TEXT-INDEPENDENT VOICE CONVERSION BASED ON STATE MAPPED CODEBOOK

Meng Zhang[1], Jianhua Tao[2], Jilei Tian[3], Xia Wang[4]

[1,2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
Beijing, China
[3]Interaction Core Technology Center, Nokia Research Center, Tampere, Finland
[4]Nokia Research Centre, China
{[1]mzhang, [2]jhtao}@nlpr.ia.ac.cn [3]jilei.tian@nokia.com [4]xia.s.wang@nokia.com

ABSTRACT

Voice conversion has become more and more important in speech technology, but most of current works have to use parallel utterances of both source and target speaker as the training corpus, which limits the application of the technology. In the paper, we propose a new method of text-independent voice conversion which uses non-parallel corpus for the training. The Hidden Markov Model (HMM) is used to represent the phonetic structure of training speech and to generate the training pairs of source and target speakers by mapping the HMM states between source and target speeches. Then, HMM state mapped codebooks are generated to create the mapping function for the text-independent voice conversion. The subjective experiments based on ABX tests and MOS tests show that the method proposed in the paper gets the similar conversion performance and better speech quality compared to the conventional voice conversion systems.

Index Terms— text-independent, voice conversion, hidden Markov model, state mapped codebook

1. INTRODUCTION

Voice conversion is a technique which is used to transform one speaker's voice in order to make it perceived as if uttered by another speaker. By far, most of current voice conversion methods are based on text-dependent corpus, which means the training set has to be extracted from parallel utterances of both source and target speakers. The source and target speech can be automatically aligned by dynamic time warping (DTW) [1]. However, the preparation of parallel speech database is very inconvenient in real-life applications. In addition, it is unlikely to have parallel utterances of both source and target speakers in some applications, e.g. cross-lingual voice conversion. Obviously it is important to develop the text-independent voice conversion based on non-parallel database. In recent years some approaches of text-independent voice conversion were proposed [2][3][4].

Comparing to text-dependent voice conversion, the biggest challenge for text-independent voice conversion is how to align the corresponding training data for obtaining mapping function or codebook. In the text-dependent case, aligned data can be derived from parallel speech by DTW. Parallel speech is necessary because otherwise there is no guarantee that in the conversion part the phonetic contents remain unchanged [4]. However, in text-independent case, the database doesn't have a high level of inherent time alignment. What is more troublesome is sometimes the phoneme sets of source and target speech are different from each other. For example, in incomplete speech database or cross-lingual cases, some source phonemes can't align to suitable target.

Hidden Markov model (HMM) has been successfully applied to speech recognition systems for its excellent ability of characterizing the spectral parameter sequence and modeling phonetic structure. When modeling speech, each HMM corresponds to a phonetic unit with phonetic signification. Considering these facts, the HMM is used to represent the phonetic structure of training speech. The transformation between source and target characteristics is accomplished by establishing a mapping between the source and target HMM states from generated the training pairs. Model similarity is used to measure the correlation between source and target states for state alignment. Then, HMM state mapped codebooks are generated to create the mapping function for the text-independent voice conversion.

The paper is organized as follows. In Section 2, our new technique of text-independent voice conversion based on state mapped codebook is derived. Subjective experimental results are presented in Section 3. For the future work, we give a preliminary discussion about the effect of different combinations of phoneme sets on text-independent data alignment in Section 4. Finally, conclusion is drawn in Section 5

2. TEXT-INDEPENDENT VOICE CONVERSION BASED ON STATE MAPPED CODEBOOK

2.1 Data alignment

HMMs are separately trained on the source and target speech data in order to represent speakers' characteristics. The transformation of these characteristics is accomplished by establishing a mapping between the source and target models. We propose a codebook-based mapping method to improve the performance since the linear transformation tends to cause over-smoothing problem [6][7]. The states of each HMM are used as structure unit in mapping approach, so we call it as state-book mapping. Fig. 1 shows the block diagram of the new approach.

After training the HMMs for source and target speech respectively, a mapping between source and target states structure which are described by a single Gaussian distribution is built. We can find the corresponding target state given a source state by comparing the similarity between the Gaussian distribution of state models. For each source state $x_k(m_k^x, v_k^x)$ $(k=1,\ldots,K)$, we want to find the corresponding target state $y_l(m_l^y, v_l^y)$ $(l=1,\ldots,L)$ by

$$l(k) = \arg\min_{l=1,\ldots,L} D(x_k, y_l) \tag{1}$$

$$D(x_k, y_l) = \int P(X \mid x_k) \cdot P(X \mid y_l) dX$$
$$= \int N(X \mid m_k^x, v_k^x) \cdot N(X \mid m_l^y, v_l^y) dX$$

where m_k^x, v_k^x are mean and variance of the Gaussian distribution of state x_k respectively. The state-book can be obtained between source and target composed by mapped state index pairs. It is worth noted that there is only one corresponding target state given a source state. But a target state can be pointed by several source states.

However, it seems not very reasonable because the distinction between features of different speaker's speech includes the difference of physiological individuality and phonetic difference. Mapped pair should be established based on contents of speech as much as possible for a phonetic parallel mapping. In order to remove the effectiveness of speaker characteristics during state mapping procedure, we apply dynamic frequency warping to source state parameter before alignment, $m_k^{x'} = w(m_k^x)$. So equation (1) is modified to be

$$l(k) = \arg\min_{l=1,\ldots,L} D(x_k', y_l) \tag{2}$$

where x_k' is warped source state generated from $x_k(m_k, v_k)$ and defined as $x_k'(m_k^{x'}, v_k^x)$.

2.2 Transforming

For a given input feature sequence from source speech, we label each frame of data X with a state index k based on source HMM's state parameter probability distributions.

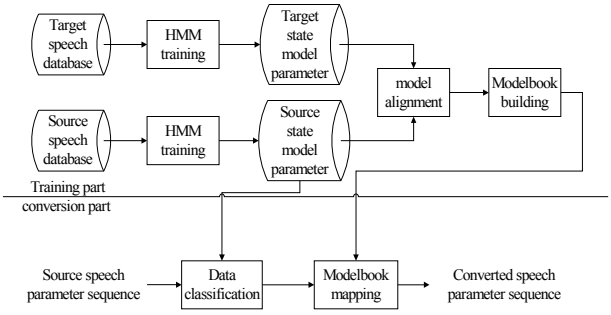

Fig.1 Flow chart of voice conversion based on state mapped codebook

$$k(X) = \arg\max_{k=1,\ldots,K} P(x_k \mid X)$$
$$= \arg\max_{k=1,\ldots,K} (\alpha_k^x N(X \mid m_k^x, v_k^x)) \tag{3}$$

By doing this we can get a sequence of source state index which implies feature sequence of source speech through state sequence. In order to achieve transformation between source and target speech, we replace each index in the source state index sequence by the corresponding target state index according to the state-book we have built in Section 3.2.

$$L = T(K) = T(\{k_1, \ldots k_N\}) = \{l(k_1), \ldots l(k_N)\} \tag{4}$$

As the same with source state index sequence, achieved target state index sequence is a description of converted speech feature sequence. But this description is discontinuous because of the influence of vector quantization during the labeling process. To handle this, we applied a parameter generation algorithm [8] used in HMM based speech synthesis (HTS) to generate parameter sequence of converted speech. By using this parameter generation algorithm, a continuous and smooth converted parameter sequence is achieved.

3. EXPERIMENT AND DISCUSSION

3.1 Linear transformation for excitation information

In this paper, we generally focus on the spectral parameter mapping. For excitation information transforming, we simply apply the linear transformation to convert the pitch contour of input speech into the target speaker's pitch range by

$$f_y = \mu_y + \frac{\sigma_y}{\sigma_x}(f_x - \mu_x) \tag{5}$$

where (μ_x, σ_x) and (μ_y, σ_y) are mean and variance of source and target speech respectively.

The STRAIGHT-based vocoder [9] can model the vocal tract and vocal source information successfully. Thus it is used to extract acoustic features from speech and synthesize speech from converted parameters.

3.2 Subjective evaluation

In this section, an experiment is designed to compare original codebook mapping based text-dependent voice conversion with state mapped codebook based text-independent voice conversion regarding both conversion performance and speech quality. We also test our method on parallel database to analyze its sensibility to the category of database it used

For the original text-dependent voice conversion, we selected 180 parallel UK English sentences from source and target speaker respectively. The same setting was also applied to the proposed text-independent for performance comparison by using unparallel database. For text-independent case, we also prepared 180 sentences from each speaker respectively for HMM training. The training sentences are difference between source and target speeches.

Each HMM was defined as 3-state left-to-right with no skip. The acoustic features are composed of 24-order LSF (Linear spectral frequency) coefficients obtained by STRAIGHT filter with a 5ms shift. Finally the spectrum parameter vector consists of 25-order LSF coefficients with gain, delta and delta-delta coefficients for the dynamic comparison and HTS based parameter generation. The triphone HMMs are trained and resulted in 3099 source states for source speaker and 2985 states for target speaker to establish the state-book.

The subjective test was performed with 11 subjects participating. Each subject was asked to value the converted speech regarding conversion performance by ABX test and speech similarity test and speech quality by MOS test respectively.

In speech similarity test (denoted as SS), subjects listened to speech sample pairs of converted speech and target speech and were asked to score their similarity on a five-point scale (1 for different to 5 for similar).

In ABX test, subjects were asked to score each converted speech by 0 for similar to source and 1 for similar to target.

MOS test is a test for speech quality. Subjects rated the speech quality on five-point scale (1 for bad, 2 for poor, 3 for fair, 4 for good, 5 for excellent). And MOS is the average of all the subjects.

3.3 Results and discussion

Fig. 2-4 shows the results of subjective evaluations for four gender combinations with the three techniques we mentioned in section 4.2. T-d denotes text-dependent voice conversion, T-i-p stands for text-independent voice conversion using parallel database and T-i-nonp means the one using unparallel database.

From Fig. 2 and Fig. 3 we can see that conversion performance of original text-dependent voice conversion is marginally better than text-independent voice conversion but not much, nearly 0.4 SS point and 0.1 ABX point. The difference maybe caused by the data alignment procedure of the text-independent training. In this new text-independent method, data is aligned by state mapping based on similarity of Gaussian distribution. That may not be reasonable enough for some case that state parameters for the same phoneme from source and target respectively is much different from each other. Because the more different the states are, the less possible they align together under the criterion we used, which doesn't take phonetic contents information into account.

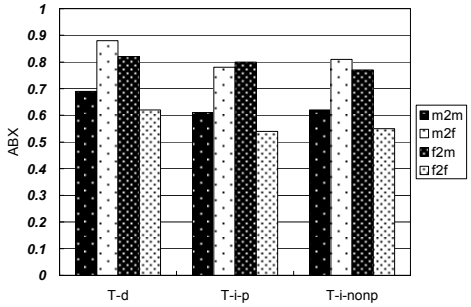

Fig.2 Results of the subjective ABX test (0 for failure, 1 for success)

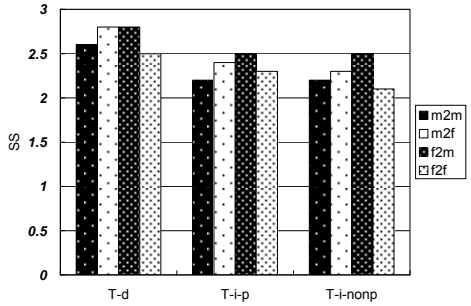

Fig.3 Results of the subjective test for speech similarity (1 for failure, 5 for success)

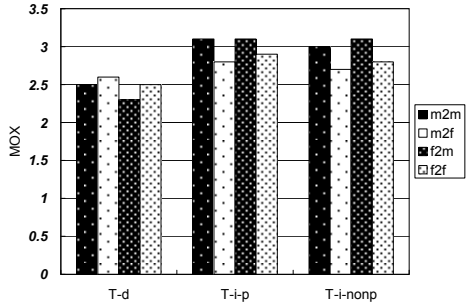

Fig.4 Results of the subjective MOS test

Comparing the converting performance of T-i-p with T-i-nonp, we can easily find that they are nearly the same. This means our new method is not sensitive to the category of the database it used.

Fig. 4 shows us that the new approach gives a better speech quality than the original text-dependent codebook based voice conversion. In conversional codebook based voice conversion, the converted spectral parameter is generated by codeword. That causes the decline of the speech quality. Although the new method we proposed is also based on codebook, we apply it in state level and use parameter generation algorithm presented in HTS to generate a smooth spectral parameter sequence from the converted state sequence. Thus the new approach can give a high quality converted speech whose quality should be similar to the results derived from HTS.

4. DISCUSSION ABOUT DIFFERENT COMBINATIONS OF PHONEME SETS

Our future work will focus on how to use the phonetic contents of speech more efficiently for supervisory data alignment in text-independent situation. Here we give a preliminary discussion about the different combinations of source and target phoneme sets. In some case, the phoneme sets of source and target speech are different from each other. Different combinations of source and target phoneme sets may have different influence on the data alignment procedure. According to different components of source and target phoneme sets there are generally five kinds of source and target phoneme sets combinations:

1. $S_x\{C\}:S_y\{C\}$ 2. $S_x\{X,C\}:S_y\{C\}$
3. $S_x\{C\}:S_y\{C,Y\}$ 4. $S_x\{X,C\}:S_y\{C,Y\}$
5. $S_x\{X\}:S_y\{Y\}$

Where S_x, S_y are source and target speech phoneme set respectively. C represents set of common phonemes in both source and target phoneme sets. X, Y represent set of the mismatched phonemes in source and target phoneme sets respectively.

Because the conversion is a transformation of data from source speech to target speech, target mismatched phonemes don't have much effectiveness to the alignment procedure. According to this, all combinations can be re-sorted into three kinds:

1. $S_x\{C\}:S_y\{C\}$ 2. $S_x\{X,C\}:S_y\{C\}$
3. $S_x\{X\}:S_y\{Y\}$

For combination 1, the simplest case, the phoneme sets of source and target is the same. We can directly align source and target speech data for each pair of mapped phonemes according to the phonetic contents. For combination 2 there are some mismatched phonemes in source which don't have mapping targets. We could firstly train mapping function or build mapped codebook using shared phonemes' speech data and then adapt the mapping criterion to those mismatched phonemes. Combination 3 is the most difficult case where phonetic contents of speech are useless because there aren't any shared phonemes between source and target speech. For this case we could align data according to the acoustic feature similarity.

5. CONCLUSION

In this paper, we proposed a new text-independent voice conversion based on state mapped codebook. This new approach uses HMM to present the phonetic structure of training speech and achieve the conversion by mapping states between source and target HMMs. Finally we apply the parameter generation algorithm presented in HTS to generate a smoothed spectral parameter sequence from the converted state sequence. Subjective experiment shows that the new approach has nearly the same conversion performance as original text-dependent voice conversion and better speech quality.

6. ACKNOWLEDGEMENT

The work was supported by the National Natural Science Foundation of China (No. 60575032, No. 70611120555) and 863 Programs (No. 2006AA01Z138, No. 2006AA01Z194) and Nokia

7. REFERENCES

[1] Y. Stylianou, O. Capp´e, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, 1998.

[2] D. S¨undermann, H. Ney, and H. H¨oge, "VTLN-Based Cross-Language Voice Conversion," in *Proc. of the ASRU'03*, Virgin Islands, USA, 2003.

[3] H. Ye and S. J. Young, "Voice Conversion for Unknown Speakers," in *Proc. of the ICSLP'04*, Jeju Island, South Korea, 2004.

[4] D. S¨undermann, H. H¨oge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. 2006a. "Text-Independent Voice Conversion Based on Unit Selection". In *Proc. of the ICASSP'06*, Toulouse, France.

[5] D. S¨undermann, A. Bonafonte, H. Ney, and H. H¨oge, "A First Step Towards Text-Independent Voice Conversion," in *Proc. of the ICSLP'04*, Jeju Island, South Korea, 2004.

[6] T. Toda, H.Saruwatari, and K. Shikao, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum". In *Proc. Of ICASSP*, 2001, pp.841.944.

[7] T. Toda, Alan W Black, Keiici Tokuda. "Spectral conversion based on maximum likelihood estimation considering globel variance of converted parameter", In *Proc. Of ICASSP,* 2005

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. of ICASSP*, pp. 1315-1318, 2000.

[9] H.Kawahara, I. Masuda-Katsuse and A. deCheveigne, "Restructuring speech representations using pitch-adaptive timefrequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187-207, 1999