Speaker and Language Adaptive Training for HMM-Based Polyglot Speech Synthesis

Heiga Zen

Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, UK
heiga.zen@crl.toshiba.co.uk

Abstract

This paper proposes a technique for speaker and language adaptive training for HMM-based polyglot speech synthesis. Language-specific context-dependencies in the system are captured using CAT with cluster-dependent decision trees. Acoustic variations caused by speaker characteristics are handled by CMLLR-based transforms. This framework allows multi-speaker/multi-language adaptive training and synthesis to be performed. Experimental results show that the proposed technique achieves better synthesis performance than both speaker-adaptively trained language-dependent and language-independent models.

Index Terms: HMM-based speech synthesis, polyglot synthesis, adaptive training

1. Introduction

There have been several attempts to synthesize polyglot speech based on hidden Markov model (HMM)-based speech synthesis [1, 2]. Although the quality of HMM-based speech synthesis is still not as good as that of the best unit-selection synthesis, it has been improving in these years. Furthermore, HMM-based speech synthesis can provide sufficient flexibility to realize polyglot synthesis. The two main technical challenges posed by the implementation of HMM-based polyglot speech synthesis are how to combine data from multiple speakers in multiple languages into a single HMM-based speech synthesizer. For this challenge, Latorre *et. al* incorporated language-specific questions in addition to phonetic ones during tree-based clustering [1]. Furthermore, they applied phone mapping to transform the phone sets of adaptation data to that of training data, and then performed maximum likelihood linear regression (MLLR) with mapped transcriptions to adapt polyglot HMMs [1]. Although this approach was shown to be effective, the following problems still need to be fixed;

- All speech data from different languages and speakers are mixed to estimate models. Though good performance has been obtained, the acoustic variability between languages and speakers is not well addressed. It would be preferable to use other training schemes that are more powerful to handle the variability between different languages and speakers in the training data.
- Only a single decision tree per state is used to represent all languages. It is expected that each language has its own context-dependency, especially for prosody.

To address these problems, a technique for speaker and language adaptive training (SLAT) is proposed. Language-specific context-dependencies in the system are captured using cluster adaptive training (CAT) [3] with cluster-dependent decision trees [4]. Acoustic variations caused by speaker characteristics are handled by constrained maximum likelihood linear regression (CMLLR)-based transforms [5]. This framework allows multi-speaker/multi-language adaptive training and synthesis to be performed.

The rest of this paper is organized as follows. Section 2 describes the proposed technique. Section 3 shows the experimental results. Concluding remarks are presented in Section 4.

2. Speaker and language adaptive training

2.1. SLAT model

Figure 1 shows the block diagram of the SLAT model. The SLAT model is a combination of CMLLR-based speaker-adaptive training (CMLLR-SAT) [5] and CAT with cluster-dependent decision trees [4]; CMLLR-SAT is to normalize acoustic variations caused by speaker differences; CAT with cluster-dependent decision trees is to capture language-specific context dependency.

Cluster adaptive training has been used in speech recognition mainly for speaker adaptation [3]. It can be viewed as a "soft" version of speaker clustering. Unlike the traditional "hard" speaker clustering, CAT expresses mean vectors of a speaker-dependent model set as linear combinations of basis vectors which represent underlying "proto-type" speakers, while keeping covariance matrices and mixture weights unchanged across clusters and speakers. This paper extends this idea to represent languages; mean vectors of a language-dependent model set are represented as linear combinations of underlying "proto-type" languages. Because each language has its own context-dependency, the model should be able to represent context-dependency of proto-type languages. The use of cluster-dependent decision trees [4] enables this.

The left side of Fig. 1 illustrates the language-adaptation part of the SLAT model. There are cluster-dependent decision trees at the leftmost part of this figure. Cluster mean vectors, $\{\boldsymbol{\mu}_n\}$, are associated with the leaf nodes of these trees. A set of mean vectors in a language-adapted model set, $\{\boldsymbol{\mu}_m^{(l)}\}$, is generated by combining the P sets of cluster mean vectors with a set of language-dependent CAT interpolation weights, $\{\lambda_{i,r}^{(l)}\}$, as

$$\boldsymbol{\mu}_m^{(l)} = \sum_{i=1}^{P} \lambda_{i,r_c(m)}^{(l)} \boldsymbol{\mu}_{c(m,i)}, \tag{1}$$

where $m \in \{1, \ldots, M\}$, $l \in \{1, \ldots, L\}$, and $i \in \{1, \ldots, P\}$ are indexes for Gaussian component, language, and cluster (proto-type), respectively, and M, L, and P are the total number of Gaussian components, languages, and clusters, respectively. $r_c(m) \in \{1, \ldots, R_c\}$ denotes the CAT regression class, R_c is the total number of CAT regression classes, $c(m, i) \in$

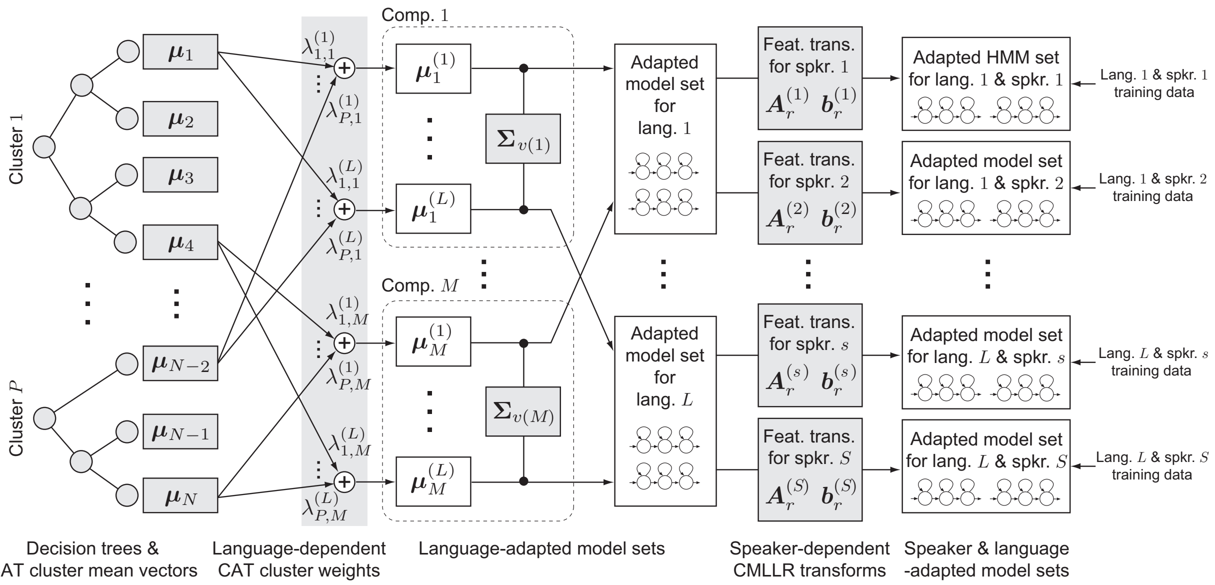
Cluster 1
Cluster P
$\boldsymbol{\mu}_1$
$\boldsymbol{\mu}_2$
$\boldsymbol{\mu}_3$
$\boldsymbol{\mu}_4$
$\boldsymbol{\mu}_{N-2}$
$\boldsymbol{\mu}_{N-1}$
$\boldsymbol{\mu}_N$
$\lambda_{1,1}^{(1)}$
$\lambda_{P,1}^{(1)}$
$\lambda_{1,1}^{(L)}$
$\lambda_{P,1}^{(L)}$
$\lambda_{1,M}^{(1)}$
$\lambda_{P,M}^{(1)}$
$\lambda_{1,M}^{(L)}$
$\lambda_{P,M}^{(L)}$
Comp. 1
$\boldsymbol{\mu}_1^{(1)}$
$\boldsymbol{\mu}_1^{(L)}$
$\boldsymbol{\Sigma}_{v(1)}$
Comp. M
$\boldsymbol{\mu}_M^{(1)}$
$\boldsymbol{\mu}_M^{(L)}$
$\boldsymbol{\Sigma}_{v(M)}$
Adapted model set for lang. 1
Adapted model set for lang. L
Feat. trans. for spkr. 1 $\boldsymbol{A}_r^{(1)}$ $\boldsymbol{b}_r^{(1)}$
Feat. trans. for spkr. 2 $\boldsymbol{A}_r^{(2)}$ $\boldsymbol{b}_r^{(2)}$
Feat. trans. for spkr. s $\boldsymbol{A}_r^{(s)}$ $\boldsymbol{b}_r^{(s)}$
Feat. trans. for spkr. S $\boldsymbol{A}_r^{(S)}$ $\boldsymbol{b}_r^{(S)}$
Adapted HMM set for lang. 1 & spkr. 1
Adapted model set for lang. 1 & spkr. 2
Adapted model set for lang. L & spkr. s
Adapted model set for lang. L & spkr. S
Lang. 1 & spkr. 1 training data
Lang. 1 & spkr. 2 training data
Lang. L & spkr. s training data
Lang. L & spkr. S training data
Decision trees & CAT cluster mean vectors
Language-dependent CAT cluster weights
Language-adapted model sets
Speaker-dependent CMLLR transforms
Speaker & language -adapted model sets

Figure 1: Block diagram of the proposed SLAT technique. Shaded blocks correspond to those to be updated.

$\{1,\ldots,N\}$ indicates the leaf node in decision trees for CAT cluster mean vectors which the i-th cluster mean vector at the component m belongs to, and N is the total number of leaf nodes in all decision trees for CAT cluster mean vectors. The generated set of mean vectors, together with a set of covariance matrices (one for an entire model set), $\{\boldsymbol{\Sigma}_k\}$, forms the language-dependent model set. Note that there are decision trees for covariance matrices as well.

The right half of Fig. 1 illustrates the speaker adaptation. In addition to the language adaptation by CAT, a set of speaker-dependent CMLLR feature-space transforms, $\{\boldsymbol{A}_r^{(s)}, \boldsymbol{b}_r^{(s)}\}$, is applied to generate the speaker- and language-adapted model set. These CMLLR feature transforms give

$$\hat{\boldsymbol{o}}_{r_a(m)}^{(s)}(t) = \boldsymbol{A}_{r_a(m)}^{(s)}\boldsymbol{o}(t) + \boldsymbol{b}_{r_a(m)}^{(s)}, \tag{2}$$

where $t \in \{1,\ldots,T\}$ and $s \in \{1,\ldots,S\}$ are indexes for time and speaker, respectively. $\boldsymbol{o}(t)$ is an observation vector at frame t, $r_a(m) \in \{1,\ldots,R_a\}$ is the CMLLR regression class, and R_a is the total number of CMLLR regression classes. Finally, a set of speaker- and language-adapted model sets are generated. Its state-output probability[1] can be expressed as

$$p(\boldsymbol{o}(t) \mid m, s, l, \mathcal{M}) = \left|\boldsymbol{A}_{r_a(m)}^{(s)}\right| \mathcal{N}\left(\hat{\boldsymbol{o}}_{r_a(m)}^{(s)}(t)\ ;\ \boldsymbol{\mu}_m^{(l)}, \boldsymbol{\Sigma}_{v(m)}\right), \tag{3}$$

where $\mathcal{M}$ is the set of model parameters. $v(m) \in \{1,\ldots,V\}$ denotes the leaf node in the decision trees which the covariance matrix of the component m belongs to and V is the total number of leaf nodes in the decision trees for covariance matrices.

The parameters of the SLAT model may be split into three distinct parts. The first part is the parameters of the canonical model $\{\boldsymbol{\mu}_n\}$ and $\{\boldsymbol{\Sigma}_k\}$. The second part is the parameters associated with the CMLLR linear transforms, $\{\boldsymbol{A}_r^{(s)}, \boldsymbol{b}_r^{(s)}\}$. The third one is the CAT interpolation weights, $\{\lambda_{i,r}^{(l)}\}$. This paper refers to the first one as canonical model parameters and the second/third ones as transform parameters.

2.2. Parameter estimation of SLAT model

The goal is to estimate the parameters of the SLAT model that maximize the likelihood given the training data with associated transcriptions and speaker/language labels. Like standard CAT and CMLLR-SAT, the EM algorithm is used. An iterative approach is used where first the transform parameters are estimated, then the canonical model parameters. The whole process is then repeated.

From Eq. (3), the auxiliary function of the EM algorithm for the SLAT model is given as

$$\begin{aligned}\mathcal{Q}\left(\mathcal{M}, \hat{\mathcal{M}}\right) \propto \sum_{m,t,s,l} \gamma_m(t) \Big\{ \log\left|\boldsymbol{\Sigma}_{v(m)}\right| - \log\left|\boldsymbol{A}_{r_a(m)}^{(s)}\right|^2 \\ + \left(\hat{\boldsymbol{o}}_{r_a(m)}^{(s)}(t) - \boldsymbol{\mu}_m^{(l)}\right)^\top \boldsymbol{\Sigma}_{v(m)}^{-1} \left(\hat{\boldsymbol{o}}_{r_a(m)}^{(s)}(t) - \boldsymbol{\mu}_m^{(l)}\right)\Big\},\end{aligned} \tag{4}$$

where $\gamma_m(t)$ is the posterior probability of component m generating the observation $\boldsymbol{o}(t)$ given the current model parameters $\hat{\mathcal{M}}$.

By taking the first partial derivative of Eq. (4) with respect to $\{\boldsymbol{\mu}_n\}$ and setting it to $\boldsymbol{0}$, a set of linear equations to determine the ML estimates of all cluster mean vectors are derived as

$$\begin{bmatrix} \boldsymbol{G}_{11} & \ldots & \boldsymbol{G}_{1N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{N1} & \ldots & \boldsymbol{G}_{NN} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}}_1 \\ \vdots \\ \hat{\boldsymbol{\mu}}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{k}_1 \\ \vdots \\ \boldsymbol{k}_N \end{bmatrix}, \tag{5}$$

where

$$\boldsymbol{G}_{n_1 n_2} = \sum_{\substack{m,t,l,i,j \\ c(m,i)=n_1 \\ c(m,j)=n_2}} \gamma_m(t) \lambda_{i,r_c(m)}^{(l)} \boldsymbol{\Sigma}_{v(m)}^{-1} \lambda_{j,r_c(m)}^{(l)}, \tag{6}$$

$$\boldsymbol{k}_{n_1} = \sum_{\substack{m,t,s,l,i \\ c(m,i)=n_1}} \gamma_m(t) \lambda_{i,r_c(m)}^{(l)} \boldsymbol{\Sigma}_{v(m)}^{-1} \hat{\boldsymbol{o}}_{r_a(m)}^{(s)}(t). \tag{7}$$

Although the dimensionality of Eq. (5) can be hundreds of thousands, it's sparse. Therefore, it can be stored and solved efficiently using a sparse matrix storage and solver. The update

[1] Its state-duration probabilities are defined in the same manner.

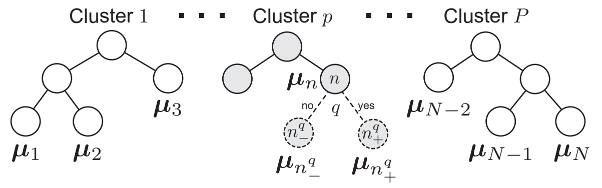

Figure 2: Overview of tree-reconstruction process. The n-th node is associated with the decision tree for cluster p. Shaded parts correspond to parameters and a tree to be updated.

formula of covariance matrices is straightforward thus it is omitted in this paper.

Reestimation of transform parameters is a simple iterative process, where given the language-dependent CAT interpolation weights, the adapted mean vectors, $\{\boldsymbol{\mu}_m^{(l)}\}$, are used to estimate the speaker-dependent CMLLR transforms, $\{\boldsymbol{A}_r^{(s)}, \boldsymbol{b}_r^{(s)}\}$, as described in [3]. Then the language-dependent CAT interpolation weights, $\{\lambda_{i,r}^{(l)}\}$, are estimated using the transformed feature vectors, $\{\hat{\boldsymbol{o}}_r^{(s)}(t)\}$, as described in [5].

2.3. Tree reconstruction

The conventional cluster-based methods assume that all clusters have the same parameter tying structure, *i.e.*, decision trees. However, there is no explicit restriction in these methods; each cluster can have its own parameter tying structure. Recently, cluster-based methods with cluster-dependent decision trees have been proposed [4, 6]. These techniques enable us to construct different decision trees for each cluster. The proposed technique uses this framework to build cluster-dependent decision trees to capture language-specific context dependency.

Because building multiple trees simultaneously [4] is computationally expensive, an iterative approach over trees [6] is used. While reconstructing a decision tree for a cluster the structure of other trees and values of parameters including cluster mean vectors and covariance matrices associated with leaf nodes of these other trees are fixed. The goal is to build decision trees and estimate associated parameters that maximize the likelihood given the training data, while maintaining the balance between model complexity and accuracy.

As illustrated in Fig. 2, the n-th terminal node associated with the decision tree for cluster p is divided into two new terminal nodes, n_+^q and n_-^q, by a question q. The log likelihood gain by this split is given by

$$\delta\mathcal{L}(n,q) = \frac{1}{2}\left(\boldsymbol{r}_{n_+^q}^\top \boldsymbol{R}_{n_+^q} \boldsymbol{r}_{n_+^q} + \boldsymbol{r}_{n_-^q}^\top \boldsymbol{R}_{n_-^q} \boldsymbol{r}_{n_-^q} - \boldsymbol{r}_n^\top \boldsymbol{R}_n \boldsymbol{r}_n\right), \tag{8}$$

where

$$\boldsymbol{R}_n = \boldsymbol{G}_{nn}^{-1}, \tag{9}$$

$$\boldsymbol{r}_n = \boldsymbol{k}_n - \sum_{\substack{m,t,l,i \\ c(m,p)=n \\ i\neq p}} \gamma_m(t) \lambda_{p,r_{c(m)}}^{(l)} \boldsymbol{\Sigma}_{v(m)}^{-1} \lambda_{i,r_{c(m)}}^{(l)} \boldsymbol{\mu}_{c(m,i)}. \tag{10}$$

According to the gain in log likelihood, the best question $\hat{q}$ can be chosen. Splitting can be stopped automatically based on an information criterion such as the minimum description length (MDL) criterion.

2.4. Initialization

There are several possible ways to initialize the parameters of the SLAT model. Here they are initialized using those of speaker-adaptively trained language-independent model sets.

First a language-independent model set is estimated using CMLLR-SAT with a single transform per each decision tree of each speaker. Then, a SLAT model set is initialized using this language-independent model set as follows. The number of clusters P in the SLAT model is set to $L+1$. The decision trees for cluster 1 and their associated cluster mean vectors are initialized by those of the language-independent model set. The covariance matrices, space weights for multi-space probability distributions (MSD) [7], and their parameter sharing structure are also initialized by those of language-independent model sets. A specific language tag is assigned to each of $2,\ldots,P$ clusters, *e.g.*, clusters 2, 3, and 4 are for German, English, and French, respectively. The decision trees for clusters $2,\ldots,P$ are initialized to have only root nodes, and the cluster mean vectors associated with these root nodes are initialized as $\mathbf{0}$. To initialize CAT cluster weights, they are simply set to 1 or 0 according to their assigned language tags as

$$\lambda_{i,r}^{(l)} = \begin{cases} 1 & i=1 \text{ or its language tag is } l \\ 0 & \text{Otherwise} \end{cases}.$$

The CMLLR transforms in the language-independent model set are used to initialize those of the SLAT model. As a result, a SLAT model set which gives exactly the same likelihood to the training data as the speaker-adaptively trained language-independent model set is obtained.

3. Experiments

To evaluate the performance of the proposed technique, a preliminary experiment was conducted. A newly recorded multilingual database was used in this experiment. It consisted of five languages; North American and British English, European Spanish, European French, and Standard German. There were 10 speakers (five male and five female) in each language, each speaker uttered 50 phonetically balanced sentences (common across speakers in each language) and 50–100 sentences which were selected from various domains (different across speakers). To avoid the effect of recording condition variations, the same microphone and recording studio were used while recording speech from all speakers. The training data consisted of 6,100 utterances, and 250 sentences excluded from the training data were used as test data. Please refer to [8] for details.

The speech analysis conditions and model topologies used in this experiment were similar to those of HTS 2008 [9]. Note that phoneme segmentation, linguistic labels (*e.g.*, syllabification, lexical stress), and $\log F_0$ values were automatically extracted and not manually corrected. A universal phoneset and context-dependent label format, which can cover phonemes and contexts in the training languages, were defined and used. A language-independent, speaker-adaptively trained (LI-SAT) model was first estimated to initialize the SLAT model.

After initializing the SLAT model in the way shown in Section 2.4, its parameters and decision trees were iteratively updated. The CAT interpolation weights for cluster 1 were fixed to 1.0 (bias cluster [3]) during the training to make cluster 1 represent the common factors across languages. The parameter sharing structure of the covariance matrices and MSD weights was assumed to be the same as that of cluster 1. Simple two class (silence and speech) base classes were used as CMLLR and CAT

Table 1: Numbers of leaf nodes for mel-cepstral coefficients, $\log F_0$, band aperiodicity, and state durations in the LD-SAT (German, UK and US English, Spanish, and French) and LI-SAT models.

Language	mel-cep.	$\log F_0$	band ap.	dur.
LI-SAT	4,359	31,201	2,244	2,259
German	1,330	8,446	740	460
UK English	1,179	8,635	683	422
US English	1,182	9,003	629	374
Spanish	1,057	5,567	512	296
French	1,147	6,196	641	346
Total	5,895	37,847	3,205	1,898

Table 2: Numbers of cluster mean vectors for mel-cepstral coefficients, $\log F_0$, band aperiodicity, and state durations in the SLAT model.

Cluster	mel-cep.	$\log F_0$	band ap.	dur.
1	4,537	12,894	1,866	1,724
2	165	1,954	306	65
3	244	1,970	173	59
4	200	1,940	226	127
5	208	1,119	227	52
6	161	1,421	261	94
Total	5,515	21,298	3,059	2,121

regression classes. Five iterations of SLAT training were run. No language-specific questions were incorporated during tree-based clustering. To improve the numerical stability and relax overfitting, L_2 regularization was performed while solving Eq. (5) and computing Eq. (8). A set of language-dependent, speaker-adaptively trained (LD-SAT) models were also trained using the same dataset to compare the quality of them against that of the SLAT model. Table 1 shows the numbers of leaf nodes for spectrum (mel-cepstral coefficients), $\log F_0$, excitation (band aperiodicity), and state durations in the LD-SAT and LI-SAT models. Table 2 shows those of the SLAT models. It can be seen from the tables that the total sizes of these models are comparable. It can also be seen from the tables that cluster 1 was dominant for mel-cepstral coefficients, band aperiodicity, and state durations even after the SLAT training. However, that for $\log F_0$ significantly reduced after the SLAT training, and clusters $2, \ldots, P$ for $\log F_0$ covered a relatively larger portion than those for other speech parameters. They suggest that common factors across languages were dominant for mel-cepstral coefficients, band aperiodicity, and state durations but they had smaller effect for $\log F_0$.

After training the models, speech parameters for the test sentences were generated from the models using the speech parameter generation algorithm considering global variance. From the generated speech parameters, speech waveforms were synthesized using the source-filter model.

A paired-comparison preference listening test was conducted. This test compared LI-SAT, LD-SAT, and SLAT models over 250 sentences from the evaluation set. Fourteen subjects participated in the test. All subjects evaluated their native or near-native languages only (three of them evaluated two languages and one of them evaluated three languages). For each subject, 15 sentences were randomly chosen from the evaluation sentences in the language which the subject selected. Orders of pairs and samples were also randomized. Before starting the test, the subjects listened to speech samples of one sentence to become familiar with the task. This sentence was randomly chosen for each subject and excluded from the actual test. After listening to each test sample, the subjects were asked to choose their preferred one. Note that the subjects could select "No preference" if they had no preference.

Table 3: Preference scores (%) between LD-SAT and SLAT, LI-SAT and SLAT, and LI-SAT and LD-SAT.

LI-SAT	LD-SAT	SLAT	No preference
–	26.7	**45.6**	27.7
36.2	–	**37.6**	38.2
33.3	36.2	–	30.5

Table 3 shows the preference test result. It can be seen from the table that the SLAT model achieved the best preference score among the three systems. The differences between LI-SAT/LD-SAT and SLAT were statistically significant.

4. Conclusion

This paper proposed the technique for speaker and language adaptive training for HMM-based polyglot speech synthesis. Language-specific context-dependencies in the system are captured using CAT with cluster-dependent decision trees. Acoustic variations caused by speaker characteristics are handled by CMLLR-based transforms. This framework allows multi-speaker/multi-language adaptive training and synthesis to be performed. Experimental results showed that the proposed technique achieves better synthesis performance than both speaker-adaptively trained language-dependent and language-independent models.

5. References

[1] J. Latorre, *et al.*, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Commun.*, vol. 48, no. 10, pp. 1227–1242, 2006.

[2] A. Black and T. Schultz, "Speaker clustering for mulitilingual synthesis," in *Proc. ISCA ITRW MULTILING*, no. 024, 2006.

[3] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 417–428, 2000.

[4] H. Zen and N. Braunschweiler, "Context-dependent additive $\log F_0$ model for HMM-based speech synthesis," in *Proc. of Interspeech*, 2009, pp. 2091–2094.

[5] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[6] K. Saino, "A clustering technique for factor analyzed voice models," Master thesis, Nagoya Institute of Technology, 2008, (in Japanese).

[7] K. Tokuda, *et al.*, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[8] H. Zen, *et al.*, "HMM-based polyglot speech synthesis by speaker and language adaptive training," in *Proc. ISCA SSW7*, 2010, (submitted).

[9] J. Yamagishi, *et al.*, "The HTS2007' system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge Workshop*, 2008.