A Mutual Information Algorithm for Text-Independent Voice Conversion

Seyed Mehdi Iranmanesh[#1], Behnam Dehghan[*2]

[#]*Lane Department of Computer Science and Electrical Engineering, West Virginia University*
[*] *Industrial and Management Systems Engineering, West Virginia University*

Abstract*—Most of voice conversion systems require parallel corpora for their training stage which means that source and target speakers should utter the same sentences. But in most practical applications, it is impossible to obtain parallel corpora. To solve this problem, text-independent voice conversion has been introduced. The main problem in text-independent voice conversion is data alignment. In this paper we introduce a novel algorithm based on mutual information for data alignment which shows the similar results to those of text-dependent systems. This algorithm does not require phonetic labeling and can be used in practical applications.*

Key words: *Text-Independent Voice conversion, Mutual Information, Frame alignment, Mel cepstral frequency warping.*

I. INTRODUCTION

Voice conversion is a technique that modifies the voice of a speaker (source) which listeners think another speaker (target) utters that voice. All systems, including voice conversion systems (VCSs), have 2 stages: training and conversion. The goal of the training stage is gathering some matched data for source and target speakers and training a mathematical model to do a conversion. Since effectively training a mathematical model highly depends on training data, gathering matched data is a crucial stage in voice conversion (VC). However this is not limited to VC, many applications need matched data for their training e.g. [1]. If parallel corpora from source and target speakers, which means that both speakers utter the same sentences, are available, the matching procedure is easy to obtain. The only challenging part is timing difference between parallel corpora, one speaker may utter a sentence slower than another speaker. To solve this problem, researchers use dynamic time warping DTW [2], which has effectively used in so many applications e.g. [3], [4], and [5], for finding the exact correspond data for source and target speakers.

If parallel corpora are not available, gathering matched data will be more challenging. In this paper, we will address this problem.

II. RELATED WORK

Recently machine learning paradigms have been widely used in speech processing, including voice conversion, [6]. Specifically neural networks and dynamical systems, which have been used in many other applications [7], [8], and [9], have been used in voice conversion [10]. Vector quantization approach [11] which provides codebook for voice conversion is one of the first algorithms in this field. Another approach is use of Gaussian Mixture Model which provides smooth and stable results[12] and [13]. Please refer the papers [14] and [15] for comprehensive survey on VC.

In most text-independent systems, minimum distance between source and target features is thefirst and important step for feature alignment. Minimum distance works in this way: First speech features are extracted from source and target speeches and thenfor each source feature, the nearest neighbor (NN) feature in target space will be determined. In the rest of the paper we will show how the minimum distance works and will propose a new algorithm based on minimum distance.

III. MINIMUM DISTANCEAPPROACH

From first research on text-independent voice conversion to latest researches, minimum distance approachplays an important role in these systems. After framing and parameterization, source vectors $X = [x_1, x_2, \dots, x_m]$ which m is the number of source frames and target vectors $Y = [y_1, y_2, \dots, y_n]$ which n is the number of target frameswill be obtained. Without the loss of generality, consider for each of source and target speakers, training speech signals are concatenated. It means that m and n are total number of training frames for source and targets respectively. Also timing order has been preserved.

For each source vector from X, the goal is to find a target vector from Y that phonetically correspond to the source vector. Minimum distance approach is as:

$$p(k) = \arg\min_j d(x_k, y_j) \quad (1)$$

So the set of paired vectors$\{x_k, y_{p(k)}\}, k = 1, \dots, k$ are obtained.

By a simple test, we will show that this approachdoes not work as we expect. We have considered 2 parallel sentences for 2 male speakers. Parallel sentences have been aligned using DTW and Mel frequency cepstral coefficients (MFCCs) are extracted from all aligned frames (2500 frames in this case) and they have formed paired feature vectors. ($\{x_i, y_i\}, i = 1,2, \dots,2500$). Let'skeepthis paired vectors set as a baseline to compare minimum distance approach with.Now we will find corresponding vectors using

minimum distance approach and compared the results with those obtained from DTW. The results show in table1.

TABLE 1. THE NUMBER OF CORRECT SELECTED FRAMES USING MINIMUM DISTANCE APPROACH (1). ONLY 108 FRAMES ARE CORRECTLY OBTAINED.

Total frames	2500
correct selected frames	108

Table1 shows that only 108 frames out of 2500 frames have been foundcorrectly. However, there is an important note. Adjacent frames have approximately same phoneticallycontents. So if minimum distance approach finds the adjacent frames (in this case the frame after or the frame before), we do not loss any information. Table.2 shows the number of corrected frames considering the selection of adjacent frames as a true selection.

TABLE 2. THE NUMBER OF CORRECT SELECTED FRAMES USING (1) AND CONSIDERING THE SELECTION OF ADJACENT FRAMES AS A TRUE SELECTION.

Total frames	2500
correct selected frames	510

Although there is a major improvement, there is still a large error in finding corresponding frames. To improve the quality of this approach, some approaches have been introduced:

In [16], an approach has been introduced. In this approach, source and target vectors are clustered separately. Then corresponding target and source clusters are obtained by minimum distance between clusters.$\{\mu_i^x, \Sigma_i^x\}$and$\{\mu_j^y, \Sigma_j^y\}$ denote source and target clusters respectively which $\mu_i^x, \mu_j^y, \Sigma_i^x$and Σ_j^y represents mean and covariance of source and target clusters.

$$p(i) = \arg\min_j d_{DFW}(\mu_i^x, \mu_j^y) \quad (2)$$

So the set of paired cluster $\{\mu_i^x, \mu_{p(i)}^y\}$ is obtained. Then for each vector in ith source cluster, corresponding target vector in $p(i)$th target cluster was obtained by minimum distance approach.

$$x = \min_\chi |\chi - y - \mu_i^x - \mu_{p(i)}^y| \quad (3)$$

Also, another way to improve the minimum distance approach is using eq. (4) instead of eq. (1):

$$p(j) = arg\min_j([x_{t-1}, x_t, x_{t+1}], [y_{j-1}, y_j, y_{j+1}]) \quad (4)$$

Another main improvement of minimum distance approach was introduced in [17] which called INCA algorithm. This algorithm has 5 stages which they are mentioned shortly in following:

1. **Initialization.** In this stage, an auxiliary vector set, $X' = \{x_k'\}$, is described. Which it initialized as $x_k' = x_k$.
2. **NN alignment:** for each vectorx_k', the nearest neighbor vector in Y (target vectors) is found. Similarly, for each vector y_j, the nearest neighbor vector in X' is found.
$$p(k) = \arg\min_j d(x_k', y_j)$$
$$q(j) = \arg\min_k d(y_j, x_k') \quad (4)$$
Duplicated pairs are eliminated. Then $\{x_{q(j)}, y_j\}$ and $\{x_k, y_{p(k)}\}$ are obtained and concatenated.
3. **Training.**For concatenated vector, mth-order General Mixture Model (GMM) is fitted. Then $F_{aux}(x)$ is found.
4. **Transformation.** For each vector x_k', the updated vector is found by:
$$x_k' = F_{aux}(x_k) \quad (5)$$
5. **Convergence checking.** This algorithm is finished when convergence is reached according to a certain approach.

Although these approaches improved the quality of voice conversion using minimum distance area, there are still some problems. For example, since INCA [17] algorithm searches total target space for finding NN, definitely some wrong vectors is selected. These wrong vectors destroy the quality of voice conversion.

As mentioned before, by using minimum distance approach, some wrong target vectors are selected. These wrong vectors usually have deferent phonetic content with source vectors. Consider we a have source vector a^s which s denotes source. The corresponding a target vector for a^s is b^t which $t$denotes target. But, there is a vector in target space c^twhich has deferent phonetic content with a^s and b^t and also:

$$d(a^s, b^t) > d(a^s, c^t) \quad (6)$$

If we use the minimum distance approach, c^t is selected instead of b^t. This wrong selection destroys the quality of voice conversion. In this paper, we introduce a novel algorithm for alignment in order to decreasing the probability of finding wrong vectors. In the next section the new algorithm will be described.

IV. Mutual Information Clustering Algorithm (MIC)

Inspired by matching score presented in [18] we have designed an algorithm based on mutual information for feature alignment. Mutual information[19] of two random variables X and Y is defined as below:

$$I(X,Y) = \sum_y \sum_x p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) \quad (7)$$

And for continuous random variables mutual information is defined as:

$$I(X,Y) = \iint p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})\, dx\, dy \quad (8)$$

Mutual information between two variables X and Y measures the information that X and Y share. If X and Y are independent, knowing X does not give any information about Y, mutual information for these two variables is zero.Mutual information clustering based algorithm is an iterative algorithm which is containing clustering stage. Clustering stage is as follow:

1. **Initialization:** Mutual information was defined for random variables. Since adjacent frames in speech signals usually have the same phonetic content, they have to be in a same cluster. So after framing and parameterization of the speech signal, we considereach 50-70 consecutive frames as a cluster. To guarantee that all frames belong to correct cluster, we consider overlapping between clusters. We have a lot of small clusters now.
2. **Calculating mutual information:** For each clusters, we calculate mutual information with other clusters.
3. **Merging:** In this section we merge each cluster with a cluster that has maximum mutual information with.
4. **Stopping:** If the number of clusters is below than a constant number which is defined by a user, the algorithm is finished. Otherwise, the process is repeated from step 2.

Figure.1 shows this algorithm schematically.

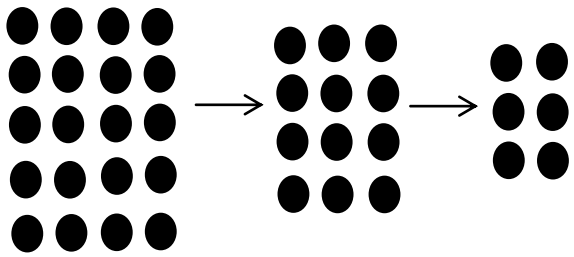

Fig.1 each circle correspond to a cluster. In each iteration, each cluster merge to a cluster that have maximum mutual information with.

For voice conversion, source and target speech signals are clustered separately with this clustering algorithm.$C_i^S \;\; i = 1,..,m$and$C_j^T \;\; i = 1,..,n$ are source and target clusters respectively. Then for each source cluster, we select a target cluster that has maximum mutual information with. Some target clusters may not be selected so for unselected target clusters, we select source clusters that havemaximum mutual information with (Fig. 2).

$$p(i) = \arg\max_j MI(C_i^S, C_j^T)$$

$$q(j) = \arg\max_i MI(C_j^T, C_i^S) \quad (9)$$

WhichMIstands for mutual information.

Now for each vector in ith source cluster, we find NN target vector in $p(i)$th target cluster. And for each vector in jth target cluster, we find NN source vector in $q(j)$th source cluster. Duplicated pairs are eliminated. So $Z = [X\ Y]^T$ was obtained.

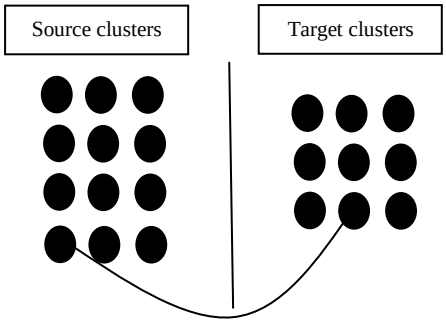

Fig.2.Curve shows the corresponding clusters.

Since adjacent frames are clustered in a same cluster, to increase the quality of alignment we can use eq. (4).

V. Conversion Function

For conversion function we use GMM [20]. For Z, Mth-order General Mixture Model (GMM) is fitted.

$$P(Z) = \sum_i^M \alpha_i \mathrm{N}(Z; \mu_z^i, \Sigma_{zz}^i), \qquad Z = \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$\mu_z^i = \begin{bmatrix} \mu_x^i \\ \mu_y^i \end{bmatrix}, \qquad \Sigma_{zz}^i = \begin{bmatrix} \Sigma_{xx}^i & \Sigma_{xy}^i \\ \Sigma_{yx}^i & \Sigma_{yy}^i \end{bmatrix} \quad (10)$$

Where α_i, μ_z^i and Σ_{zz}^i are the weights, the mean vectors, and covariance matrices of the m Gaussian componentsrespectively and $N()$ denotes the probability density function These parameters are obtained by using EM algorithm. Then conversion function can be defined as:

$$F(x) = \sum_{i=1}^m p_i(x)[\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx-1}(x - \mu_i^x)] \quad (11)$$

Where

$$p_i(x) = \frac{\alpha_i N(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{q=1}^M \alpha_q N(x; \mu_q^x, \Sigma_q^{xx})} \quad (12)$$

VI. EXPERIMENT AND DISCUSSION

In this paper, we focus on spectral parameter mapping. 24-dimensional mel frequency cepstral coefficient (MFCC) vectors are used as the spectral feature. The MFCCs are calculated from spectra obtained using STRAIGHT [21].

We used a logarithm Gaussian normalized transformation [20] to transform the $F_{0\,s}$ to the $F_{0\,t}$ as indicated below:

$$\log\left(F_{0\,conv} = \mu_t + \frac{\sigma_t}{\sigma_s}(\log(F_{0\,s}) - \mu_s\right) \quad (13)$$

Where μ_s and μ_t are the mean of the fundamental frequency in logarithm domain for source and target speakers respectively. $\sigma_s$andσ_t are the variance of the fundamental frequency in logarithm domain for source and target speakers respectively.

To compare these results with conventional parallel training [13], we used 60parallel English sentences for MIC algorithm and 45 parallel English sentences for parallel algorithm. The sentences have been spoken by two male and one femalespeakers. Also 20 English sentences selected for subjective test. The sentences are sampled at $16KHZ$.

MOS test is a subjective test for speech quality. Subjects rated the speech quality as below:

1 for bad, 2 for poor, 3 for fair, 4 for good, 5 for excellent

20 persons are asked to rate. Then MOS is the average of all persons' scores. Table 3 shows the results for male to male conversion (m2m) and male to female conversion(m2f). Both algorithms approximately have the same scores.

Table 1.Comparison MIC and Parallel by MOS test for m2m. Both algorithms approximately have the same score.

VC algorithm	MOS – m2m	MOS – m2f
MIC	3.5	2.7
Parallel	3.6	2.8

VII. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel data alignment algorithm for text-independent algorithm based on mutual information.

A subjective test shows that the new algorithm obviates the need of parallel corpora for voice conversion and also need a few sentences more than parallel algorithm for training stage. Since these extra sentences are not parallel, sentences can be easily obtained.Not requiring phonetic labeling is another advantage of this algorithm [22]. So it can be used in practical applications.In this paper we clustered source and target features separately, while it can be done jointly. This job would be our next approach.

References

[1] Z. Cao and N. Schmid. Matching heterogeneous periocular regions: Short and long standoff distances. In Image Processing (ICIP), 2014 IEEE International Conference on, pages 4967–4971, Oct 2014.

[2] Rabiner L, Juang B-H. Fundamental of Speech Recognition. NJ: Prentice Hall; 1993

[3] Motiian S, Pergami P, Guffey K, Mancinelli C.A, Doretto G. Automated extraction and validation of children's gait parameters with the Kinect. BioMedEngOnLine. 2015;14:112.

[4] S. Sempena, N. Maulidevi, P. Aryan, Human action recognition using dynamic time warping, ICEEI, IEEE (2011) 1–5.

[5] M. Toman, M. Pucher, S. Moosmuller, and D. Schabus, "Unsupervised interpolation of language varieties for speech synthesis," Speech Communication, 2015.

[6] J. Dean et aI., "Large scale distributed deep networks," NIPS, 2012. [II] L. Deng and X. Li. "Machine learning paradigms for speech recognition: An overview," IEEE Trans. Audio, Speech & Lang. Proc., Vol. 21, No. 5, May 2013.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," Proc. ICASSP, pp. 3893–3896, 2009.

[8] S. Motiian, K. Feng, H. Bharthavarapu, S. Sharlemin, and G. Doretto, "Pairwise kernels for human interaction recognition," in Advances in Visual Computing, 2013, vol. 8034, pp. 210–221.

[9] C. W. Han, T. G. Kang, D. H. Hong, N. S. Kim, K. Eom, and J. Lee, "Switching linear dynamic transducer for stereo data based speech feature mapping," in Proc. IEEE ICASSP, May 2011, pp. 4776–4779.

[10] N. S. Kim, T. G. Kang, S. J. Kang, C. W. Han, and D. H. Hong, "Speech feature mapping based on switching linear dynamic system," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 2, pp. 620–631, Feb. 2012.

[11] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on, 1988, pp. 655-658 vol.1.

[12] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 131-142, 1998.

[13] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, 1998, pp. 285-288 vol.1.

[14] Machado, A.F. and Queirozm M.: 'Voice Conversion: A Critical Survey', Proceedings of Sound and Music Computing (SMC) (2010).

[15] Y. Stylianou, "Voice transformation: a survey," in ICASSP 2009.

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

[17] D. Erro, A. Moreno, and A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 944-953, 2010.

[18] Z. Cao and N. A. Schmid, "Fusion of operators for heterogeneous periocular recognition at varying ranges," IEEE International Conference on Image Processing, 2014, pp. 4967-4971.

[19] Fazlollah M. Reza, "An Introduction to Information Theory. Dover Publications," Inc, New York. ISBN 0-486-68210-2.

[20] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 2222-2235, 2007.

[21] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in *Acoustics, Speech, and Signal Processing*, 2000

[22] J. Tao, M. Zhang, J. Nurminen, J. Tian, X. Wang, "Supervisory Data Alignment for Text-Independent Voice Conversion," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 5, JULY 2010