International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

To Investigate the Accuracy of the Dynamic Time Warping Based Transformation Function for Voice Conversion

Radhika Khanna[1], Parveen Lehana[2]

DSP Lab, University of Jammu, Jammu

pklehanajournals@gmail.com

Abstract

Voice conversion involves transformation of speaker characteristics in a speech uttered by a speaker called source speaker so as to generate a speech having voice characteristics of a desired speaker called target speaker. Voice conversion technology 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, and low bit-rate speech coding. There are various models used for voice conversion such as Hidden Markov Model (HMM), Artificial Neural Network (ANN), and Dynamic Time Warping (DTW) based. The quality of transformed speech depends upon the accuracy of the transformation function. For obtaining an accurate transformation function, the alignment of the passages spoken by source and target speakers should be properly aligned. The objective of the paper is to investigate the effect of DTW based transformation function estimation on the closeness of the transformed speech towards the target.

***Keywords*-- HMM, DTW, ANN, speech synthesis, voice conversion.**

I. Introduction

Speech signal is a very important biomedical signal which carries two type 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. Voice conversion is a technique of modifying a speech uttered by a source speaker into the voice of target speaker [1]. Voice conversion technology 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, and 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 modified parameters. The transformation function is obtained by analyzing the aligned source and target speaker's utterances. Precise estimation of the transformation function 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.

Various techniques are used for voice conversion such as codebook based transformation [2-4], dynamic frequency warping technique [5-7], Speaker interpolation [8], artificial neural networks [9], Gaussian mixture models (GMMs) [10-12], Hidden Markov Models (HMMs) based [13], Vector quantization based [14] and so on.

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 synthesised. The quality of the synthesised speech depends 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 [15-16].

The quality of transformed speech depends upon the accuracy of the transformation function. For obtaining an accurate transformation function, the alignment of the passages spoken by source and target speakers should be properly aligned. Alignment is necessary to determine corresponding units in the source and target voices. This is due to the fact that the durations of sound units (i.e. phonemes or sub-phonemes) can be quite different among speakers. Generally DTW is used for this purpose [22-23]. 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 [25]. Let the source and target feature vectors be represented by X of length n and Y of length m, respectively [24], 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

$$d\left(x_i, y_j\right) = \left(x_i, y_j\right)^2 \quad (3)$$

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: boundary condition, monotonicity condition, and step size condition [23]. In boundary condition, starting and ending point of the warping path must be the first and the last point of the aligned sequence. Monotonicity condition preserves the time ordering of the points. Step size condition limits the warping path from long shifts in time.

The objective of the paper is to investigate the effect of DTW based transformation function estimation on the closeness of the transformed speech towards the target. 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. Methodology

Methodology is divided into three phases: recording, estimation of transformation function, transformation of the source speech parameters and error estimation.

Recording

Speech data is required for both training and testing. Speech material was recorded from eight speakers (4 male and 4 female, ages: 20-23 years). The male speakers are referred to as M1, M2, M3, and M4 and the female as F1, F2, F3, and F4. Twenty-five utterances were recorded from each speaker. 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.

Estimation of transformation function

The recorded speech which is in the form of wave files is converted into mel frequency cepstral coefficient (MFCCs) speech vectors. Davis & Mermelstein (1980) pointed out that MFCC representation is a beneficial approach for speech recognition [31]. 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 [33], [34]. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency [28-30]. MFCC is perhaps the best known and most popular, and are more robust to background noise [27], [29], [30], so we use MFCCs for our investigations.

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,

$$y_i = f_i(x_1, x_2, ...x_i, ..., x_p) \quad (4)$$

$$y_i = c_{0,i} + c_{1,i}x_1 + c_{2,i}x_2 + ... + c_{n,i}x_{p,} \quad (5)$$

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 [17] [18] [19] [20] [21][26]

$$g({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) = f({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) + \varepsilon_n$$
$$\text{where} \quad 0 \le n \le q-1 \qquad (6)$$

The multivariate function can be written as

$$f(w_1, w_2, \cdots, w_m) = \sum_{k=0}^{p-1} c_k \phi_k (w_1, w_2, \cdots, w_m)$$
(7)

Where p is the number of terms in the polynomial of m variables. By combining (6) and (7), we get a matrix equation

$$\mathbf{b} = \mathbf{A}\mathbf{z} + \boldsymbol{\varepsilon}$$

Where vectors **b**, **z**, and **ε** are given by

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

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

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

Matrix **A** is a $q \times p$ matrix, with elements given as

$$a(n,k) = \phi_k ({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m)$$

Where $0 \le n \le q-1$ and $0 \le k \le p-1$

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

$$E = \sum_{n=0}^{q-1} \begin{bmatrix} g({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) \\ -f({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) \end{bmatrix}^2$$

and we get the solution

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

Where $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is known as pseudo–inverse of **A** [21].

Table I.
Averaged Mahalanobis distances between different speaker pairs.

Speaker pair	ST	TT'	Reduction (%)
F1F2	3.4	2.7	20.6
F3M3	4.2	2.7	35.7
M4F4	4.1	2.8	31.7
M1M2	3.7	2.7	27.0

Transformation of the source speech parameters and error estimation

The source speech is converted into MFCCs feature vectors. The spectral parameters MFCCs are transformed using the transformation function, (5) obtained in the transformation function estimation block, for the given speaker pair.

The error is estimated by finding the percentage of reduction in the spectral distance [26]. The reduction in the spectral distance is carried out by calculating the cepstral Mahalanobis distance [34-35]. 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*) are taken as a measure of decrease in the distance between spectral envelopes.

III. Results And Discussions

Using the technique described in the methodology, four transformation functions were estimated from the MFCCs derived from 20 utterances of aligned source and target speech material. Four different pairs (F1F2, F3M3, M4F4, and M1M2) were taken (6) 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 (7) was quantified using Mahalanobis distance.

The mean distance between the source and target (*ST*), target to transformed speech (*TT'*) is shown in Table I. In case of female to female transformation the original source to target distance is 3.4 and the corresponding distance between the target and the transformed speech is 2.7 giving a reduction of about 20.6%.

Machine Intelligence & Advancement Group (MIRAG)

IJETAE EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 2, January 2013)

National conference on Machine Intelligence Research and Advancement (NCMIRA, 12), INDIA.

In case of female to male transformation the original source to target distance is 4.2 and the corresponding distance between the target and the transformed speech is 2.7 giving a reduction of about 35.7%. In case of male to female transformation the original source to target distance is 4.1 and the corresponding distance between the target and the transformed speech is 2.8 giving a reduction of about 31.7%. In case of male to male transformation the original source to target distance is 3.7 and the corresponding distance between the target and the transformed speech is 2.7 giving a reduction of about 27.0%. It may be observed that the reduction for cross gender conversion is maximum.

Analysis of the frame wise reduction showed that it is phoneme dependant. For example for sentence “/kksfcu tc lksdj mBrh rks ns[krh fd pkSdk lkQ iMk gs vkSj crZu e>s gq, gs“ phoneme wise reduction is shown in TABLE II. For the phoneme /B/, there is the minimum reduction of about 18% and for /g/ phoneme there is maximum reduction i.e of 44.4%.

TABLE II.
Phoneme wise averaged Mahalanobis distances.

Phonemes	ST	TT’	Reduction (%)
v	5.1	3.7	27.4
vk	5.7	3.9	31.6
g	7.0	4.1	41.4
bZ	6.2	4.5	27.4
m	5.5	4.0	27.2
,	6.7	3.7	44.8
,s	5.9	4.0	32.2
vks	6.4	4.2	34.4
vkS	5.1	3.5	31.4
d	5.3	3.8	28.3
[k	5.3	4.1	22.6
p	5.8	4.1	29.3
>	6.0	4.1	31.7
t	5.8	3.9	32.7
B	6.1	5.0	18.0
M	6.0	4.6	23.3
r	5.9	4.2	28.8
n	5.1	3.8	25.5
/k	5.3	3.8	28.3
i	5.9	4.7	20.3
Q	5.7	4.2	26.3
c	5.1	3.6	29.4
u	5.9	3.7	37.3
e	5.3	3.3	37.7
j	5.5	4.0	27.8
l	8.2	5.3	35.4
g	7.2	4.0	44.4

IV. CONCLUSIONS

Investigations were carried out to study the effect of DTW 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 target speakers. The analysis of the results showed that the transformation function may be satisfactorily estimated after aligning the source and target utterances with the help of DTW. It was found that only the phonemes /B/ and /i/ showed minimum closeness towards the target. Subjective evaluation of the transformed speech using ABX test is on our future plan.

REFERENCES

[1] E. Moulines and Y. Sagisaka, Eds., 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.

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

[4] O. Turk and L. M. Arslan, “Robust processing techniques for voice conversion,” Comput. Speech Lang., vol. 20, no. 4, pp. 441–467, 2006.

[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, École Nationale Superieure Des Télécommunications, 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.

[13] M. Abe, S Nakamura, K Shikano, H. Kuwabara, “Voice conversion through vector quantization” in proc. IEEE ICASSP, 1988, p.p 565-568.

[14] 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.

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

[16] 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.

[17] G. M. Philips, "Interpolation and Approximation by Polynomials", New York: Springer-Verlag, 2003.

[18] V. Pratt, "Direct least-squares fitting of algebraic surfaces," Computer Graphics, vol. 21, no. 4, pp. 145–152, 1987.

[19] 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, and Language Processing, vol. 17, no. 2, pp. 277–286, 2009.

[20] R. L. Branham Jr., Scientific, " Data Analysis:An Introduction to Overdetermined Systems", New York: Springer-Verlag, 1990.

[21] F. Itakura , "Line spectrum representation of linear predictor coefficients of speech signals", Journal of Acoust. Soc. of America, vol. 57, S35 (A), 1975b.

[22] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no.1, pp. 43–49,1978.

[23] 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.

[24] 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 Dec 2010.

[25] P.K . Lehana, P.C. Pandey, " Transformation of short-term spectral envelope of speech signal using multivariate polynomial modeling", National Conference on Communications (NCC), 2011 , pp. 1-5.

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

[27] Jr., J. D., Hansen, J., and Proakis, J., " Discrete-Time Processing of Speech Signals", second ed. IEEE Press, New York, 2000.

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

[29] S.M. Kamruzzaman, A. N. M. Rezaul Karim, Md. Saiful Islam, Md. Emdadul Haque, "Speaker Identification using MFCC-Domain Support Vector Machine" CoRR abs/1009.4972, 2010).

[30] Davis, S. & Mermelstein, P. (1980), "Comparison of Parametric Representation for Monosyllable Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustic, Speech and Signal Processing, pp. 357-366.

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

[32] E. Helander, J. Nurminen, M. Gabbouj, "LSF mapping for voice conversion with very small training sets", in Proc. ICASSP 2008, Las Vegas, Nevada, pp. 4669-4672.

[33] 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.

[34] R.E. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers", in Proc. ISCA Tutorial Research Workshop Speech Synthesis 2001, Perthshire, Scotland, paper no.: 123.