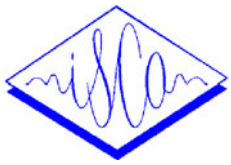

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

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

Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Kate Knill, Sacha Krstulović, Javier Latorre

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

Abstract

This paper describes a technique for speaker and language adaptive training (SLAT) for HMM-based polyglot speech synthesis and its evaluations on a multi-lingual speech corpus. The SLAT technique allows multi-speaker/multi-language adaptive training and synthesis to be performed. Experimental results show that the SLAT technique achieves better naturalness than both speaker-adaptively trained language-dependent (LD-SAT) and language-independent (LI-SAT) models. In cross-lingual adaptation speaker similarity tests SLAT and LI-SAT outperform LD-SAT but there are still significant differences between polyglot adaptation and intra-language adaptation.

1. Introduction

In the last few years there have been several approaches to synthesize speech in multiple languages [1–8]. With respect to multilingualism, most of them have adopted the concept given in [1], according to which a system is multilingual if it uses a set of common algorithms for all the languages and stores the language-specific information in separated data tables. For such multilingual synthesizers, it is irrelevant whether all the languages are synthesized by the same voice or a different voice is used for each language. However, for many applications, to have a common voice characteristic for all the languages is much more important than to share the same synthesis algorithms. For example, when synthesizing text that contains language-switching, changing from one voice to another every time the language of the input text changes can be extremely confusing. Other examples of such applications are portable devices such as a car-navigation system or an e-mail reader integrated in a cell-phone that have to be multilingual while keeping a small memory footprint.

The main purpose of this research is the implementation of a system that can speak multiple languages with multiple speakers' voice characteristics. To distinguish such a system from the traditional multilingual one described in [1], the term 'polyglot synthesizer' [8, 9] is used. Here, it becomes possible to use the voice of someone who only speaks English to synthesize speech in French, Spanish, German, or any other language.

Recently, statistical parametric polyglot synthesis techniques were proposed [8, 10]. These techniques tried to solve the main drawbacks of previous approaches based on a polyglot speaker [9] and phone mapping [6], while being easily adaptable in order to permit the imitation of the voice characteristics of any speaker. The main hypothesis of these techniques is that the average vocal tract characteristic over a sufficient number of speakers is similar for any language. According to this hypothesis, it should be possible to create an artificial polyglot speaker either by combining the average voices of all the languages under consideration, or by mixing the speech data of multiple speakers of those languages into a single polyglot average voice. Using such an artificial polyglot speaker, it is possible to solve the main problems of the two previous approaches to polyglot synthesis:

- Since no real polyglot speaker is required, the system can be expanded to any new language by just making an average voice for that new language, or by including speech data from some speakers of that new language into the statistical polyglot speaker.
- Since for none of the languages "spoken" by the statistical polyglot speaker a phone mapping is required, the level of foreign accent in the synthesis of these languages is considerably reduced. Hence, much better intelligibility can be obtained.

Obviously, a statistical speaker can only be created by a statistical parametric speech synthesis method [11]. For this purpose, HMM-based speech synthesis (HTS) [12] can be used. Although the naturalness of synthesized speech from HMM-based speech synthesis is still not as good as that of the best unit-selection synthesis, it has been improving in recent years. Furthermore, it provides sufficient flexibility to realize polyglot synthesis. The two main technical challenges posed by the implementation of HMM-based polyglot speech synthesis approach are:

1. How to combine data from multiple speakers in different languages into a single HMM-based speech synthesizer in such a way that despite using a limited number of speakers per language, the system speaks with the same voice for all the languages.
2. How to get the output voice sound like the target speaker, even when the language spoken by the target speaker and the language being synthesized are different.

For the first challenge, Latorre *et al.* incorporated language-specific questions in addition to phonetic ones into the phonetic decision tree clustering [8].[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 [8]. Although this approach worked effectively, the following problems still need to be addressed:

- All speech data from different languages and speakers is mixed to estimate models. Although good performance has been obtained, the acoustic variability between languages and speakers is not well addressed. It would be

[1] That system used phonetic contexts and generated only spectral sequences from HMMs, unlike the full HMM-based system [12]. It used external modules to predict prosody.

preferable to use other training schemes that are more powerful to handle the variability between different languages and speakers in the training data. Adaptive training is a powerful solution for building systems on non-homogeneous training data [13]. Rather than dealing with all the data as a single block, the training data is split into several homogeneous blocks, for example speaker or language. The effectiveness of adaptive training in HMM-based speech synthesis has been demonstrated [14].

- Only a single phonetic decision tree per state is used to represent all languages. It is expected that each language has its own context-dependency, especially for prosody. Latorre *et al*. used only phonetic contexts and generated only spectral parameters from HMMs, unlike full HMM-based systems [12, 15]. In this case, using a single phonetic decision tree per state is acceptable. However, if a full HMM-based polyglot speech synthesis system is required, one single phonetic decision tree per state is not enough to capture language variability.

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) [16] with cluster-dependent decision trees [17]. Acoustic variations caused by speaker characteristics are handled by constrained MLLR (CMLLR)-based transforms [18]. This framework allows multi-speaker/multi-language adaptive training and synthesis. By adapting both language-dependent CAT interpolation weights and speaker-dependent CMLLR transforms, it should be possible to construct an adapted model set for the target speaker and target language rapidly, using only a small amount of adaptation data. Furthermore, by using a target speaker's CMLLR transforms with the pre-estimated language-dependent CAT interpolation weights, new model sets for these languages may be obtained with the target speaker's voice characteristic.

The rest of this paper is organized as follows. Section 2 describes the SLAT technique. Section 3 shows the experimental results. Finally, concluding remarks and future plans are presented in Section 4.

2. Speaker & Language Adaptive Training

2.1. SLAT model

Figure 1 shows the block diagram of the SLAT model. The SLAT model uses the structured transform framework [19] to combine CMLLR-based speaker-adaptive training (CMLLR-SAT) [18] and CAT [16] with cluster-dependent decision trees [17]. Acoustic variations caused by speaker characteristics are handled by CMLLR-based transforms. Language-specific context-dependencies in the system are captured using CAT with cluster-dependent decision trees. Note that the SLAT model has multiple clusters each of which has a different decision tree-based parameter tying structure, in contrast to the standard structured transform setup.

Cluster adaptive training has been used in speech recognition mainly for speaker adaptation [16]. It can be viewed as a "soft" version of speaker clustering. Unlike the traditional "hard" version of speaker clustering, CAT expresses the mean vectors of a speaker-dependent model set as linear combinations of basis vectors which represent the underlying prototype 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 prototype languages. Because each language has its own context-dependency, it is expected that these prototype languages also have their own context-dependencies. The use of cluster-dependent decision trees [17] enables us to capture the context-dependencies of prototype languages.

The left side of Fig. 1 illustrates the language-adaptation part of the SLAT model. The cluster-dependent decision trees are located 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, $\{\lambda_{i,q}^{(l)}\}$, as

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

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 (prototype), respectively, and M, L, and P are the total number of Gaussian components, languages, and clusters, respectively. $q(m) \in \{1, \ldots, Q\}$ denotes the CAT regression class,[2] Q is the total number of CAT regression classes, $c(m, i) \in \{1, \ldots, N\}$ indicates the leaf node in decision trees for the 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 the decision trees for the 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 in the SLAT model but not shown in the figure.

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

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

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, $r(m) \in \{1, \ldots, R\}$ is the CMLLR regression class, and R is the total number of CMLLR regression classes. Finally, a set of speaker- and language-adapted model sets are generated. The emission probability of the observation $\boldsymbol{o}(t)$, which is uttered by speaker s in language l, from component m can be expressed as

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

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 state-duration probabilities in the SLAT model can be expressed in the same manner.

[2] The CAT and CMLLR regression classes define the parameter sharing structure of CAT and CMLLR transforms.

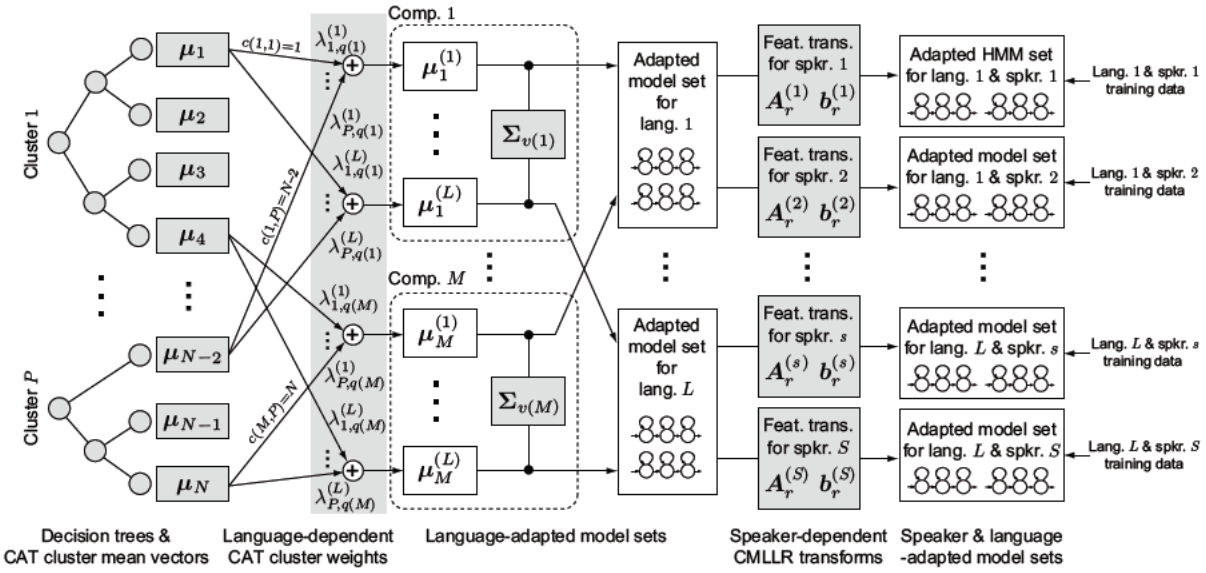

Figure 1: Block diagram of the proposed SLAT technique. Shaded blocks correspond to those to be updated.

The parameters of the SLAT model can 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 transforms, $\{\boldsymbol{A}_r^{(s)}, \boldsymbol{b}_r^{(s)}\}$. The third part 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. Due to the limitation of space, details of training process of the SLAT model are omitted. Please refer to [20] for details.

2.2. Adaptation

The SLAT adaptation process uses two sub-steps estimating speaker-dependent CMLLR transforms and language-dependent CAT interpolation weights. The final CMLLR transforms and the CAT interpolation weights are used to construct the adapted model for synthesis.

2.2.1. Adapting to a new speaker

To adapt the model to a target speaker who can speak a language included in the training data, first a language-adapted model set is composed using the pre-estimated language-dependent CAT interpolation weights. Then, speaker-dependent CMLLR transforms are estimated as described in [18]. The estimation process is illustrated in Fig. 2. Using the estimated speaker-dependent CMLLR transforms with the pre-estimated language-dependent CAT interpolation weights, any language included in the training data can be synthesized with the target speaker's voice characteristic.

2.2.2. Adapting to a new language

To estimate the language-dependent CAT interpolation weights of a new language, a set of target language adaptation data uttered by multiple speakers is required. This is because it is difficult to separate language and speaker variations if there is only speech data from a single speaker. The estimation process is illustrated in Fig. 3. For each speaker, first CMLLR transforms are estimated to normalize the speaker variations. Using the normalized features, the language-dependent CAT interpolation weights for the target language are estimated. Estimations of the CMLLR transforms and the CAT interpolation weights are interleaved until they converge.

3. Experiments

3.1. Data preparation

There are a couple of multilingual speech databases available, such as GlobalPhone [21]. However, none of them were designed for the speech synthesis purpose. Therefore, a new database was recorded for the SLAT experiment.[3] The database consists of five languages; North American English, British English, European Spanish, European French, and Standard German. There were 10 non-professional speakers (five male and five female) in each language. To cover various ages, these speakers were selected from four age ranges (1 from 13–18, 2 from 20–30, 1 from 30–50, and 1 from 50–70) for each gender. Speakers did not have strong regional accents and spoke close to the standard of each of the languages selected.[4] Each speaker uttered the same 50 phonetically rich sentences which covered all phones in the language and another set of 50 or more sentences which were selected from various domains (and differed between speakers). The total recording duration for each speaker was between eight and fifteen minutes. A headset microphone was used to record the voices and EGG recordings were also made simultaneously. All recordings were in a normal recording room with low reverberation without any background noise. To avoid the effect of recording condition variations, the same microphone and recording room were used while recording speech from all speakers. The sampling frequency was 48 kHz, later downsampled to 16 kHz. These recordings are used in the experiment reported here.

A universal phone set, which covers all training languages,

[3]The authors are planning to increase the size of training data for SLAT using GlobalPhone corpus in the future.

[4]Within a language, the speakers varied somewhat in accent but their speech did not have strong regional accent.

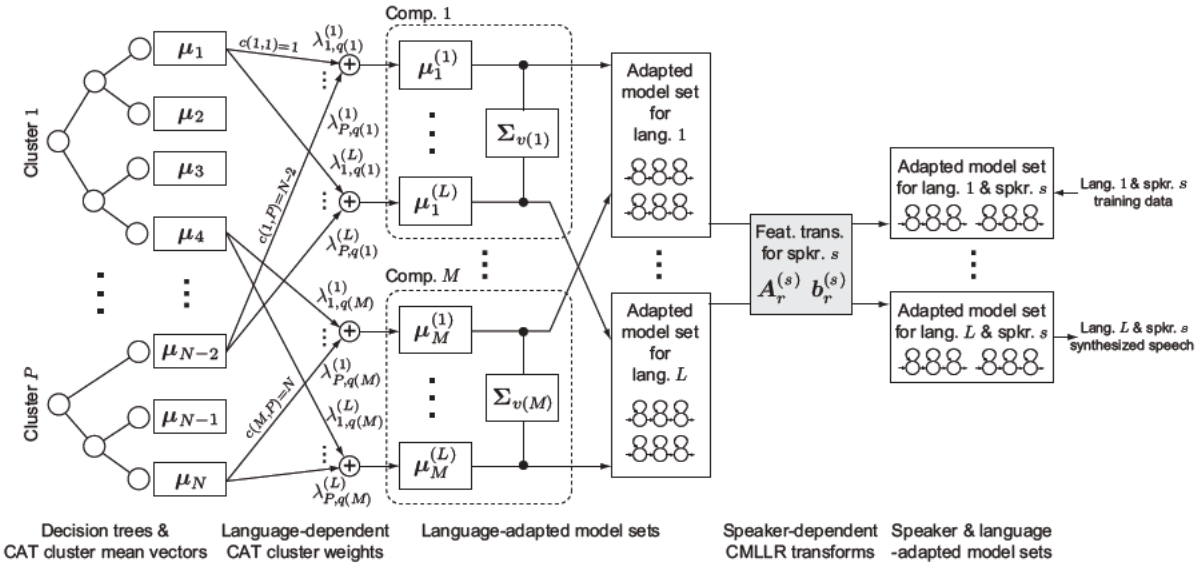

Figure 2: Block diagram of speaker adaptation and polyglot synthesis. Shaded blocks correspond to those to be updated.

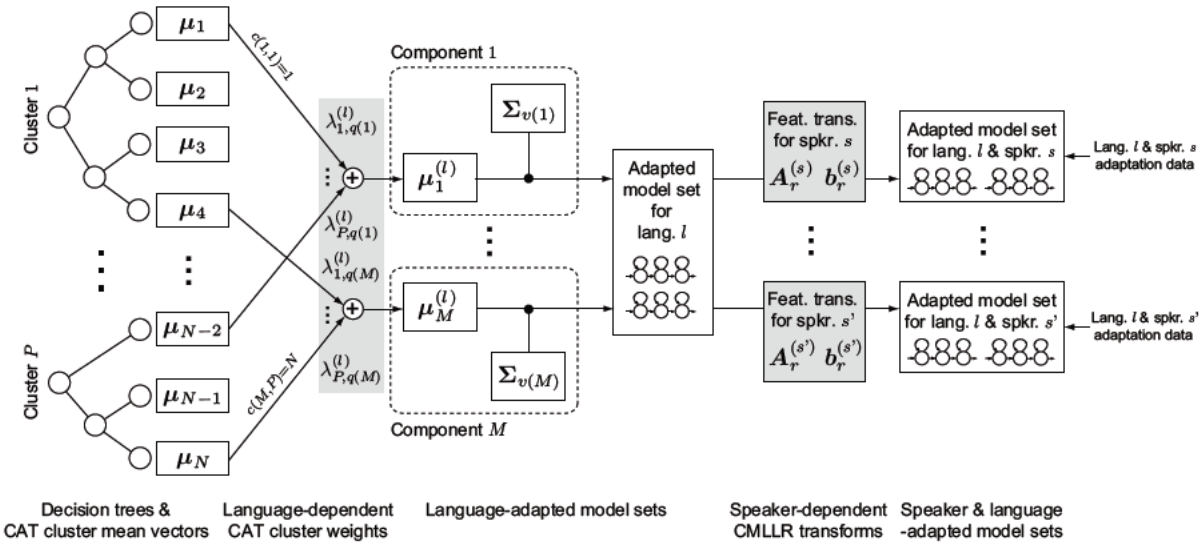

Figure 3: Block diagram of language adaptation. Shaded blocks correspond to those to be updated.

was defined and used. It defines a set of phone symbols in alphabetical characters that map to unique speech units across all training languages. Each phone symbol in this phone set has an equivalent transcription in the IPA alphabet [22]. This ensures that the phone set is well grounded in phonetic sciences. The recording scripts were automatically converted into the corresponding phone sequences using Toshiba's proprietary text analysis engine, and then Toshiba's proprietary HMM-based automatic aligner was used to extract the phone segmentations. Only a few severe text normalization problems were manually corrected and no manual correction was performed to the extracted segmentations. A universal context-dependent label format, which can cover possible contexts in the training languages, was also defined. The contexts used in this format are similar to those in [23]: they include phonetic, prosodic, and grammatical contexts. The fundamental frequency (F_0) values of the recordings were automatically extracted from the EGG recordings using the voting method [24] with proprietary and publicly available F_0 extraction techniques. No manual correction of the extracted F_0 values was performed.

3.2. Experimental setup

To evaluate the performance of the proposed technique, a preliminary experiment was conducted. The speech analysis conditions and model topologies used in this experiment were similar to those of HTS 2008 [24]. A language-independent, speaker-adaptively trained (LI-SAT) model was first estimated to initialize the SLAT model.

After initializing the SLAT model by the LI-SAT model, its parameters and decision trees were iteratively updated. The CAT interpolation weights for cluster 1 were fixed to 1.0 (bias cluster [16]) during the training to make cluster 1 represent the common factors across languages. The parameter sharing struc-

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

ture of the covariance matrices and the MSD weights were assumed to be the same as that of cluster 1. Simple two class (silence and speech) base classes were used as CMLLR and CAT regression classes. Five iterations of SLAT training were run. One iteration of SLAT training consisted of

1. Rebuild decision trees;
2. Update canonical model parameters;
3. Update transform parameters.

To improve the numerical stability and relax overfitting, L_2 regularization was performed while training the SLAT model. The MDL criterion was used to control the size of decision trees. 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. The LI-SAT and SLAT models were trained using data from all languages, while the LD-SAT models were trained using data from individual languages. After training the models, speech parameters for the test sentences were generated from the models using the speech parameter generation algorithm considering global variance [25]. From the generated speech parameters, speech waveforms were synthesized using the source-filter model.

3.3. Experimental results

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 were 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. This suggests that common factors across languages were dominant for mel-cepstral coefficients, band aperiodicity, and state durations but that they had a smaller effect for $\log F_0$.

A paired-comparison preference listening test was conducted. This test compared the naturalness of synthesized speech generated from LD-SAT, LI-SAT, and SLAT models over 250 sentences excluded from the training data. 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

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

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
33.3	36.2	–	30.5
24.2	–	**37.6**	38.2
–	26.7	**45.6**	27.7

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 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 at $p < 0.01$ level by the two-tailed t-test.

A differential mean opinion score (DMOS) test was conducted to evaluate the speaker similarities in polyglot synthesis. The target speaker was a German male speaker and the target language was US English. Ten subjects participated in the test. For each subject, fixed five sentences from the US English evaluation sentences were used. Seven pairs of speech samples were presented for each sentence. The first sample in each pair was a reference natural speech utterance from the database and the second one was a speech sample to be evaluated. The speech samples to be evaluated were generated from

1. US English LD-SAT model without adaptation (`AVM`);
2. US English LD-SAT model adapted with CMLLR transforms from a training speaker who sounded subjectively most similar to the target speaker (`TRAIN`);
3. US English LD-SAT model adapted with CMLLR transforms for the target speaker estimated by the data mapping-based cross-lingual adaptation technique [26] (`CROSS`);
4. LI-SAT model adapted with CMLLR transforms for the target speaker (`LI-SAT`);
5. SLAT model adapted with CMLLR transforms for the target speaker and CAT interpolation weights for the target language (`SLAT`);
6. German LD-SAT model adapted with CMLLR transforms for the target speaker (`INTRA`);

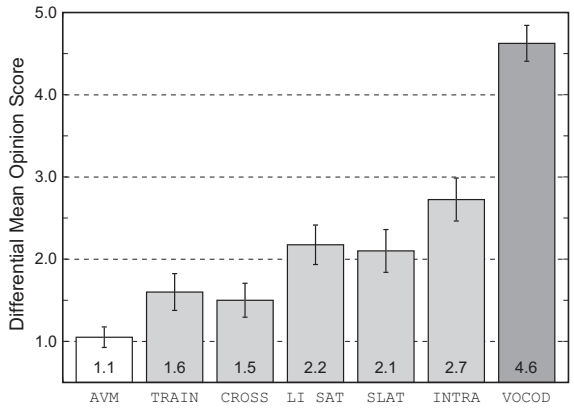

Figure 4: Differential MOS test results of synthesized speech from the adapted models with different techniques. Error bars show 95% confidence intervals.

7. Vocoded natural speech (VOCOD).

After listening to each pair, subjects were asked to give a five-scale opinion score for the second sample relative to the first one, expressing how similar the speaker identity was (1: very dissimilar – 5: very similar). Note that the reference utterances, vocoded natural speech, and synthetic speech generated from the German LD-SAT model were in German but other samples were in US English. To reduce the effect of language differences as much as possible, the subjects were asked to ignore other factors (*e.g.*, intelligibility or naturalness) while rating the speaker similarity.

Figure 4 shows the experimental results. It can be seen from the table that all the adaptation techniques achieved better similarity scores than the AVM. Although LI-SAT and SLAT achieved better similarity scores than CROSS, there are still significant differences between polyglot synthesis and intra-language adaptation INTRA. Furthermore, the gap between INTRA and VOCOD indicates that the statistical modeling process caused the largest degradation in speaker similarity.

4. Conclusion

This paper proposed the technique of speaker and language adaptive training for HMM-based polyglot speech synthesis. While training the canonical model, language-specific context dependency is captured through CAT with cluster-dependent decision trees, and acoustic variations caused by speaker characteristics are normalized by CMLLR-SAT. Experimental results showed that the proposed approach achieved better synthesis than both speaker-adaptively trained language-dependent and language-independent models.

Future work includes evaluation of language adaptation, increasing the amount of training data per language, and adding more languages from different language families.

5. Acknowledgments

The authors would like to thank Dr. M. J. F. Gales for helpful comments and discussions. The authors are also grateful to Dr. Art Blokland for helping data preparation.

6. References

[1] R. Sproat, Ed., *Multilingual text-to-speech synthesis: The Bell labs approach*. Kluwer Academic Publisher, 1998.

[2] S. Quazza, *et al*., "Actor: A multilingual unit-selection speech synthesis system," in *Proc. ISCA SSW4*, 2001.

[3] A. Black and K. Lenzo, "Multilingual text-to-speech synthesis," in *Proc. ICASSP*, vol. 3, 2004, pp. 761–764.

[4] M. Chu, *et al*., "Microsoft Mulan – A bilingual TTS system," in *Proc. Interspeech*, vol. 1, 2003, pp. 264–267.

[5] F. Deprez, *et al*., "Introduction to multilingual corpus-based concatenative speech synthesis," in *Proc. Interspeech*, 2007, pp. 2129–2132.

[6] N. Campbell, "Talking foreign – Concatenative speech synthesis and the language barrier," in *Proc. Eurospeech*, 2001, pp. 337–340.

[7] H. Liang, *et al*., "A cross-language state mapping approach to bilingual (Mandarin-English) TTS," in *Proc. ICASSP*, 2008, pp. 4641–4644.

[8] 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.

[9] C. Traber, *et al*., "From multilingual to polyglot speech synthesis," in *Proc. Eurospeech*, 1999, pp. 835–838.

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

[11] H. Zen, *et al*., "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[12] T. Yoshimura, *et al*., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[13] T. Anastasakos, *et al*., "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[14] J. Yamagishi, "Average-voice-based speech synthesis," Ph.D. dissertation, Tokyo Institute of Technology, 2006.

[15] K. Tokuda, *et al*., "The HMM-based speech synthesis software toolkit," http://hts.sp.nitech.ac.jp/.

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

[17] 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.

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

[19] K. Yu and M. Gales, "Adaptive training using structured transforms," in *Proc. ICASSP,*, 2004, pp. 317–320.

[20] H. Zen, "Speaker and language adaptive training for HMM-based polyglot speech synthesis," in *Interspeech*, 2010, to appear.

[21] T. Schultz, "Globalphone: a multilingual speech and text database developed at Karlsruhe University," in *Proc. ICSLP*, 2002, pp. 345–348.

[22] International Phonetic Association, *Handbook of the international phonetic association*, Cambridge University Press, 1999.

[23] K. Tokuda, *et al*., "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002, CD-ROM Proceeding.

[24] 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.

[25] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst*., vol. E90-D, no. 5, pp. 816–824, 2007.

[26] Y.-J. Wu, *et al*., "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 528–531.