AA SPECTRAL SPACE WARPING APPROACH TO CROSS-LINGUAL VOICE TRANSFORMATION IN HMM-BASED TTS

Hao Wang[1], *Frank Soong*[1,2] *and Helen Meng*[1]

[1] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Speech Group, Microsoft Research Asia, Beijing, China

ABSTRACT

This paper presents a new approach to cross-lingual voice transformation in HMM-based TTS with only the recordings from two monolingual speakers in different languages (e.g. Mandarin and English). We aim to synthesize one speaker's speech in the other language. We regard the spectral space of any speaker to be composed of universal elementary units (i.e. tied-states) of speech in different languages. Our approach first forces the spectral spaces of the two speakers to have the same number of tied-states. Then we find an optimal one-to-one tied-state mapping between the two spectral spaces. Hence, the mapped speech trajectory in the spectral space of the target speaker can be found according to that generated in the spectral space of the reference speaker. Consequently, we can synthesize high-quality speech for the target monolingual speaker's voice in the other language. This can also be used as training data for a new TTS system.

Index Terms— cross-lingual, voice transformation, spectral space warping, HMM-based TTS

1. INTRODUCTION

With the sound development of speech technology in recent years, various speech service and applications are beginning to have an impact on people's daily lives. Text-to-speech (TTS) synthesis is widely used to output response and feedback in voice user interfaces. However, a cross-lingual TTS where a well-trained TTS in one language is converted to a new monolingual speaker's voice in another language remains a challenge. Such technology would be of great use in many cross-lingual applications such as computer-assisted language learning (CALL), where a user can listen to the synthesized utterances in the foreign language of his/her own voice.

This paper deals with the following problem: we have a monolingual speaker (i.e. target speaker) of one language (e.g. English) and we want to build a personalized synthesizer that enables this target speaker to speak in another language (e.g. Mandarin Chinese). The resources we have are English recordings from this target speaker and Mandarin recordings from a reference speaker. Our approach to cross-lingual voice transformation is applicable to different language pairs. But hereafter in this paper, we will simply refer to the English and Mandarin language pair (in the context of enabling a monolingual English speaker to speak Mandarin) for the sake of clarity.

Previous approaches to cross-lingual voice transformation include the state mapping approach proposed by Qian et al. [1]. They used recordings from a bilingual (English and Mandarin) speaker to build two language-specific decision trees separately. Every leaf node (tied-state) in one tree (in Mandarin) can be mapped to its nearest neighbor leaf node in the other tree (in English). This state mapping process bridges across languages. This mapping information was applied to a monolingual English speaker to synthesize her Mandarin speech. Wu et al. [2] proposed a similar approach. The difference is that they established a state mapping between two Average Voice models in different languages instead of language-specific models trained with recordings from a bilingual speaker. Then they used the state mapping information to either map the data or to map the transform in speaker adaptation. A frame mapping based approach was proposed by Qian et al. [3] and He et al. [4]. In both efforts, only recordings in different languages from two speakers were needed. Spectral frequency warping techniques (piecewise-linear warping based upon formant frequency mapping in [3] and bilinear warping in [4]) were used for speaker equalization. Then the warped reference speaker's speech parameter trajectory was used as a guide to select the most appropriate frame from the target speaker's speech and then concatenate the selected frames together to generate the target speaker's speech in a new language.

Our proposed approach is enlightened by all the cited work above which show that there exists universal elementary unit of a speaker's speech in different languages, if the speech segments considered are small enough. Hence, we regard that the spectral space of any speaker is composed of a set of the universal elementary units in different languages. If the spectral spaces of two speakers have the same number of the universal elementary units, then every elementary unit in one spectral space has a unique counterpart in the other spectral space. Thus we should be able to find a one-to-one elementary unit mapping between the spectral spaces of the two speakers. With the mapping information, the spectral space of one speaker can be warped towards the spectral space of the other speaker, which we call "spectral space warping". This idea leads to our approach. In this paper, we treat tied-states as the universal elementary units. Unlike the tied-state mapping in [1] where each independent tied-state in one language is mapped to its nearest tied-state in the other language for the same speaker or tied-state mapping between two Average Voices in [2], our approach aims to find an optimal one-to-one tied-state mapping between the spectral spaces of the two speakers directly. The advantage is that it does not need an intermediate speaker adaptation process, which was used in [3] and [4]. The details of our approach are given in Section 2. A baseline approach for comparison will be presented in Section 3. Experiments and evaluations are presented in Section 4. Conclusions are drawn in Section 5.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

2. CROSS-LINGUAL VOICE TRANSFORMATION IN HMM-BASED TTS

We have a reference speaker's Mandarin corpus and a target speaker's English corpus. We want to use these two corpora to build a TTS which enables the target speaker to speak Mandarin.

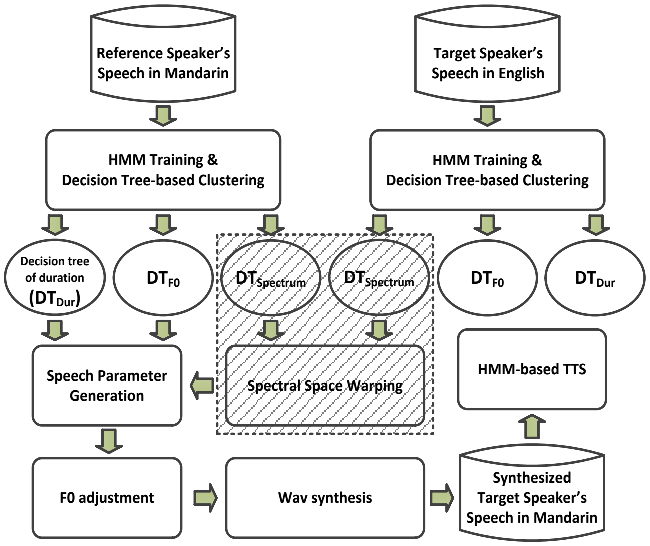
Reference Speaker's Speech in Mandarin
Target Speaker's Speech in English
HMM Training & Decision Tree-based Clustering
HMM Training & Decision Tree-based Clustering
Decision tree of duration (DT_{Dur})
DT_{F0}
$DT_{Spectrum}$
$DT_{Spectrum}$
DT_{F0}
DT_{Dur}
Speech Parameter Generation
Spectral Space Warping
HMM-based TTS
F0 adjustment
Wav synthesis
Synthesized Target Speaker's Speech in Mandarin

Fig. 1: *Schematic diagram of our spectral space warping approach to cross-lingual voice transformation in HMM TTS.*

Fig. 1 is an overview of our approach. For both corpora, we perform decision tree-based clustering through the standard HMM-based TTS training procedure, to obtain single Gaussian context-dependent tied models. During model training, the spectral, F0 and durational features are separated into different streams. We only focus on decision trees of the spectrum ($DT_{Spectrum}$) for spectral space warping. Then, warped models of the spectrum in Mandarin are obtained. Along with the original models of the F0 and duration in Mandarin, the warped models of the spectrum are used to generate speech parameters. The generated F0 needs to be adjusted, and then the target speaker's speech in Mandarin is synthesized and further used to build an HMM-based TTS system.

2.1. Spectral space warping

The core of our approach is to bridge across speakers and across languages by spectral space warping. Regarding the tied-states (leaf nodes of decision trees) as the universal elementary units of speech in different languages, the spectral space of a speaker is represented by a $DT_{Spectrum}$. Two language-specific $DT_{Spectrum}$ are created separately for the two speakers. If the two $DT_{Spectrum}$ have the same number of leaf nodes, we may warp one spectral space towards the other one by finding the optimal one-to-one leaf node mapping between the two $DT_{Spectrum}$. Fig. 2 zooms in on the area of the shaded box in Fig. 1 to illustrate how spectral space warping works. Two steps are involved in spectral space warping:

Step 1: Equalizing the number of leaf nodes between the two language-specific $DT_{Spectrum}$.

Step 2: Finding the optimal one-to-one leaf node mapping between the two $DT_{Spectrum}$.

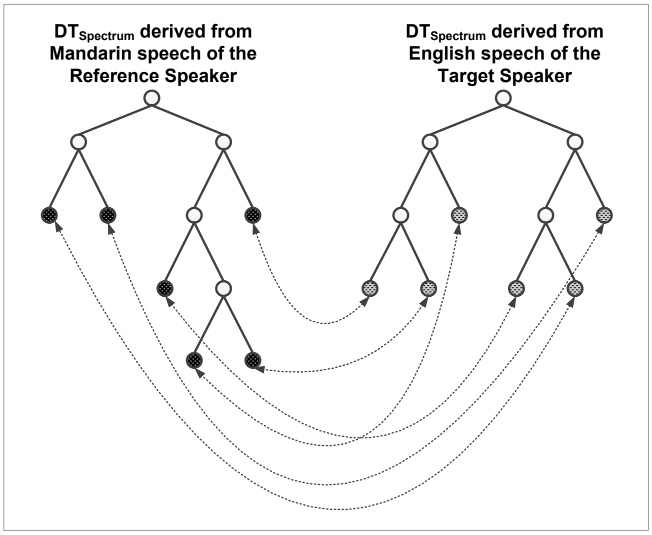
$DT_{Spectrum}$ derived from Mandarin speech of the Reference Speaker
$DT_{Spectrum}$ derived from English speech of the Target Speaker

Fig. 2: *Illustration of spectral space warping — one-to-one leaf node mapping between two language-specific $DT_{Spectrum}$.*

2.1.1. Equalizing the numbers of unvoiced and voiced leaf nodes

All the leaf nodes (tied-states) of $DT_{Spectrum}$ are categorized into three types — unvoiced, voiced and silence. In this paper, we are not interested in the silence leaf nodes. We use the trained language-specific models to obtain the state-level alignment for the recordings of each corpus. We first disregard the silence leaf nodes. For the remaining leaf nodes, we can label them as unvoiced or voiced by examining the speech frames in that node and taking the majority.

We would like to find mappings between the same type of leaf nodes, i.e. an unvoiced leaf node in one tree is mapped to an unvoiced leaf node in the other tree, while a voiced leaf node in one tree is mapped to a voiced leaf node in the other tree. The two language-specific $DT_{Spectrum}$ probably have different numbers of unvoiced and voiced leaf nodes. To equalize the respective numbers of unvoiced and voiced leaf nodes between the two trees, we merge the same type of (unvoiced/voiced) leaf nodes pairwise, for the tree that has more (unvoiced/voiced) leaf nodes, until equality is attained. We perform merging according to the following two principles: (i) the merged pair of nodes must belong to the same state in the (5-state) topology of the phone HMM models; (ii) the merged pair must have the least decrease in log likelihood (which can be calculated with the tied-state means, variances and occupation count [5]).

2.1.2. Finding the optimal one-to-one leaf node mapping

The Kullback-Leibler divergence (KLD) is used to measure the distance between a pair of leaf nodes, each of which comes from one of the two language-specific $DT_{Spectrum}$. This is similar to that in [1] and [2]. Then, we apply the Hungarian algorithm to find the optimal mapping for unvoiced and voiced leaf nodes, respectively. The Hungarian algorithm solves "the assignment problem", that is, given *n* individuals, *n* jobs and an $n \times n$ matrix containing the cost of assigning each individual to a job, it finds a way of assigning exactly one job to one individual, such that the overall total cost is minimized [6,7]. Our problem can be seen as an assignment problem, where the KLD distance is regarded as the cost of assignment. The optimal assignment achieves the minimum total sum of distances (measured as KLD). Under this condition, we expect the neighborhood of tied-states (leaf nodes of $DT_{Spectrum}$) in a spectral space to be preserved after mapping, i.e. neighbors for the tied-states in a spectral space are still neighbors for the corresponding mapped tied-states in the warped spectral space. This guarantees the continuity of the tied-state trajectory in the warped spectral space during the generation of speech parameters.

2.2. Parameter generation for warped Mandarin speech

Leaf nodes of decision trees give rise to tied-states. To synthesize the target speaker's voice speaking the same Mandarin sentences of the reference speaker's corpus, we first use the trained reference speaker's Mandarin model to align the tied-state sequences of all the Mandarin recordings. Then, according to the spectral mapping information obtained in the previous subsection, we replace the spectral tied-states of the Mandarin model by their mapped spectral tied-states of the English model (we do not touch the silence tied-states). Using these mapped spectral tied-states and the aligned state-level durations, we can generate spectrum parameters of the warped Mandarin speech. The F0 parameters are generated using the original Mandarin models of the F0 and adjusted according to the following linear transformation equation:

$$\widehat{F0} = \frac{(F0_r - \mu_r)}{\sigma_r} \cdot \sigma_t + \mu_t \qquad (1)$$

where μ_r, μ_t, σ_r and σ_t are means and standard deviations of the F0 for the corpora of the reference and target speakers, respectively. Using the generated warped speech parameters of spectrum and the adjusted F0, a set of the target speaker's Mandarin utterances can be synthesized. Then we treat the synthesized speech as training data in developing an HMM-based TTS system.

3. BASELINE APPROACH: THE NEAREST TIED-STATE MAPPING

We build a baseline system using vocal tract length normalization (VTLN) and nearest tied-state mapping techniques.

3.1. Frequency warping for speaker equalization

A bilinear transform based VTLN is used to minimize the differences between speakers. The bilinear transform can be represented as follows [4,8]:

$$\psi_\alpha(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} = e^{-j\beta_\alpha(\omega)}, |\alpha| < 1 \qquad (2)$$

where α is the warping factor, ω denotes the input frequency and the frequency transformation $\beta_\alpha(\omega)$ is obtained by making the substitution $z = e^{j\omega}$. We implement the frequency warping in the spectral features of the Mel-cepstral coefficients (MCEPs) [9]. The MCEP features are warped using the following matrix representation [8,10,11]:

$$\boldsymbol{c}_\alpha = \boldsymbol{B}_\alpha \boldsymbol{c} \qquad (3)$$

where $\boldsymbol{c}$ and $\boldsymbol{c}_\alpha$ are the MCEP feature vectors before and after the bilinear warping, $\boldsymbol{B}_\alpha$ is the matrix transformation with the warping factor α. MCEPs already have a bilinear transform with a warping factor $\alpha_1 = 0.42$ approximating the Mel-scale frequency warping. Thus another stage of bilinear transform is needed and can be cascaded with the existing one by combining the warping factors α_1 and α_2 according to the following equation [8]:

$$\alpha = \frac{\alpha_1 + \alpha_2}{1 + \alpha_1 \alpha_2} \qquad (4)$$

We compute the MCEP means of all voiced frames in the Mandarin and English recordings respectively [4]. Then, we estimate the warping factor α_2 by a grid search to minimize the MCEP distortion between the MCEP mean of the English speaker and the warped MCEP mean of the Mandarin speaker that is obtained by Equation (3).

3.2. Nearest tied-state mapping across languages

The reference speaker's Mandarin speech is warped as described in Section 3.1 with the F0 adjusted according to Equation (1). Two language-specific $\text{DT}_{\text{Spectrum}}$ are created separately for the warped reference speaker's Mandarin speech and the target speaker's English speech. All leaf nodes are categorized into three types, i.e. unvoiced, voiced and silence. For each type (unvoiced/voiced), every leaf node of the warped Mandarin $\text{DT}_{\text{Spectrum}}$ has a mapped (nearest) leaf node of the English $\text{DT}_{\text{Spectrum}}$ in the minimum KLD sense. We use the mapping information to generate a set of target speaker's Mandarin utterances and further used to train an HMM-based TTS. Compared with our spectral space warping approach, this baseline approach implements VTLN before spectral tied-state mapping and it does not require a one-to-one mapping between tied-states (leaf nodes of $\text{DT}_{\text{Spectrum}}$) of two spectral spaces.

4. EXPERIMENTS AND EVALUATIONS

4.1. Experimental setup

Our Mandarin (MAN) recordings were made by a female reference speaker in broadcast news reading style. To verify that our proposed approach is applicable to different speaker pairs in cross-lingual voice transformation, two sets of English recordings are selected from the CMU_ARCTIC_BDL (BDL) and the CMU_ARCTIC_SLT (SLT) corpora that were recorded by a male target speaker and a female target speaker, respectively. Thus, we have two speaker pairs in our experiments — MAN-BDL and MAN-SLT. Table 1 shows the numbers of utterances we used for training.

Table 1. *The numbers of utterances used for training.*

Speaker Language	Reference (Female)	Target (Male)	Target (Female)
Mandarin	1000	N/A	N/A
English	N/A	1000	1000

All speech signals are sampled at 16 kHz, windowed with a 25-ms Hamming window, shifted every 5ms. We extract the 24th-order MCEPs plus log-energy and log-F0 together with their delta and delta-delta that are modeled by multi-stream HMMs. Each phone HMM has a five-state, left-to-right topology with single Gaussian, diagonal covariance distributions.

For our spectral space warping approach, after equalizing the numbers of unvoiced and voiced leaf nodes between the two language-specific $\text{DT}_{\text{Spectrum}}$, there are 327 unvoiced and 882 voiced leaf nodes for MAN-BDL speaker pair, 278 unvoiced and 900 voiced leaf nodes for MAN-SLT speaker pair. The average KLDs we obtain for the optimal one-to-one mapping between each speaker pair are 37.87 and 37.94, respectively.

For the baseline approach, we use the estimated warping factors −0.31 and 0.42 to warp the reference speaker's Mandarin speech for each speaker pair, respectively. Then, for MAN-BDL speaker pair, we have 383 unvoiced and 1129 voiced leaf nodes for the warped Mandarin $DT_{Spectrum}$, 327 unvoiced and 882 voiced leaf nodes for the English $DT_{Spectrum}$. For MAN-SLT speaker pair, we have 363 unvoiced and 1269 voiced leaf nodes for the warped Mandarin $DT_{Spectrum}$, 278 unvoiced and 900 voiced leaf nodes for the English $DT_{Spectrum}$. The average KLDs obtained for the optimal one-to-one mapping between each speaker pair are 21.51 and 24.93, respectively.

In the synthesis stage, speech parameters including MCEPs and log-F0 are obtained by the maximum likelihood parameter generation algorithm [12], and later used for waveform generation through the MLSA filter.

4.2. Subjective evaluation

Subjective evaluations of the synthesized Mandarin speech are conducted to determine mean opinion scores (MOS) for speech quality and speaker similarity. For each speaker pair, 15 Mandarin utterances are synthesized by the proposed spectral space warping approach and the baseline approach, respectively. Six experienced listeners are asked to rate speech quality of each synthesized utterance on a 5-point scale (1:bad, 2:poor, 3:fair, 4:good, 5:excellent). The same five-point scale is applied to the speaker similarity test where each of the same six subjects is asked to measure how similar the voice of the synthesized speech is to the target speaker's voice. The results of the evaluations are shown in Table 2.

Table 2. *Speech Quality and Speaker Similarity scores with the 95% confidence interval (CI) of Mandarin utterances synthesized by our proposed spectral space warping approach and the baseline nearest tied-state mapping approach.*

Speaker Pair	Approach	Speech Quality (95% CI)	Speaker Similarity (95% CI)
MAN-BDL	Spectral Space Warping	3.28 (±0.20)	3.16 (±0.21)
	Nearest Tied-State Mapping	2.74 (±0.19)	2.22 (±0.19)
MAN-SLT	Spectral Space Warping	3.36 (±0.13)	2.83 (±0.20)
	Nearest Tied-State Mapping	1.97 (±0.17)	2.13 (±0.15)

The results of the subjective evaluations show that our proposed spectral space warping approach significantly outperforms the baseline approach in both speech quality and speaker similarity. A possible reason that our proposed approach achieves better speech quality is that in the baseline approach, the leaf nodes are assumed independent of one another during the mapping process, but our proposed approach uses the Hungarian algorithm (See Section 2.1.2) to warp all the (unvoiced/voiced) leaf nodes as a whole towards another set of leaf nodes, which implies that no independence assumption is used. Therefore, to a certain extent, the proposed approach can probably guarantee the continuity of the tied-state trajectory in the warped spectral space during the generation of speech parameters, which makes the generated speech sound smoother. However, the gap between the two speaker similarity scores for each speaker pair is not expected to be this wide since despite their different tied-states mapping schemes, both approaches use the spectral tied-states trained from the target speaker's English recordings to generate Mandarin speech. Such a wide gap is possibly due to the fact that no monolingual English target speaker's Mandarin recordings can be provided for comparison and it is difficult to compare across languages. Thus the subjects' rating on speaker similarity may tend to be greatly affected by speech quality.

5. CONCLUSIONS

In this paper, we propose a spectral space warping approach to cross-lingual voice transformation in HMM-based TTS. This proposed approach only requires monolingual corpora in two different languages from two speakers, and can directly warp the spectral space of the reference speaker towards the spectral space of the target speaker. No intermediate speaker adaptation process is required. Subjective evaluations with MOS show that our proposed spectral space warping approach significantly outperforms the nearest tied-state mapping baseline approach in both speech quality (more than 0.5 absolute improvement on a 5-point scale) and speaker similarity (more than 0.7 absolute improvement on a 5-point scale).

6. ACKNOWLEDGEMENTS

This project is partially sponsored by a grant from the Hong Kong SAR Government General Research Fund (Project Number: 14205814).

7. REFERENCES

[1] Y. Qian, H. Liang and F. K. Soong, "A Cross-language State Sharing and Mapping Approach to Bilingual (Mandarin–English) TTS", IEEE Transactions on Audio, Speech, and Language Processing, VOL. 17, NO. 6, pp.1231-1239, 2009.

[2] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State Mapping based Method for Cross-lingual Speaker Adaptation in HMM-based Speech Synthesis", in Proc. of Interspeech, pp. 528–531, 2009.

[3] Y. Qian, J. Xu, and F. K. Soong, "A Frame Mapping based HMM Approach to Cross-lingual Voice Transformation", in Proc. ICASSP, pp. 5120-5123, 2011.

[4] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a Monolingual Speaker into Multilingual for a Mixed-language TTS", in Proc. Interspeech, 2012.

[5] S. J. Young et al., The HTK Book (for HTK Version 3.4). Cambridge, U.K.: Cambridge University Press, 2009.

[6] H. W. Kuhn, "The Hungarian algorithm for the assignment problem", Naval Research Logistics Quarterly, 2:83–97, 1955.

[7] J. Munkers, "Algorithms for the assignment and transportation problems", Journal of the Society for Industrial and Applied Mathematics, 5:32–38, 1957.

[8] L. Saheer, P. N. Garner, J. Dines and H. Liang, “VTLN Adaptation for Statistical Speech Synthesis”, Proc. of ICASSP, pp. 4838-4841, March 2010.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech”, in ICASSP, pp. 137–140, 1992.

[10] N. Nocerino, F. K. Soong, L. R. Rabiner, D. H. Klatt, “Comparative study of several distortion measures for speech recognition”, In Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Atlanta, GA, pages 25-28. Apr., 1985.

[11] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space”, IEEE Transactions on Speech and Audio Processing,, vol. 13, pp. 930–944, 2005.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, in ICASSP, pp. 1315–1318, 2000.