ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

IMPROVING THE PERFORMANCE OF MGM-BASED VOICE CONVERSION BY PREPARING TRAINING DATA METHOD

Guo-yu ZUO[1, 2], Wen-ju LIU[1], Xiao-gang RUAN[2]

[1]National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences, Beijing
{gyzuo,lwj}@nlpr.ia.ac.cn
[2]School of Electronics Information and Control Engineering,
Beijing University of Technology, Beijing
adrxg@bjut.edu.cn

ABSTRACT

This paper proposes an approach to improve both the target speaker's individuality and the quality of the converted speech by preparing the training data. In mixture Gaussian spectral mapping (MGM) based voice conversion, spectral features representations are analyzed to obtain the right feature associations between the source and target characteristics. A voiced and unvoiced (V/UV) decision scheme for time-alignment is provided to obtain the right data for training mixture Gaussian spectral mapping function while removing the misaligned data. Experiments are conducted in terms of the applications of spectral representation methods and V/UV decisions strategies to the MGM functions. When linear predictive cepstral coefficients (LPCC) are used for time-alignment and the V/UV decisions are adopted for removing bad data, results show that the conversion function can get a better accuracy and the proposed method can effectively improve the overall performance of voice conversion.

1. INTRODUCTION

Voice conversion (speaker conversion) is a technique to makes one speaker's voice sound as if it were uttered by another speaker [1]. Various applications are available that range from automated telephone querying system to medical healthcare. Recently, much attention has been paid to build a multi-speaker database from a one-speaker corpus, which has potential to get an unlimited amount of speech data without costing any more than the conventional data-acquisition approaches [2].

This technique has been developed over the past twenty years and many conversion approaches have been proposed. Abe et al proposed a VQ-based codebook mapping method [3]. Stylianou described a probabilistic mixture Gaussian mapping (MGM) method [4], which has the useful property of being continuous. Neural networks [5] as well as the adaptive filtering method [6] have also been used for mapping spectral characteristics of the speakers. In general, the MGM method has shown a superior performance to the other transformation approaches.

Spectral features play an indispensable part in determining the speaker's identity. In all these conversion algorithms noted as above, the spectral characteristics are represented by LPCs, MFCCs, and LSFs and the others [3-6]. To some extent, however, these representations are different from each other in the listener's perceptive distances. In conversion processing, various time-alignment approaches such as dynamic time warping (DTW) and force-alignment [7] are performed to get the feature associations between the source and target characteristics. In this work, different presentations are analyzed and studied to find their effects on spectral distance measures and on DTW alignment results. A voiced and unvoiced decision strategies specified for DTW use to get right data are presented for training mixture Gaussian mapping function. The experi-ments and subjective evaluations on speech quality and speaker identity are carried out to observe the improvements in voice conversion performance by preparing training data as above, which will be addressed in detail in the next sections.

2. VOICE CONVERSION BASED ON MIXTURE GAUSSIAN MAPPING METHOD

In mixture Gaussian mapping method, the joint density approach is applied to the density $p(\boldsymbol{x},\boldsymbol{y})$ and predict the target $\boldsymbol{y}$ from the source $\boldsymbol{x}$ by finding $E[\boldsymbol{y}|\boldsymbol{x}]$, the expected value of $\boldsymbol{y}$ given $\boldsymbol{x}$.

In this method, a Gaussian mixture model is fit to the probability distribution of acoustic features, which is given by

$$p(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \ \sum_{i=1}^{m} \alpha_i = 1, \ \alpha_i \geq 0 \qquad (1)$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, m denotes the total number of Gaussian mixtures. α_i denotes the weight of class i.

The features of the source speaker are converted into those of the target speaker using the mapping function as follows:

$$\hat{\boldsymbol{y}} = E[\boldsymbol{y} \mid \boldsymbol{x}] = \sum_{i=1}^{m} h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i^x)]$$
$$h_i(\boldsymbol{x}) = \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})} \quad (2)$$

where

$$\boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}, \boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix} \quad (3)$$

are the mean vector and covariance matrix of class *i*. These parameters are trained from the joint vectors $\boldsymbol{z} = [\boldsymbol{x}^T, \boldsymbol{y}^T]^T$, which are composed of the time-aligned source vectors $\boldsymbol{x}$ and target vectors $\boldsymbol{y}$ and probabilistically described by a GMM whose parameters are trained by joint density distribution [8]. $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ denote the mean vectors of class *i* for the source and target speakers. $\boldsymbol{\Sigma}_i^{xx}$ denotes the covariance matrix of class *i* for the source speaker, $\boldsymbol{\Sigma}_i^{yx}$ is the cross covariance matrix of class *i* for the source and target speakers. The EM algorithm can be used to find the most likely GMM parameters $(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a given set of data and the least conversion error $\min E\{\| \boldsymbol{y} - \hat{\boldsymbol{y}} \|^2\}$ is obtained on all the training data.

3. PREPARING DATA FOR TRAINING MGM FUNCTION

3.1 Spectral Representations Used for DTW and GMM Training

In order to get an accurate spectral transformation function, different representations of spectral features are extracted for time-alignment, training MGM function and analysis-synthesis respectively.

Before training MGM function, the training data of joint vectors, which are composed of the source and target vectors, is obtained using the DTW approach. Linear predictive cepstral coefficients (LPCCs) are adopted to perform time-alignment, since cepstral coefficients have been found to give a better associated relations than LPCs, MFCCs and the others between the associated source and target speakers' utterances. Fig.1 plots three DTW paths using three feature representations LPCs, LPCCs, and MFCCs respectively. The speech comes from the 863DB synthesis database with the source speaker named Guo L and the target named Peng GW. The associated script is "达斡尔族" (Da2 wo4 er3 zu2). It is noted in the plot that the alignment path in respect of LPCs is far away from those of LPCCs and MFCCs. it is therefore reasonable to think that selecting feature representation is an important factor in the problem of getting the right joint vectors before training MGM function.

Each pair of the time-aligned source and target vectors is

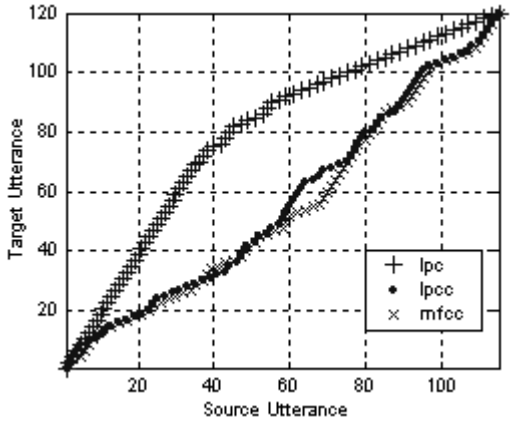

Fig.1 DTW paths using different feature representations

then composed into a joint vector. Bark-warped LSFs are used as the representation of spectral features in voice conversion for several reasons as follows. LSFs are closely relative to formant frequencies, but in contrast to formant frequencies they can be estimated quite reliably and easily derived from the associated LPC coefficients; the spectral mapping function conducts linear combination on spectra, and LSFs have been shown to possess good linear interpolating performance; since human ears are sensitive to the low frequencies of speech, each LSF pair for one speech frame are bark-scaled in order to better reflect the perceptual distance between LSF pairs. The bark-warping function is described by

$$bark(f) = 6.0 \log\left\{\frac{f}{1200} + \sqrt{\left(\frac{f}{1200}\right)^2 + 1}\right\} \quad (4)$$

In the conversion system, the barked LSFs are extracted to train the a Gaussian mixture model as Eq.(1) represents, whose parameters are used to get the spectral mapping function Eq.(2).

3.2 Decision Strategies of Voiced/Unvoiced Speech

In DTW processing, there is a possibility in the joint vectors that the source speech frame is voiced and the target is unvoiced and vice versa. This means that the speech of different classes of acoustic or voicing properties is liable to be classified into the same class. The misaligned potential usually occurs to the boundaries of Initials and Finals in Mandarin speech, which will lead to speech classifying errors and, in the end, cause a negative influence on the target speaker identities and qualities of converted speech. The DTW processing is carried out to remove the misaligned feature vector pairs due to their voicing properties. Many V/UV decision methods are available, based on which, however, the V/UV decision approach for DTW use is proposed. It takes into account some voice phenomena such as such as vocal fry, aperiodic phonation and others, which can remove the misaligned or bad data as

possible as it can, although its absolute accuracy is not so high as some state-of-the-art methods. In this proposed approach, the frames below the silence energy threshold are decided to be silences and some variables are determined:

ZCR_TH and EN_TH denote the zero-cross rate threshold and the energy one respectively.

VUVZC and VUVEN denote voicing flags respectively associated with zero-cross rate and energy.

The strategies are summarized as follows:

1) Calculate ZCR_TH and EN_TH and get the original decision results VUVZC and VUVEN.

2) For each frame whose VUVZC is 0 (unvoiced), count the number of samples whose amplitude is larger than 0.4 times the utterance maximum. If it is larger than 3, set its VUVZC to be 1.

3) For each frame whose VUVEN is 0, count the number of its samples whose amplitude is larger than one third of the utterance maximum. If it is larger than 3, set its VUVEN to be 1.

4) For each frame whose VUVZC is 1, if its energy is large than 0.75 times EN_TH, set its VUVEN to be 1.

5) For each frame whose energy is larger than 2 times EN_TH, if the zero-cross rate is less than 1.2 times ZCR_TH, set its VUVZC to be 1.

6) For each frame whose energy is less than 2.5 times EN_TH, if zero-cross rate is less than 1.4 times ZCR_TH, set VUVZC to be 0.

7) If each frame's VUVEN and VUVZC are both voiced, then let this frame be voiced and otherwise unvoiced.

8) If the number of the continuous voiced frames is less than 5, set them to be unvoiced. And then, if the number of the continuous unvoiced frames is less than 5, set them to be voiced.

4. IMPLEMENTATION OF VOICE CONVERSION SYSTEM

Plotted in Fig.2 is the framework of voice conversion system, which includes the training and conversion parts. In training stage, the source and target speech are framed into half the frame-length of overlapped blocks with a Hanning window. LPCCs and bark-warped LSFs are calculated for DTW and MGM respectively. After the V/UV decision and DTW processing, the joint vectors are used to train the MGM function, which will associates the source and target characteristics spaces.

In conversion stage, each frame of the input speech is analyzed into a frame of barked LSFs and the associated LPC residual signals, and the latter can be obtained by inverse filtering each frame of speech using the associated LPCs. The input vectors are converted into the expected target vectors using the spectral conversion function (to reduce the discontinuity of speech caused by MGM, a low-pass filter is applied to each of LSF coefficients to smooth the value differences between neighboring frames). The converted spectra are combined with the analyzed residual signals to get the converted speech frames.

After energy modification, the overlapped frames are added into converted speech and finally used TD-PSOLA technique match the average fundamental frequency (F0) of input speech to that of target speech. The F0 mapping function is given by

$$f_0^{'} = \frac{\mu_t}{\mu_s} \times f_0 \tag{5}$$

in which f_0 and f_0' denote log scale of F0 of the input and converted speech, and the μ_s and μ_t denote mean log scale F0 of the source and target speech.

5. EXPERIMENTS AND EVALUATIONS

5.1 Experimental Data and MGM Training Schemes

The experiments on speech quality and speaker identity are conducted to evaluate the performance of the proposed method. Speech of 3 (two male and one female) speakers from the 863DB synthesis database is used for evaluation. The training data sampled at 16 kHz consist of 200 4-word sentences. Another 5 sentences are used for test, each phone of which has occurred in the training data more than three

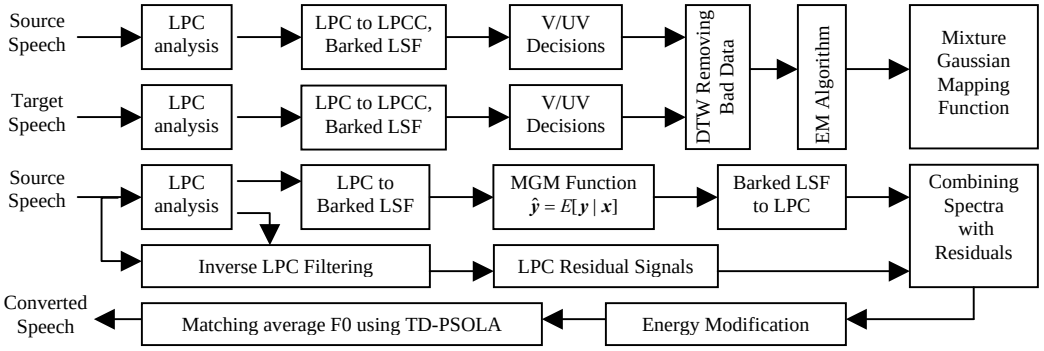

Fig.2 Training and conversion structures of voice conversion system

times. The feature vector's order of all the types of representations is 20 with the number of Gaussian mixtures set to be 62. The male-to-male (M2M) and male-to-female (M2F) conversions are carried out for the speech quality experiments, and only the M2M conversion is implemented for the individuality experiments. Experiment results for three kinds of MGM function training schemes are evaluated under these conditions; what kind of feature representation is adopted and whether the V/UV decision strategies are applied:

1) LPC: LPC for DTW and no V/UV decisions.

2) LPCC: LPCC for DTW and no V/UV decisions.

3) LPCCVUV: both LPCC for DTW and V/UV decisions.

5.2 Evaluations on Speech Quality

The evaluations on speech quality are conducted both on the male-to-male and male-to-female conversions. By adopting the strategies of both the LPCC parameters for DTW use and the V/UV decisions for removing bad data, Fig.3 shows that their qualities (with probabilities higher than 80%) outperform those without V/VU decisions. The listeners believed that the latter speech is perceived rather dull in contrast with the former. For the male-to-male conversion, the generated speech sounds a little better than that for the male-to-female one. It is found in these tests that the qualities with LPC are a little worse than those with LPCC.

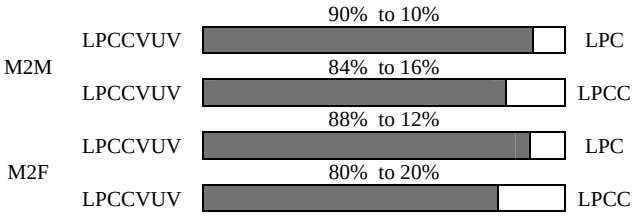

Fig.3 Evaluation results for speech quality

5.3 Evaluations on Speaker Identity

The speaker individuality of the converted speech is evaluated by ABX test. In the first test, decide which of the source speech A and the target B is closer to the converted speech X which is generated by each scheme. In the second test, decide which of the converted speech with LPCC and LPCCVUV respectively is closer to the target speech. The evaluation results are reported in Fig.4.

It is evident that the converted speech in general sounds more like the target speaker than the source one. The comparative results by LPC with LPCC underlie the positive influence of the proper spectral representation on the converted speech's speaker identification. As reported from the middle two rows in Fig.4, the MGM function, accompanied with an obvious improvement in speech

Fig.4 Evaluation results for speaker identity

quality, gives a probabilistic increase of 8% in the target speaker's individuality of the converted speech when the V/UV decision strategies are applied. It is also found from the last row that the converted speech by LPCCVUV sounds more like the target speech than that by LPCC only.

6. CONCLUSION

An approach is presented to improve the quality and the target speaker's individuality of the converted speech. The DTW processing for getting the right training data of joint vectors takes into account the influence of spectral feature's representation. The proposed V/UV decision strategies for DTW use are to prevent the rapid degradation of the conversion model's accuracy from the misaligned data. Results show that the proper representation type of spectral features plays a part in getting the right data, which can help correctly associate the source feature vectors with the target ones. With the V/UV decisions adopted, further evaluations show that there is an evident improvement in the speech quality and target speaker's individuality in the MGM-based voice conversion.

REFERENCES

[1] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*. 1995, 16 (2): 165-173.

[2] G. Zuo, W. Liu and X. Ruan. Voice Conversion Technology and Its Development. *Acta Electronica Sinica*, 2004, 32(7): 1165-1172.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice Conversion through Vector Quantization. *Proc. ICASSP*, pp. 655-658, 1988.

[4] Y. Stylianou, O. Cappe and E. Moulines. Continuous Probabilistic Transform for Voice Conversion. *IEEE Tran. Speech and Audio Proc.*, vol.6, No.2, pp.131-142, March 1998.

[5] T. Watanabe, et al. Transformation of Spectral Envelope for Voice Conversion Based on Radial Basis Function Networks. *Proc. ICSLP*. Denver, USA. Sept. 2002: 285-288.

[6] O. Salor, M. Demirekler and B. Pellom. A System for Voice Conversion based on Adaptive Filtering and Line Spectral Frequency Distance Optimization for Text-to-Speech Synthesis. *Proc. Eurospeech'03*, Switzerland, Sept. 2003.

[7] L M Arslan. Speaker Transformation Algorithm using Segmental Codebooks (STASC). *Speech Communication.* 1999, 28(3): 211-226.

[8] A Kain and M Macon. Spectral Voice Conversion for Text-to-Speech Synthesis. *Proc. ICASSP*. Seattle, USA, May 1998(1): 285-288.