INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Cross-Lingual Speaker Adaptation for Statistical Speech Synthesis Using Limited Data

Seyyed Saeed Sarfjoo, Cenk Demiroglu

Electrical and Computer Engineering Department
Ozyegin University, Istanbul, Turkey

saeed.sarfjoo@ozu.edu.tr, cenk.demiroglu@ozyegin.edu.tr

Abstract

Cross-lingual speaker adaptation with limited adaptation data has many applications such as use in speech-to-speech translation systems. Here, we focus on cross-lingual adaptation for statistical speech synthesis (SSS) systems using limited adaptation data. To that end, we propose two techniques exploiting a bilingual Turkish-English speech database that we collected. In one approach, speaker-specific state-mapping is proposed for cross-lingual adaptation which performed significantly better than the baseline state-mapping algorithm in adapting the excitation parameter both in objective and subjective tests. In the second approach, eigenvoice adaptation is done in the input language which is then used to estimate the eigenvoice weights in the output language using weighted linear regression. The second approach performed significantly better than the baseline system in adapting the spectral envelope parameters both in objective and subjective tests.

Index Terms: statistical speech synthesis, speaker adaptation, nearest-neighbor, cross lingual speaker adaptation, eigenvoice adaptation

1. Introduction

Cross-lingual speaker adaptation (CLSA) for statistical speech synthesis is used for adapting to a target speaker in an output language, using adaptation data from the speaker in an input language. CLSA algorithms have many applications such as deployment in speech-to-speech translation systems [1, 2].

In a commonly used approach, state mapping between the acoustic models of the input and output languages has been done [3, 4, 5] and adaptation data or transformation functions are mapped to output language states using the state-map.

Mismatch between the language-dependent average voice models (AVMs) degrade the adaptation performance in transformation mapping [6, 7] since speaker-specific transformations of similar states may be different in different languages and voice models. To alleviate the problem, transform mapping using shared decision tree context clustering is proposed in [8] where not only acoustic-similarity but also contextual similarity of states are taken into account during mapping.

AVM can also be trained using data from multiple languages and adapted to a target speaker that speaks one of the languages used in training [9]. However, the adaptation performance is not always sufficient because some of the leaf nodes of the decision tree is trained with only one of the languages in the training set. A speaker and language factorization technique to alleviate the problem is proposed in [10] where cluster adaptive training (CAT) is used to build an AVM using data from different languages. For a target language, cluster weights are estimated for building a language-dependent model before adapting to the speakers of that language.

A language-independent approach is proposed in [11] by using a manually-developed, language-independent space of perceptual characteristics (PC) [11]. In this method, a new target speaker model in the input language speaker space is first projected to the PC space and then projected from the PC space to the output language.

A factor analysis based CLSA using bilingual speech data is proposed in [12]. In this method, model parameters representing language-dependent acoustic features and factors representing speaker characteristics are simultaneously optimized using a maximum likelihood approach and a single statistical model is trained using bilingual speech data. Performance improves compared to training each eigenvoice spaces independently.

Recently, deep neural networks methods have been used for training multilingual acoustic models [13, 14, 15]. However, such models need significant amount of data for training and adaptation whereas the focus here is adaptation with limited data.

In this paper, we focused on cross-lingual adaptation when only a few utterances are available from a target speaker. We propose two methods to outperform the baseline system in the limited data case. In the first method, we propose a speaker-specific state-mapping algorithm. In that approach, a bilingual database was used in which data in both English and Turkish were available from the same speakers. After generating speaker-adapted models in both languages, speaker-specific state-mapping is done for each of the bilingual speakers in the speaker pool. Then, for a target speaker outside the pool, a nearest-neighbor is found in the pool and state-map of that nearest-neighbor is used for adaptation. Performance was found to be significantly higher than the baseline target-independent state-mapping algorithm for the excitation parameters both in objective and subjective tests.

In the second method, an eigenvoice approach is used for rapid adaptation. Eigenvoice weights computed for the input language are linearly transformed into output language weights. Transformation matrix is learned using the bilingual database with a least-squares approach. To further boost the performance, weighted linear regression is done where weights are estimated based on similarity of the target speaker to the training speakers.

In our previous work, we proposed a Bayesian eigenvoice adaptation algorithm for rapid intralingual adaptation where nearest-neighbors were used for estimating the hyperparameters of the a priori probability distribution function (pdf) of

This work was supported by FP7 Marie Curie International Reintegration Grant CLSASTS.

http://dx.doi.org/10.21437/Interspeech.2016-345

the eigenvoice weights [16]. Here, we exploit the similarity of the target speaker to nearest-neighbors for more accurate state-mapping and eigenvoice weight transformation as introduced above.

This paper is organized as follows. Baseline cross-lingual speaker adaptations method is described in Section 2. Proposed algorithms are described in Section 3. Experimental results are presented and discussed in Section 4. Finally, conclusion is done in Section 5.

2. Baseline Adaptation algorithm

Cross-lingual speaker adaptation through state-mapping is one of the most successful methods [5]. In that approach, first, two average voice models (AVM) in the input and the output languages are trained. Then, a state-map between these two models is established. Kullback-Leibler divergence (KLD) [17] is typically used for computing the distance between states. All input language AVM states are mapped to one of the output language AVM states using the KLD measure. Once the adaptation data is mapped to output language AVM states, any of the existing intra-lingual adaptation algorithms such as CMLLR or CSMAPLR can be used.

After state-mapping, in one approach, adaptation data that is collected for the input language AVM states can be mapped to output language AVM states for adaptation. In a second approach, speaker transformation matrices can be mapped to output AVM states instead of the data. The data mapping approach has better speaker similarity, and the transform mapping approach has better speech quality after cross-lingual speaker adaptation [5]. Because our goal was to improve speaker similarity, we used data mapping in the baseline system.

3. Proposed algorithms

3.1. Data mapping method using nearest-neighbors

The baseline algorithm does state-mapping using the AVMs once and uses the same map for all target speakers. However, data mapping can be done more effectively if state-mapping is done in a speaker-specific manner. To that end, we created a bilingual database where data in both the input and the output languages were collected from each training speaker. Then, intra-lingual speaker adaptation was done and models in the input and output languages were generated for those training speakers. For each bilingual speaker s_i in the pool of training speakers, a seperate map $\boldsymbol{M}_i$ was produced between the speaker-dependent models for the input and output languages.

In the next step, the problem is to select which one of those pre-trained maps to use for adaptation of a target speaker. Here, similarity between the target speaker and the training speakers was used to select the nearest training speaker, s_{nn}, to the target speaker s_{tar}. As the similarity measure, we use the L_2 distance $(\boldsymbol{\mu}_{nn} - \boldsymbol{\mu}_{tar})^T(\boldsymbol{\mu}_{nn} - \boldsymbol{\mu}_{tar})$ where $\boldsymbol{\mu}_{nn}$ is the supervector of mean vectors of the states that are in the acoustic model of the nearest training speaker in the input language. Similarly, $\boldsymbol{\mu}_{tar}$ is the supervector of the target speaker.

Once s_{nn} is selected, the state-map $\boldsymbol{M}_{nn}$ is used for mapping the adaptation data to output language states. Then, similar to the baseline approach, intra-lingual adaptation is performed.

3.2. Eigenvoice adaptation

Eigenvoice-based adaptation has been used for SSS when limited adaptation data is available [16, 18]. In this approach, given a set of R eigenvectors, $\boldsymbol{e}_r$, the supervector for speaker s is modelled with

$$\boldsymbol{\mu}^{(s)} = \boldsymbol{\mu}_{\text{SI}} + \boldsymbol{E}\boldsymbol{w}_s + \boldsymbol{\epsilon}_s \tag{1}$$

where $\boldsymbol{E} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ ...\ \boldsymbol{e}_R]$, $\boldsymbol{w}_s \in \mathbb{R}^{R\times 1}$ is weight vector of the speaker s, and $\boldsymbol{\epsilon}_s$ is the approximation error. Principal component analysis (PCA) is used here for training $\boldsymbol{E}_{in}$ and $\boldsymbol{E}_{out}$ for the input and output languages respectively. The supervector $\boldsymbol{\mu}^{(s)} = [\boldsymbol{\mu}_1^{(s)^T}\ \boldsymbol{\mu}_2^{(s)^T}\ ...\ \boldsymbol{\mu}_{N_{st}}^{(s)^T}]^T$ where N_{st} is the total number of states in all decision trees in the acoustic model.

In the ML-based eigenvoice approach, given some adaptation data $\boldsymbol{\chi}_a = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(N_{o,s})}\}$, $N_{o,s}$ is the total number of observations from speaker s, the likelihood function

$$p(\boldsymbol{\chi}_a|\boldsymbol{w}_s, \boldsymbol{E}) \propto exp\Bigg(-\frac{1}{2}\sum_{c=1}^{N_{st}}\sum_{i=1}^{N_c^{(s)}}(\boldsymbol{x}_c^{'(i)} - \boldsymbol{E}_c\boldsymbol{w}_s)^T \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}_c^{'(i)} - \boldsymbol{E}_c\boldsymbol{w}_s)\Bigg) \tag{2}$$

where $\boldsymbol{E}_c \in \mathbb{R}^{F\times R}$ is the c^{th} block of the $\boldsymbol{E}$ matrix corresponding to state c, F is the size of $\boldsymbol{\mu}_c$, $\boldsymbol{x}_c^{'(i)} = \boldsymbol{x}_c^{(i)} - \boldsymbol{\mu}_c$, $\boldsymbol{x}_c^{(i)}$ is i^{th} observation that is aligned with state c, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the speaker independent mean vector and covariance matrix of the Gaussian emission (pdf) of state c, and $N_c^{(s)}$ is the number of observations aligned with state c for speaker s. Here, Viterbi alignment is used for likelihood estimation.

Weight vector of speaker s, $\boldsymbol{w}_s$, is estimated as

$$\hat{\boldsymbol{w}}_s = \boldsymbol{G}_w^{(s)^{-1}}\boldsymbol{k}_w^{(s)} \tag{3}$$

where

$$\boldsymbol{G}_w^{(s)} = \sum_{c=1}^{N_{st}} N_c^{(s)}\boldsymbol{E}_c^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{E}_c \tag{4}$$

$$\boldsymbol{k}_w^{(s)} = \sum_{c=1}^{N_{st}} \boldsymbol{E}_c^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{S}_{x,c}^{(s)} \tag{5}$$

$$\boldsymbol{S}_{x,c}^{(s)} = \sum_{i=1}^{N_c^{(s)}} \boldsymbol{x}_c^{'(i)} \tag{6}$$

Eq (3) can be used for calculating the weights $\boldsymbol{w}_{s,in}$ for the input language. However, output language weights $\boldsymbol{w}_{s,out}$ should be estimated to calculate

$$\hat{\boldsymbol{\mu}}_{\text{out}}^{(s)} = \boldsymbol{\mu}_{\text{SI},out} + \boldsymbol{E}_{out}\hat{\boldsymbol{w}}_{s,out}. \tag{7}$$

To estimate $\boldsymbol{w}_{s,out}$, a linear regression approach is used here. Weights of the bilingual training speakers for the input and output languages are first computed using Eq (3). Then, a linear regression matrix $\boldsymbol{A}$ is trained such that $\boldsymbol{w}_{s,out} = \boldsymbol{A}\boldsymbol{w}_{s,in} + \boldsymbol{\epsilon}_s$. Least-squares approach is used for training $\boldsymbol{A}$. Once A is trained using the bilingual training speakers, it is used for transforming the eigenvoice weights of the target speaker in the input language into weights in the output language.

3.2.1. Speaker-specific Regression of Eigenvoice Weights

Linear regression is used here to transform the eigenvoice weights since nonlinear methods such as neural networks require significantly more data and collection of large bilingual databases is expensive. However, to improve the performance of the linear model, the $\boldsymbol{A}$ matrix can be constructed in a target-specific manner. To that end, we propose a weighted linear regression approach as described below.

Given adaptation data from a target speaker, a speaker-specific $\boldsymbol{A}_{tar}$ matrix is computed using:

$$\boldsymbol{A}_{tar} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \sum_{i=1}^{N_p} \boldsymbol{\epsilon}_{i,tar}^T \boldsymbol{\epsilon}_{i,tar} \tag{8}$$

where N_p is the number of training speakers and

$$\boldsymbol{\epsilon}_{i,tar} = \boldsymbol{L}_{tar}(i).(\boldsymbol{w}_{out}(i) - \boldsymbol{A}\boldsymbol{w}_{in}(i)) \tag{9}$$

where $\boldsymbol{L}_{tar}(i)$ is the error weight for the i^{th} training speaker, $\boldsymbol{w}_{out}(i)$ is its eigenvoice weight in the output language and $\boldsymbol{w}_{in}(i)$ is its eigenvoice weight in the input language.

Error weights in the output language are computed in three steps. In the first step, distance of the target speaker to training speakers is computed by using the L_2 distance between the mean supervectors as described in Section 3.1. Then, in the second step, distances are compressed and normalized with

$$\boldsymbol{L}_{tar}(i) = 1 - \log_2\left(\frac{d(i) - d_{min}}{d_{max} - d_{min}} + 1\right) \tag{10}$$

where $d(i)$ is the distance of i^{th} training speaker to the target, d_{max} is the maximum distance and d_{min} is the minimum distance of training speakers to the target. Only the best N_{nn} speakers are used in computing the regression matrix, $\boldsymbol{A}_{tar}$. In the last step, weight vector of the target speaker in the output language is computed with $\hat{\boldsymbol{w}}_{tar,out} = \boldsymbol{A}_{tar}\boldsymbol{w}_{tar,in}$.

4. Experiments

4.1. Experiment Setup

All systems in the experiments were trained with 78 dimensional vectors consisting of 24 Mel-Generalized Cepstrum Coefficients (MGCs), 1 log-energy, 1 log-F0 (LF0) coefficient, and their delta and delta-delta parameters. 25 msec analysis window with 5 msec frame rate is used for feature extraction. Phonemes are modeled with 5 state Hidden Semi-Markov Models (HSMM).

Turkish was used as the input language and English was used as the output language. Four speakers from the CMU-ARCTIC database with 1130 utterances from each of them were used to train the AVM in English. For training the AVM in Turkish, three female speakers with 1100 utterances from each of them were used. For the proposed state-mapping algorithm, a bilingual Turkish-English database is created that has 29 female speakers and 50 utterances per speaker. Turkish and English speaker-dependent models for each speaker were generated using CSMAPLR adaptation and an additional MAP adaptation. Leave-one-out method was used in testing for each one of the 29 training speakers.

More speakers were required for learning the linear regression matrix of the eigenvoice approach. Thus, 59 additional female speakers were used in the eigenvoice experiments. 10 bilingual utterances were recorded from each speaker. Because the performance of eigenvoice adaptation saturates much faster than CSMAPLR, significantly less data was needed from each speaker for the eigenvoice experiments. In these experiments, 2, 3, and 10 dimensional eigenvoices were used. N_{nn} was set to 40 in the weighted linear regression approach.

State-mapping [5] was used as the baseline since, for speaker similarity, it is one of the best performing algorithms in cross-lingual adaptation. Performance was measured with objective and subjective tests as discussed below.

4.2. Objective Measure Tests

Root-mean-square-error (RMSE) is used for objectively measuring the distance between the MGC and LF0 features of synthesized and reference speech samples. To remove the effect of the vocoder, for reference speech samples, synthetic speech from the speaker-dependent models were used rather than natural speech from the target speaker. English AVM is used for modelling the durations [19] to time-align the synthetic and reference states.

For each target speaker, adaptation was performed for 2, 5, and 10 utterances of adaptation data. For each of the 29 adapted models, 40 English sentences from the WSJ1 database were synthesized for testing.

For state-mapping algorithms, both CSMAPLR and CMLLR methods were tested. Results are shown in Figure 1. For the MGC features, CSMAPLR and CMLLR performed similarly. CSMAPLR algorithm is used for the MGC features for the baseline and proposed systems. For the LF0 feature, CMLLR algorithm performed better for the baseline system while the CSMAPLR algorithm performed better for the proposed system. Best algorithm is used for each system.

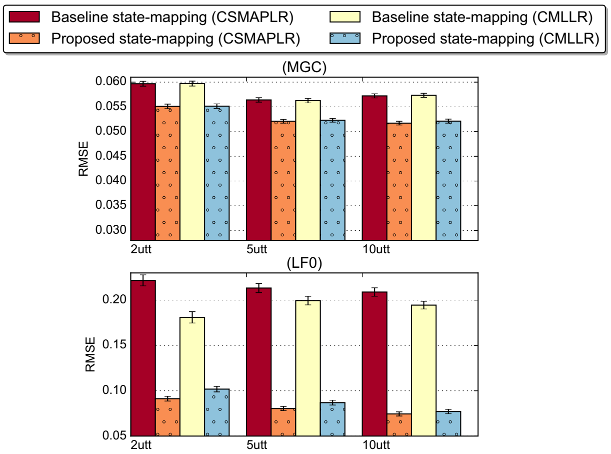


Figure 1: Comparison of objective evaluation (RMSE) of baseline and proposed state-mapping algorithms using CSMAPLR and CMLLR adaptation for the MGC and LF0 features with 95% confidence intervals for 2, 5 and 10 utterances.

For the MGC features, proposed state-mapping algorithm performed better than the baseline algorithm as shown in Figure 2. However, simply synthesizing with the nearest-neighbor without any further adaptation was significantly better than both of the state-mapping algorithms. Eigenvoice approach performed the best. Even though increasing the rank of eigenspace was helpful for intralingual eigenvoice adaptation, shown in Figure 2 for comparison purposes, that was not the case for the proposed methods. We believe that this is related to increased non-linearity and increased need for training speakers in the higher dimensional space.

For the LF0 feature, proposed state-mapping algorithm had the best performance. It also performed better with more adap-

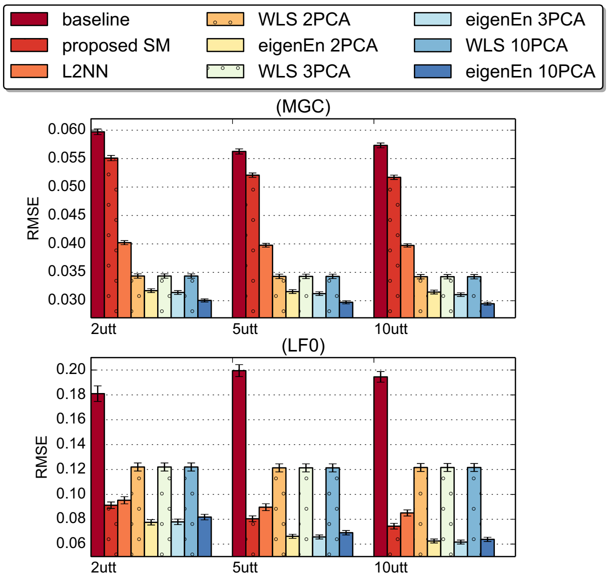


Figure 2: Comparison of objective evaluation (RMSE) of proposed cross-lingual adaptation algorithms for MGC and LF0 features with 95% confidence intervals for 2, 5 and 10 utterances. EigenEn represents intra-lingual eigenvoice adaptation. L2NN indicates synthesis of English sentences with the models of the nearest-neighbor without any further adaptation. 2, 3, and 10 dimensional eigenvoices were used. SM indicates state-mapping.

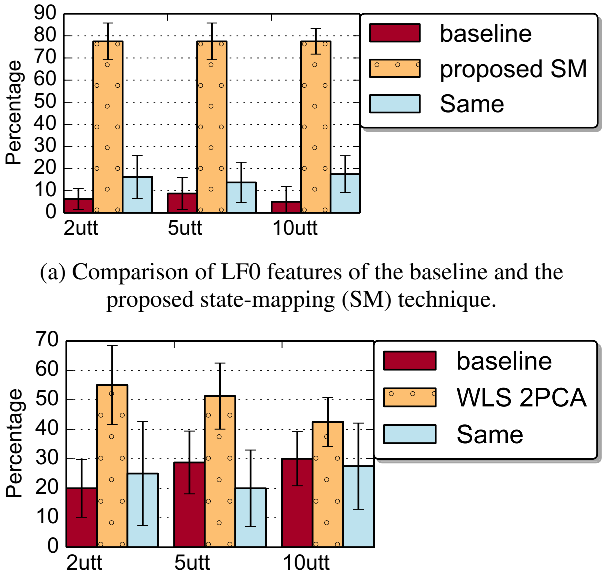
tation data which was not the case with the baseline system. Eigenvoice algorithm also performed better than the baseline. However, its performance was poorer than the proposed state-mapping algorithm. Similar to MGC features, increasing the rank of eigenspace was not helpful. Simply using the nearest-neighbor without any adaptation performed comparable to the proposed state-mapping system especially for the two utterance case.

4.3. Subjective Measure Tests

ABX test is used to subjectively measure the similarity of synthesized samples to the target speaker. Similar to objective measures, in the subjective measure tests, synthetic speech from the speaker-dependent models were used as reference. In the ABX test, listeners prefer sample A or sample B depending on perceived similarity to the reference sample X. A and B samples are synthesized from different adaptation methods randomly and X is the reference sample. 20 target speakers were used and one English sentence from the WSJ database was synthesized for each speaker and each adaptation data size. 8 listeners took the test.

Based on the result of objective experiments, two subjective ABX tests were conducted. In the first experiment, excitation features generated with the baseline and the proposed state-mapping algorithms were compared. MGC features of the baseline model were used in both systems. Preference results are shown in Figure 3a. Proposed state mapping algorithm significantly outperformed the baseline algorithm in listening tests.

In the second experiment, MGC features of the baseline algorithm is compared with the MGC features of the proposed eigenvoice algorithm with 2-dimensional eigenvectors. LF0 features were generated with the proposed state-mapping algorithm in both systems. Preference results are shown in Figure 3b. Eigenvoice algorithm significantly outperformed the baseline algorithm for the 2 and 5 adaptation utterances. However, the difference is not significant in the 10 utterance case. Thus, even though the eigenvoice adaptation technique performed well, its performance saturated fast. The CMLLR technique used in the baseline method, however, did not perform well with minimal data but reached the same performance as the eigenvoice method when 10 utterances were available.



(a) Comparison of LF0 features of the baseline and the proposed state-mapping (SM) technique.

(b) Comparison of MGC features of the baseline and the proposed two-dimensional eigenvoice technique with weighted least-squares (WLS 2PCA).

Figure 3: ABX subjective similarity test results with 95% confidence interval.

AB quality test was also performed to compare the proposed methods with the baseline method. However, a significant quality difference was not found.

5. Conclusion

In this work, we proposed two algorithms for cross-lingual speaker adaptation of statistical speech synthesis models with limited data. The proposed algorithms rely on a bilingual speech database that we collected and exploited to reduce the adaptation data requirements. In the first method, target-specific state-mapping was proposed which significantly outperformed the baseline target-independent state-mapping for the LF0 feature. In the second method, eigenvoice weights estimated in the input language were transformed into the eigenvoice weights in the output language using linear regression. To boost the performance, weighted linear regression was performed by using distance of the target speakers to the training speakers in weight estimation. The eigenvoice weight transformation using weighted linear regression performed significantly better than the baseline state-mapping algorithm for the MGC parameters.

6. References

[1] S. Matsuda, X. Hu, Y. Shiga, H. Kashioka, C. Hori, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai *et al.*, "Multi-lingual speech-to-speech translation system: Voicetra," in *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, vol. 2. IEEE, 2013, pp. 229–233.

[2] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis using kld-based transform mapping," *Speech Communication*, vol. 54, no. 6, pp. 703–714, 2012.

[3] H. Liang, Y. Qian, F. K. Soong, and G. Liu, "A cross-language state mapping approach to bilingual (mandarin-english) tts," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4641–4644.

[4] Y.-N. Chen, Y. Jiao, Y. Qian, and F. K. Soong, "State mapping for cross-language speaker adaptation in tts," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4273–4276.

[5] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis." Interspeech, 2009, pp. 528–531.

[6] H. Liang and J. Dines, "An analysis of language mismatch in hmm state mapping-based cross-lingual speaker adaptation," Idiap, Tech. Rep., 2010.

[7] X. Peng, K. Oura, Y. Nankaku, and K. Tokuda, "Cross-lingual speaker adaptation for hmm-based speech synthesis considering differences between language-dependent average voices," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 605–608.

[8] D. Nagahama, T. Nose, T. Koriyama, and T. Kobayashi, "Transform mapping using shared decision tree context clustering for hmm-based cross-lingual speech synthesis," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an hmm-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[10] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1713–1724, 2012.

[11] V. d. F. Oliveira, S. Shiota, Y. Nankaku, and K. Tokuda, "Cross-lingual speaker adaptation for hmm-based speech synthesis based on perceptual characteristics and speaker interpolation," in *Interspeech*, 2012, pp. 983–986.

[12] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for hmm-based speech synthesis," in *8th ISCA Speech Synthesis Workshop*, 2013, pp. 317–322.

[13] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," pp. 4994–4998, 2015.

[14] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7639–7643.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[16] A. Mohammadi, S. S. Sarfjoo, and C. Demiroglu, "Eigenvoice speaker adaptation with minimal data for statistical speech synthesis systems using a map approach and nearest-neighbors," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 2146–2157, 2014.

[17] P. Liu, F. K. Soong, and J.-L. Thou, "Divergence-based similarity measure for spoken document retrieval," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–89.

[18] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Seventh International Conference on Spoken Language Processing*, 2002.

[19] Y.-J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for hmm-based speech synthesis," in *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*. IEEE, 2008, pp. 1–4.