To Investigate The Accuracy of The Vector Quantization Based Transformation Function For Voice Conversion

Radhika Khanna, Parveen Lehana*
University of Jammu, Jammu

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. There are various models used for voice conversion such as hidden Markov model (HMM), artificial neural network (ANN), vector quantization (VQ) 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. These correspondences are formed by segmenting the spectral vectors of the source and target speakers into clusters using VQ-based clustering. VQ reduces the computation amount and memory size drastically. The objective of the paper is to investigate the effect of VQ based transformation function estimation on the closeness of the transformed speech towards.

1. Introduction

A speech signal consists of two main parts: one carries the speech information, and the other 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]-[4]. 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], [3], 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] .

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]-[17].

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. These correspondences are usually obtained by segmenting the spectral vectors of one or both speakers into VQ-based clustering [17]. To create smooth transition between neighboring frames, vector quantization may be used to extract codebooks for 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]. VQ reduces the

computation amount and memory size drastically. 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

2. Methodology

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

2.1. Recording

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 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.

2.2 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. MFCCs are based on the known variation of the human ear's critical bandwidths with frequency [18]-[20]. MFCC is perhaps the best known and most popular, and are more robust to background noise [17], [19], [20], 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,

$$y_i = f_i[x_1, x_2, ...x_i, ..., x_p], \quad (1)$$

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

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] [21]-[25].

$$g({}^n w_1, {}^n w_2, \cdots, {}^n w_m) = f({}^n w_1, {}^n w_2, \cdots, {}^n w_m) + \varepsilon_n, \quad 0 \le n \le q-1 \quad (3)$$

The multivariate function can be written as

$$f(w_1, w_2, \cdots, w_m) = \sum_{k=0}^{p-1} c_k \phi_k(w_1, w_2, \cdots, w_m) \quad (4)$$

where p is the number of terms in the polynomial of m variables. By combining (3) and (4), we get a matrix equation

$$\mathbf{b} = \mathbf{Az} + \boldsymbol{\varepsilon} \quad (5)$$

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 × p matrix, with elements given as

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

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 ck's can be determined for minimizing the sum of squared errors

$$E = \sum_{n=0}^{q-1} \left[\begin{array}{l} g({}^n w_1, {}^n w_2, \cdots, {}^n w_m) \\ -f({}^n w_1, {}^n w_2, \cdots, {}^n w_m) \end{array} \right]^2 \quad (6)$$

and we get the solution

$$\mathbf{z} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \quad (7)$$

where $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is known as pseudo–inverse of A [17], [26].

2.3 Transformation of 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, obtained in the transformation function estimation block, for the given speaker pair.

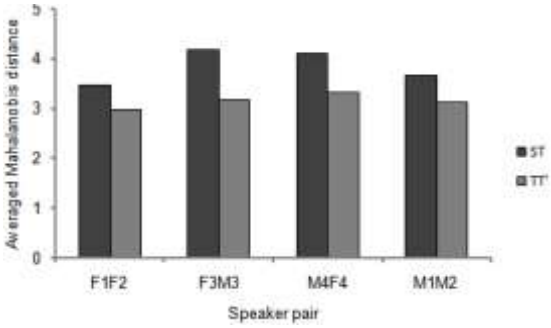
The error is estimated by finding the percentage of reduction in the spectral distance [17], [27]. The reduction in the spectral distance is carried out by calculating the cepstral Mahalanobis distance [27]-[33]. The distance between the target frames and the source frames is calculated as target-source 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.

3. Results and discussions

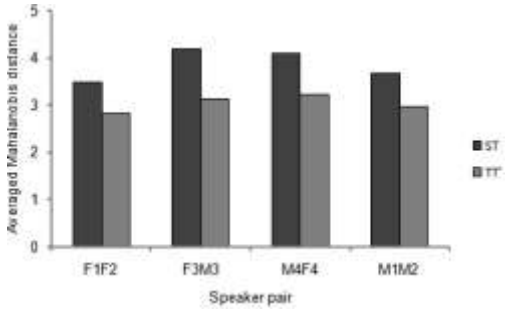
Using the technique described in the methodology, four transformation functions were estimated from the MFCCs derived from 20 utterances of aligned source and target speech material. Four different pairs (F1F2, F3M3, M4F4, and M1M2) were taken for estimating the transformation functions. The accuracy 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 of the distance between the source and target ST, target to transformed speech TT' is shown in Table1. In case of female to female transformation (F-F) for 512 classes, the original source to target distance (ST) is 3.47 and the corresponding distance between the target and the transformed speech (TT") is 2.82 giving a reduction of about 18.73%. In case of female to male transformation (F-M), the original source to target distance is 4.18 and the corresponding distance between the target and the transformed speech is 3.05 giving a reduction of about 27.03%. In case of male to female transformation (M-F), the original source to target distance is 4.09 and the corresponding distance between the target and the transformed speech is 3.17 giving a reduction of about 22.49%. In case of male to male transformation (M-M), the original source to target distance is 3.66 and the corresponding distance between the target and the transformed speech is 2.90 giving a reduction of about 20.76%. It may be observed that the reduction for cross gender conversion is maximum.

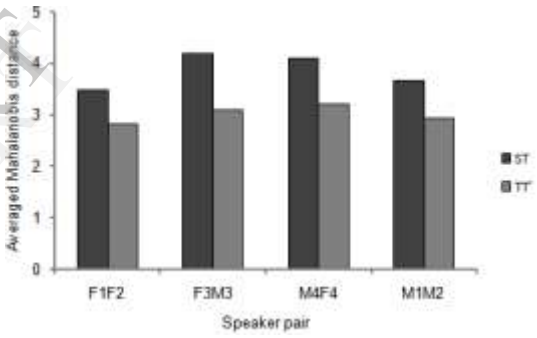
The mean of the distance between the source and target ST, target to transformed speech TT' is shown for classes 64, 128, 256, 512 as histograms in Fig. 1. In Fig. 1(a), the mean distance between ST and TT' is shown for 64 classes. It is seen from histogram that in case of female to male transformation (F3M3) the reduction is maximum i.e. 24.4% and for same gender the reduction is about 14%. In Fig. 1(b), the mean distance between ST and TT' is shown for 128 classes. It is seen from histogram that in case of female to male



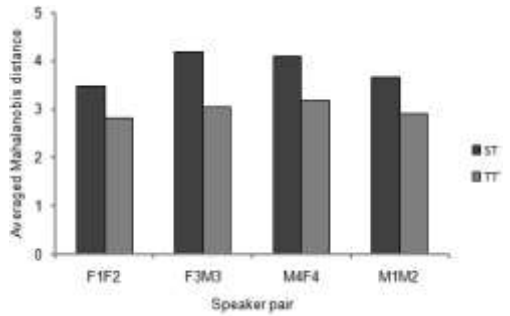
(a) Classes = 64



(b) Classes = 128



(c) Classes = 256



(d) Classes = 512

Fig. 1 Average distance between ST and TT' for four speaker pairs for different classes.

transformation (F3M3) the reduction is maximum i.e. 25.11% and for same gender the reduction is about 19%. In Fig. 1(c), the mean distance between ST and TT' is shown for 256 classes. It is seen from histogram that in case of female to male transformation (F3M3) the reduction is maximum i.e. 26.07% and for same gender the reduction is about 20%. In Fig. 1(b), the mean distance between ST and TT' is shown for 128 classes. It is seen from histogram that in case of female to male transformation (F3M3) the reduction is maximum i.e. 25.11% and for same gender the reduction is about 19%.

Table I. Averaged Mahalanobis distances between different speaker pairs.

Classes	Speaker pair	ST	TT'	Reduction (%)
64	F1F2	3.47	2.97	14.40
	F3M3	4.18	3.16	24.40
	M4F4	4.09	3.32	18.82
	M1M2	3.66	3.12	14.75
128	F1F2	3.47	2.83	18.44
	F3M3	4.18	3.13	25.11
	M4F4	4.09	3.22	21.27
	M1M2	3.66	2.96	19.12
256	F1F2	3.47	2.81	19.02
	F3M3	4.18	3.09	26.07
	M4F4	4.09	3.20	21.76
	M1M2	3.66	2.92	20.21
512	F1F2	3.47	2.82	18.73
	F3M3	4.18	3.05	27.03
	M4F4	4.09	3.17	22.49
	M1M2	3.66	2.90	20.76

4. Conclusions

Investigations were carried out to study the effect of 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 target speakers. The analysis of the results showed that the transformation function may be satisfactorily estimated even with 128 classes after aligning the source and target utterances with the help of VQ. Subjective evaluation of the transformed speech using ABX test is on our future plan.

5. 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] P. K .Lehana, P. C. Pandey, "Transformation of short-term spectral envelope of speech signal using multivariate polynomial modelling," in *Proc. National Conf. Commun.*, 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] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 565-568.

[15] W. Endres, W. Bambach, and G. Flosser, "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, 2013.

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

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

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

[21] 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.

[22] G. M. Philips. *Interpolation and Approximation by Polynomials*. Springer Verlag, New York, 2003.
[23] V. Pratt, "Direct least-squares fitting of algebraic surfaces*," Computer Graphics*, vol. 21, no. 4, pp. 145–152, 1987.
[24] P. C. Pandey and M. S. Shah, "Estimation of place of articulation during stop closures of vowel–consonant–vowel utterances," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 17, no. 2, pp. 277–286, 2009.
[25] 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.
[26] 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, vol.1, pp. 289-292.
[27] R. Curtin, N. Vasiloglou, and 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.
[28] 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.
[29] 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
[30] T. Takeshita, S. Nozawa, and F. Kimura, "On the bias of Mahalanobis distance due to limited sample size effect," in *Proc. IEEE int. Conf. Document Analysis, Recognition*, pp. 171-174, 1993.
[31] 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.
[32] T. Kamei, "Face retrieval by an adaptive Mahalanobis," in *Proc. IEEE Int. Conf. Image Process*., pp.153-156.