PHONEME CLUSTER BASED STATE MAPPING FOR TEXT-INDEPENDENT VOICE CONVERSION

Meng Zhang[1], Jiaohua Tao[1], Jani Nurminen[2], Jilei Tian[3], Xia Wang[4]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Nokia, Devices R&D, Tampere, Finland
[3]Nokia Research Center, Tampere, Finland
[4]Nokia Research Center, Beijing, China
{mzhang, jhtao}@nlpr.ia.ac.cn, {jani.k.nurminen, jilei.tian, xia.s.wang}@nokia.com

ABSTRACT

This paper takes phonetic information into account for data alignment in text-independent voice conversion. Hidden Markov Models are used for representing the phonetic structure of training speech. States belonging to same phoneme are grouped together to form a phoneme cluster. A state mapped codebook based transformation is established using information on the corresponding phoneme clusters from source and targets speech and weighted linear transform. For each source vector, several nearest clusters are considered simultaneously while mapping in order to generate a continuous and stable transform. Experimental results indicate that the proposed use of phonetic information increases the similarity between converted speech and target speech. The proposed technique is applicable to both intra-lingual and cross-lingual voice conversion.

Index Terms— text-independent voice conversion, Hidden Markov Model, state mapping

1. INTRODUCTION

Traditional text-dependent voice conversion techniques require parallel training data which must be further aligned using e.g. dynamic time warping (DTW) [1]. From the viewpoint of practical applications, the requirement of having parallel speech databases is rather inconvenient and sometimes hard to fulfill. Moreover, in some cases it may be even impossible to obtain parallel speech corpora. This is the case e.g. in cross-lingual voice conversion where the source and the target speakers speak in different languages. Due to these reasons, it is important to develop text-independent voice conversion techniques that can be trained using non-parallel databases.

During the recent years, some solutions for text-independent voice conversion have been proposed [2][3][4]. The method proposed in [2] utilized a vocal tract length normalization (VTLN) based mapping function. Speech recognition has also been applied for helping data alignment procedure [3]. Many techniques align training data based on similarity measurement, such as spectral distance. According to [5], the more similar corresponding source and target vectors are, the less speaker-dependent information can be trained from them. Thus, the use of similarity criterion has its limitations. In order to improve the performance of text-independent voice conversion, phonetic content information should also be utilized.

As is widely known, Hidden Markov Models (HMMs) have been successfully applied to speech recognition systems due to their excellent ability to characterize the spectral parameter sequence as well as to model the phonetic structure. In this paper, HMMs are used to represent the phonetic structure of training speech. The transformation between the source and target characteristics is accomplished by establishing a mapping between the source and target HMM states using phonetic cluster adaptation induced linear transform. In order to build a continuous state alignment between source and target speech using phonetic information, the distribution of the common phoneme clusters between the source and target speech are regarded as anchors when establishing a mapping between the acoustic spaces of the source and target speakers. Several nearest phonetic clusters to each vector are taken into account simultaneously in the mapping for continuity. The proposed method can be applied to both intra-lingual and cross-lingual text-independent voice conversion.

The rest of this paper is organized as follows. In Section 2, the state alignment scheme based on phoneme cluster adaptation is derived. Section 3 describes the system based on the proposed state alignment method. Experimental results and discussion are given in Section 4. Finally, concluding remarks are made in Section 5.

2. STATE ALIGNMENT USING PHONEME CLUSTER ADAPTATION

2.1. Different combination of phoneme sets

The state alignment method proposed in this paper utilizes a shared phoneme set for establishing the alignment between source and target states. Sometimes the phoneme sets of source and target speech can be different from each other, e.g. in incomplete speech database or cross-lingual cases. According to different components of source and target phoneme sets, there are generally five kinds of source and target phoneme set combination as listed below:

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. $S_x\{X\}:S_y\{Y\}$

S_x and S_y denote phoneme set of source and target speech respectively. C represents the set of common phonemes that are available in the phoneme sets of both the source and target speech. X , Y represent set of the mismatched phonemes in the source and target phoneme sets respectively. Because the conversion is a transformation from source speech to target speech, target mismatched phonemes do not have much effect to the alignment procedure. Moreover, case 5, where all the phonemes are different between the source and target speech, is not very realistic. Therefore, cases 1 and 2 are the ones that need to be studied in particular in the alignment. The technique proposed in this paper covers both of these cases within a unified framework.

2.2. Phoneme cluster adaptation based state alignment

From mathematic point of view, the training data alignment process in voice conversion aims to find a binary relation between two speech vector sets. In voice conversion, the transform should be phonetically motivated to retain the content of speech. It is also beneficial if the transform is continuous to ensure that smoothness can be preserved in conversion. Given a source vector and its neighboring vectors in the source space, their corresponding targets should also be close to each other. In other words, the conversion function should not distort vector's neighborhood structure in the conversion so that we can obtain continuous converted speech when transforming continuous source speech.

After HMM training, all states that represent same phoneme are grouped as a phoneme cluster. Only the distribution mean of each state is taken into account for simplification. In order to align states based on phonetic information, it is reasonable to assume that states representing the same phoneme from different speakers are more likely to be aligned together. Especially, the distribution of source phoneme cluster should be adapted to

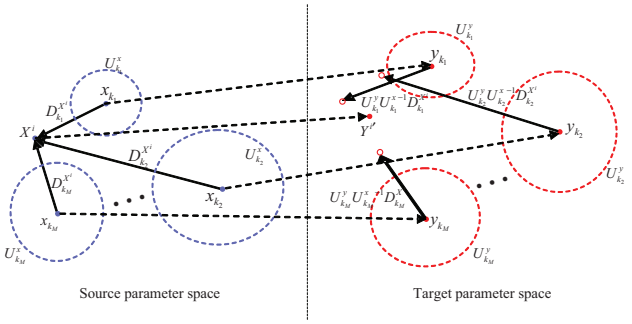

Fig. 1 phoneme cluster adaptation based state alignment

that of the corresponding target phoneme cluster. Based on this assumption, we map the distributions of the common phonemes from source and target speech in order to create the mapping between the source and target acoustic spaces.

In conventional intra-lingual scenarios, the source and target speech share the same phoneme set and all the phoneme clusters are considered. In cross-lingual case, or more generally in cases that correspond to case 2 in Section 2.1., we only deploy the clusters of the phonemes which belong to C for the adaptation.

Figure 1 illustrates the general idea of the state alignment scheme using phoneme cluster adaptation. Firstly, the distribution of each phoneme cluster is calculated. The distribution of each cluster is considered a single Gaussian so it is represented by mean and covariance of states belonging to that phoneme cluster. The acoustic parameter vectors corresponding to distribution of these phoneme clusters are denoted as $\{x_i,U_i^x\}$, $\{y_i,U_i^y\}$, $(i=1,...,N)$ for the source and target spaces, respectively. N denotes the number of common phonemes.

The acoustic parameter used in the mapping is the line spectral frequency (LSF). A perceptual spectral distance is employed as follows [6]:

$$h_k=\frac{1}{\arg\min(\left|lsf_k^1-lsf_{k-1}^1\right|,\left|lsf_k^1-lsf_{k+1}^1\right|)},\quad k=1,\ldots,P$$

$$Dis_{perceptual}(lsf^1,lsf^2)=\sum_{k=1}^{P}h_k\left|lsf_k^1-lsf_k^2\right| \qquad (1)$$

Then for each source state, we can find its M nearest source clusters $\{x_{k_j},U_{k_j}^x\},(k_j\in\{1,\ldots,N\},j=1,2,\ldots,M)$, and consider these nearest M clusters in order to generate a smooth and stable mapping. Residual vectors $D_{k_j}^{X^i}$ between the state mean vector X^i (the superscript means this state belongs to phoneme i) and centroids of nearest M source clusters are calculated. After this, each source state X^i can be described as:

$$X^i=\sum_{j=1}^{M}w_j^{X^i}(x_{k_j}+D_{k_j}^{X^i}), \qquad (2)$$

where,

$$D_{k_j}^{X^i} = X^i - x_{k_j}, (k_j \in \{1,\ldots,N\}, \quad j = 1,2,\ldots,M) \tag{3}$$

$$w_j^{X^i} = \frac{1}{\left|D_{k_j}^{X^i}\right|} \Big/ \sum_{j=1}^{M} \frac{1}{\left|D_{k_j}^{X^i}\right|}. \tag{4}$$

According to the assumption mentioned above, we adapt phoneme clusters from source and target by considering both mean and covariance of member states. Thus, the reference mapped target vector $Y^{i'}$ corresponding X^i are calculated by adapting each residual vector $D_{k_j}^{X^i}$ from the source cluster to the corresponding target cluster.

$$Y^{i'} = \sum_{j=1}^{M} w_j^{X^i} (y_{k_j} + U_{k_j}^{y} \cdot U_{k_j}^{x\,-1} \cdot D_{k_j}^{X^i}) \tag{5}$$

Finally, the state-book can be created by finding the target state Y based on the reference target state $Y^{i'}$ as

$$Y = \arg\min_{\{Y\}}(Dis_{perceptual}(Y^{i'}, Y)) \tag{6}$$

The reference target state $Y^{i'}$ is not directly used because the calculation in Equation (5) may bring some distortion to the spectral structure which may cause degradation in converted speech quality.

The entire state alignment process adapts the structure of the source parametric space to the target parametric space based on adaptation of phoneme clusters. The weighted linear transform utilizes several nearest clusters. The contribution of each phoneme cluster depends on the perceptual distance to the given vector.

3. APPLICATION FOR TEXT-INDEPENDENT VOICE CONVERSION SYSTEM

Once the state mapped codebook is established, we can use it to build a text-independent voice conversion system [7]. Figure 2 shows a flow chart of the complete system. The system can be separated into two parts: training and conversion. The training part includes data alignment and state mapped codebook building. In this paper, the proposed weighted linear transform based on phoneme cluster adaptation is deployed. International Phonetic Alphabet (IPA) is used to represent the phonetic set for both source and target speech. Thus, the common phonetic set can be easily set for both intra-lingual and cross-lingual cases.

In the conversion part, the feature vector sequence extracted from the source speech is labeled by a sequence of source state indices. Then a sequence of target state index is generated based on the source state index sequence according to the state-book made in the training step. This part operates in the similar manner as the conventional codebook based conversion.

Given the target state index sequence obtained above, a continuous speech parameter sequence can be generated

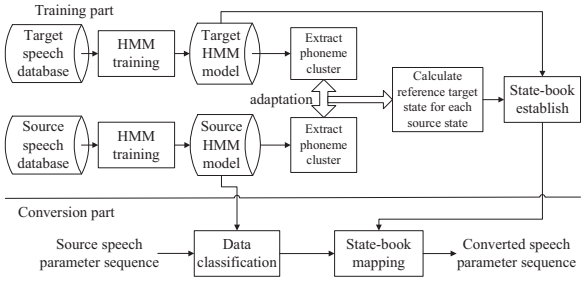

Fig. 2 Flow chart of proposed conversion system

using the parameter generation algorithm [8] developed for the HMM based speech synthesis system (HTS). Dynamic features are taken into account.

We employ a STRAIGHT-based vocoder [9] to extract acoustic features from speech and to synthesize speech from converted parameters representing the vocal tract and vocal source information. The STRAIGHT analysis generates spectra having the dimensionality of 513. For the conversion, 24th order line spectral frequency (LSF) coefficients are then calculated from the spectrum. The complete feature vector in the conversion consists of 24-order LSF, gain and their dynamic counterparts (first and second derivatives).

This paper mainly focuses on spectral conversion. The excitation features are converted by a simple mean-variance transformation which converts the pitch contour of the source speech into the converted pitch contour having characteristics of the target speech. The pitch transform is calculated as:

$$f_y = \mu_y + \frac{\sigma_y}{\sigma_x}(f_x - \mu_x) \tag{7}$$

where (μ, σ) are mean and variance of pitch respectively.

4. EXPERIMENTS

4.1. Subjective evaluations

The performance of the state alignment algorithm proposed in this paper was evaluated in a listening test. For testing intra-lingual text-independent voice conversion, 180 non-parallel British English sentences from source and target speaker were employed for the HMM training. The triphone HMM training resulted in 3099 states for the source speaker and 2985 states for the target speaker. The states of the source and target speakers are used to establish the state-book. For testing cross-lingual voice conversion, 180 Mandarin Chinese sentences of the source speaker and 180 British English sentences from target speaker were used in the training that generated 3595 states for the source speaker and 2985 states for the target speaker.

Each HMM was defined as 3-state left-to-right with no skip. The acoustic features were composed of 24th-order LSF coefficients obtained by the STRAIGHT algorithm

Table 1. Results from ABX test

ABX	Baseline intra-lingual	Proposed intra-lingual	Proposed cross-lingual
f2m	0.76	0.84	0.81
f2f	0.55	0.68	0.63

Table 2. Results from MOS test

MOS	Baseline intra-lingual	Proposed intra-lingual	Proposed cross-lingual
f2m	2.3	2.7	2.2
f2f	2.6	2.5	2.3

with a 5ms shift. Finally the spectrum parameter vector consisted of 25-order LSF coefficients with gain, as well as delta and delta-delta coefficients for the dynamic comparison and HTS based parameter generation.
The subjective test was performed with 9 subjects participating. Each listener was asked to evaluate the identity conversion performance in an ABX test and speech quality in a MOS test.

Tables 1 and 2 show the results of evaluation. The baseline system is a system which doesn't consider phonetic information but has similar other parts in state alignment process [7] comparing to technique proposed in this paper. The experiments were designed to compare the voice conversion performance with and without the phonetic information for state mapping. The results of baseline system are generated only with intra-lingual case.

4.2. Discussion

As shown in Table 1, the proposed technique offers better conversion performance than the baseline system. The proposed transformation utilizing phonetic information can retain the content of input speech and offer a mapping that reflects the relation of speaker identities of the source and target speakers.

The number of the nearest phoneme clusters has effect to the quality of converted speech. When the number of considered nearest clusters is too small, the mapping causes more discontinuous and unstable. When the number is too large, the converted speech may sound muffled caused by the spectral distortion generated from "over-average". The optimal number of considered clusters needs to be estimated separately in different use cases. The results shown in Tables 1 and 2 are given with the experimentally found optimal numbers.

There are some phonemes which can't be mapped to target language phoneme set in the cross-lingual case as shown in Section 2.1. We mapped them to imaginary phonemes in target space by weighted linear transform. This approach solves the problem but still the distortion in states belonging to these phonemes is bigger than in the states of common phonemes. Thus, it is understandable that the quality of converted speech in the cross-lingual case is worse than in the intra-lingual case.

5. CONCLUSION

In this paper, we take phonetic information into account for guiding training data alignment in text-independent voice conversion. States representing the same phoneme are grouped together as a phoneme cluster. By considering the corresponding phoneme clusters from source and target speech, a state mapped codebook is established using weighted linear transform. For each source vector, several nearest clusters are considered in the mapping in order to generate a continuous and stable transform. Experimental results indicate that the proposed scheme is a promising technique for the text-independent voice conversion. The proposed use of the phonetic information increases the similarity between converted speech and target speech. The proposed method can be applied in both intra-lingual and cross-lingual cases.

6. ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (No. 60873160) and Nokia funding.

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. Sündermann, H. Ney, and H. Höge, "VTLN-Based Cross-Language Voice Conversion," in Proc. of the ASRU'03.
[3] H. Ye and S. J. Young, "Voice Conversion for Unknown Speakers," in Proc. of the ICSLP'04.
[4] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. "Text-Independent Voice Conversion Based on Unit Selection", in Proc. of the ICASSP'06.
[5] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion", in Proc. of ICSLP, 2006.
[6] R. Laroia, N. Phamdo, N. Farvardin. "Robust and Efficient Quantization of Speech LSF Parameters Using Structured Vector Quantizers" In Proc. of ICASSP, 1991
[7] M. Zhang, J. Tao, J. Tian, X. Wang, "Text-independent voice conversion based on state mapped codebook," in Proc. of ICASSP 2008.
[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 2000.
[9] H. Kawahara, I. Masuda-Katsuse and A. deCheveigne, "Restructuring speech representations using pitch-adaptive time frequency smoothing and instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999