, 1999.

[26] D.Sündermann, A.Bonafonte, H.Ney, H.Höge, "A first step towards text-independent voice conversion", *in Proc. Int. Conf. Spoken Lang. Process*., pp.1173-1176, 2004.

[27] S.Davis and P. Mermelstein , "Comparison of Parametric Representation for Monosyllable Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustic, Speech, Signal Process.*, 28(9): 357-366.

[28] Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, Vol. 6, 1998, pp. 131-142.

[29] E. Helander, J. Nurminen, M. Gabbouj, "LSF mapping for voice conversion with very small training sets," *in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process*, 2008, Las Vegas, Nevada, pp.4669-4672.

[30] John R. Deller Jr., John H. L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals", IEEE Press, New York, 2000.

[31] F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, "AVector Quantization Approach to Speaker Recognition", AT&T Technical Journal, vol. 66, pp.14-26, 1987.

[32] S. M. Kamruzzaman, A. N. M. RezaulKarim, Md. Saiful Islam, Md. EmdadulHaque, "Speaker Identification using MFCC-Domain Support Vector Machine" *int. J. Elect., Power engg*., vol. 1, no. 3, pp. 274-278, 2007.

[33] Lawrence Rabiner and Biing-Hwang Juang, *Fundamental of Speech Recognition*, Prentice-Hall,Englewood Cliffs, N.J., 1993.

© 2013, IJARCSSE All Rights Reserved

[34] S.Salvador and P. Chan," Toward accurate dynamic time warping in linear time and space," *Intell.Data Anal.*, vol. 11, no. 5, pp. 561-580, 2007.
[35] O. Turk and L. M. Arslan, "Robust processing techniques for voice conversion," *Comput.Speech Lang.*, vol. 20, no. 4, pp. 441–467, 2006.
[36] D.Mackay, *An example inference task in information theory, inference and learning algorithm*, Cambridge University press: pp. 284-292, 2003.
[37] J. M. D. Pereira, P. M. B. S. Girão, and O. Postolache, "Fitting transducer characteristics to measured data," *IEEE Instrum. Meas. Mag.*, vol. 4, no. 4, pp. 26–39, 2001.
[38] G. M. Philips, "Interpolation and Approximation by Polynomials", New York: Springer-Verlag, 2003.
[39] V. Pratt, "Direct least-squares fitting of algebraic surfaces," *Computer Graphics*, vol. 21, no. 4, pp. 145–152, 1987.
[40] P. C. Pandey and M. S. Shah, "Estimation of place of articulation during stop closures of vowel–consonant–cowel utterances*," IEEE Transactions on Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 277–286, 2009.
[41] R. Vergin, D. O'Shaughnessy, and V. Gupta, "Compensated mel frequency cepstrum coefficients," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*,vol. 1, 1996, pp.323-326.
[42] R. L. Branham Jr., Scientific Data Analysis, "An Introduction to over determined Systems" New York: Springer-Verlag, 1990.
[43] R. Curtin, N. Vasiloglou, D.V. Anderson, "Learning distances to improve phoneme classification," *in Proc. Int. Workshop on Machine Learning for Signal Process*., 2011, Beijing, China, pp. 1-6.
[44] R.E. Donovan, A new distance measure for costing spectral discontinuities in concatenative speech synthesisers, *in Proc. ISCA Tutorial Research Workshop Speech Synthesis*, 2001, Perthshire, Scotland.
[45] T. Takeshita, F. Kimura, and Y. Miyake, "On the Estimation Error of Mahalanobis Distance," *Trans. IEICE*, vol. J70-D, no. 3, pp. 567-573, Mar. 1987
[46] T. Takeshita, S.Nozawa, and F. Kimura, "On the bias of Mahalanobis Distance due to limited sample size effect*," in proc. 2nd IEEE int. conf. computer cardiology Trans.*, vol. J70-D, no. 3, pp. 567-573, Mar. 1987.
[47] J.C.T.B. Moraes, M.O. Seixas, F.N. Vilani, and E.V.Costa, "Arealtime QRS complex classification method using Mahalanobis distance," *in Proc. IEEE Int. Conf. Computer Cardiology*, 2002, pp.201-204.
[48] T. Kamei, "Face retrieval by an adaptive Mahalanobis distance using a confidence factor," *in Proc. IEEE Int. Conf. Image process.*, pp.153-156.
[49] G. Chen, H. G. Zhang, and J . Guo," Efficient computation of Mahalanobis distance in financial hand-written Chinese character recognition," *in Proc. IEEE Int. Conf. Machine learning and cybernetics*, 2007, vol. 4, pp.2198-2201.

© 2013, IJARCSSE All Rights Reserved