STRUCTURAL KLD FOR CROSS-VARIETY SPEAKER ADAPTATION IN HMM-BASED SPEECH SYNTHESIS

Markus E. Toman
Telecommunications Research Center Vienna (FTW)
Vienna, Austria
email: toman@ftw.at

Michael Pucher
Telecommunications Research Center Vienna (FTW)
Vienna, Austria
email: pucher@ftw.at

ABSTRACT

While the synthesis of natural sounding, neutral style speech can be achieved using today's technology, fast adaptation of speech synthesis to different contexts and situations still poses a challenge. In the context of variety modeling (dialects, sociolects) we have to cope with the problem that no standardized orthographic form is available and that existing speech resources for these varieties are rare. We present recent approaches in the field of cross-lingual speaker transformation for HMM-based speech synthesis and propose a method for transforming an arbitrary speaker's voice from one variety to another one. We apply Kullback-Leibler divergence for data mapping of HMM-states, transfer probability density functions to the decision tree of the other variety and perform speaker adaptation. A method to integrate structural information in the mapping is also presented and analyzed. Subjective listening tests show that the proposed method produces speech of significantly higher quality than standard speaker adaptation techniques.

KEY WORDS

speech processing, algorithms and techniques, speech synthesis, speaker adaptation, variety modeling

1 Introduction

Speech synthesis of language varieties is a basis for many application scenarios where a natural and realistic persona design is necessary. We have been investigating this topic in the last years [1]. Here we consider the problem of variety transformation, where we transform speech data of speaker A in variety V_1 to a model of speaker A in variety V_2, without having V_2 data from speaker A. For example, Standard Austrian German speech data of a speaker could be used to build a Viennese dialect voice model for the same speaker.

This method can be applied in language learning scenarios, where a user can listen to his / her own voice in a variety that he / she wants to learn. The modeling techniques developed here can further be applied to accented speech. Our modeling technique exploits the fact that there is a significant overlap between the modeled varieties V_1 and V_2.

In previous and current projects we recorded and annotated phonetically balanced speech data in different Austrian varieties from Vienna (VD) [2], Innervillgraten - Eastern Tyrol (IVG), Bad Goisern - Upper Austria (BG) [3] and Standard Austrian German (AT). In our current work we are focusing on the Eastern Tyrol dialect from Innervillgraten (IVG). In this paper we evaluate the transformation from Standard Austrian German (AT) to Innervillgraten dialect (IVG).

Related work in the field of language transformation [4, 5, 6] will be described in Section 3.

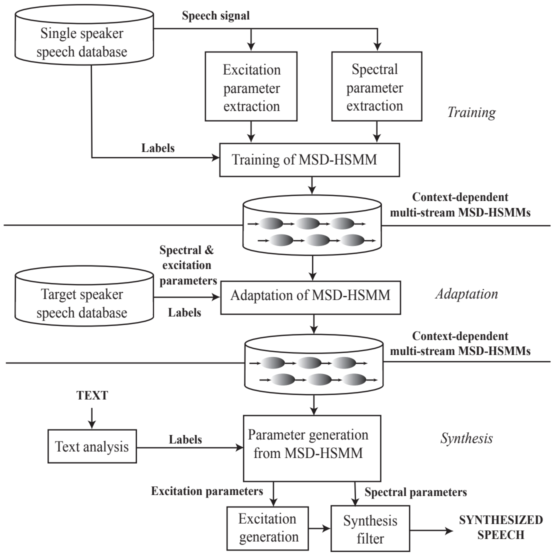
Single speaker speech database
Speech signal
Excitation parameter extraction
Spectral parameter extraction
Training
Labels
Training of MSD-HSMM
Context-dependent multi-stream MSD-HSMMs
Spectral & excitation parameters
Target speaker speech database
Labels
Adaptation of MSD-HSMM
Adaptation
Context-dependent multi-stream MSD-HSMMs
TEXT
Text analysis
Labels
Parameter generation from MSD-HSMM
Synthesis
Excitation parameters
Spectral parameters
Excitation generation
Synthesis filter
SYNTHESIZED SPEECH

Figure 1. HMM-based speech synthesis system (HTS).

2 Speaker-adaptive acoustic modeling

Figure 1 shows a block diagram of the HMM-based speech synthesis system (HTS) used for speaker adaptation. As our basic system we used a version published by the EMIME project[1]. The system input in the training phase consists of a training set of speech signal waveforms and corresponding labels. Labels contain symbolic representations (phones) of the speech signal content and contextual information like phonetic or linguistic features. This input

[1] EMIME - http://www.emime.org/

is then used to train Hidden Markov Models (HMMs). In the synthesis phase, labels are used to synthesize a corresponding speech signal from the models. New labels can be generated from text using methods of text analysis. Multiple speakers can be combined in an average voice model and speaker adaptation can then be used to derive the voice of a specific speaker from it [4, 5, 6].
For our experiments we employ 5-state Hidden Semi-Markov Models (HSMMs) [7]. We extract 40 mel-frequency cepstral coefficients [8], fundamental frequency F_0 (modeled as multi-space probability distribution [9]) and a set of 25 band-limited aperiodicity measures [10] from the speech signal. Also, dynamic features were used to improve continuity of the generated speech spectra [11]. The decision-tree based context clustering technique as described in [12] and as available in HTS has been used to share model parameters across multiple contexts. We use different sets of decision tree questions for each variety. These are partially handcrafted as well as automatically generated from our phone set definitions.

3 Language transformation

Recently, different methods for cross-lingual transformation have been developed. The goal is to transform data in language L_1 to language L_2 while retaining the original speaker characteristics. These methods operate on different levels.

3.1 Frame-level transformation

Qian et al. [13] developed a method that operates on the frame level. Using frequency warping to perform voice conversion, the voice of a speaker in L_2 is converted to the voice of another speaker in L_1. This method of transforming the voice of a speaker is also known as vocal tract length normalization (VTLN) [14]. In a second step, the waveforms generated from the acoustic model trained on the frequency warped data are used to guide a waveform unit selection process.

3.2 State-level transformation

Wu et al. [15] propose a state-level adaptation method. They use Kullback-Leibler Divergence (KLD) to generate a mapping between probability density functions of average voice models of L_1 and L_2. They describe two alternative approaches called data mapping and transform mapping. In data mapping, the KLD-mapping is applied to the adaptation data (in L_1) and then speaker adaptation is used to generate transforms from the L_2 average voice to the mapped adaptation data. In transform mapping, first transforms from the L_1 average voice to the adaptation data are generated. Then the transforms are attached to the mapped state models in L_2. As our work is based on a data mapping approach, this procedure will be described in more detail in the Section 4.2. Liang and Dines [16] propose an extended method. They apply a decision tree to cluster the probability density functions into phonetic categories. Mapping is then restricted to only occur within these clusters. They report a reduction of mel-cepstral distortion and subtle improvements detected during subjective listening tests.

3.3 Phone-level transformation

Wu et al. [17] also proposed a phone level based adaptation method. They apply a mapping on phones to achieve interlingual speaker adaptation between English and Mandarin Chinese.

4 Cross-Variety adaptation

Based on the state-level transformation method by Wu et al. [15], we integrated a state mapping mechanism into our cross-variety adaptation system. Given data from multiple speakers in varieties V_1, for which also adaptation data exist, and V_2, to which the voice model should be transformed, we build average voice models [6], denoted as AVG_1 and AVG_2 respectively. The corresponding decision trees will be denoted as DT_1 and DT_2. Note that DT_1 and DT_2 actually consist of multiple trees for mel-frequency cepstral coefficients, $F0$, aperiodicity and duration for each HMM state.

4.1 Mapping function

For every probability density function (pdf) $A \in AVG_1$, a $B \in AVG_2$ which minimizes KLD is determined.

$$M(A) = \arg\min_{B} \text{KLD}(A, B) \tag{1}$$

Equation 1 defines a mapping function M from AVG_1 to AVG_2.

Figure 2 shows an illustration of the relation between decision tree, pdf and KLD-mapping. For example, “mcep_s2_12” refers to the 40-dimensional pdf number 12 for the mel-frequency cepstral coefficients in HMM state 2. The decision tree questions used in this illustration consist of two parts. The second part is a phonetic symbol from our phone set definitions, for example “ks” as the “x”-sound in “wax”. The first part of the question can be “C” for center, “L” for left and “R” for right, referring to the position of the phone in question. Note that this is an artificial example for conceptual illustration. As previously mentioned, multiple trees for the feature streams for each HMM state have been used, resulting in 15 decision trees and 5 additional decision trees for duration modeling. For example, our experiments for transforming an Austrian German (AT) average voice to an Innervillgraten (IVG) average voice resulted in 13,808 mappings. This makes vivid visualizations difficult as this means there are 13,808 leaf nodes in the Austrian

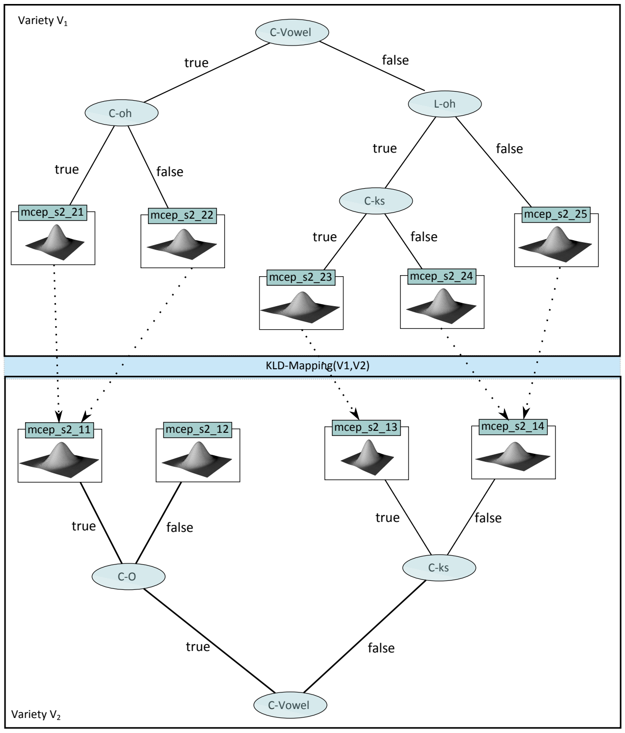

Figure 2. KLD-Mapping between probability density functions clustered by decision tree.

German decision trees. There are 1,700 different questions available for this variety.

4.2 Data mapping

In this step, we want to map the probability density functions of the speaker to be adapted from V_1 to V_2 using the mapping function M as defined in the previous sections. This is implemented as described in [15].

The first step is to classify all adaptation data labels using the decision tree DT_1. This yields pdfs for all states and feature streams for each label. We denote the set of these pdfs as S_1. Next we replace every link from the labels to every pdf $C \in S_1$ with a link to pdf $M(C) \in AVG_2$. See the doted arrows in Figure 2 for an example: the link to the pdf named “mcep_s2_22” from AVG_1 will be replaced with a link to “mcep_s2_1” from AVG_2. Now every adaptation data label is associated with a number of pdfs in AVG_2, making the model compatible with AVG_2.

4.3 Regression tree generation

Wu et al. [15] do not describe the method used to build the regression tree which is used to generate data clusters for which transformations will be trained. While the adaptation data are now compatible to AVG_2, the decision tree DT_2 is not adequate to handle labels in variety V_1, especially if the phonetic structures of V_1 and V_2 are very different. In the worst case, all labels would be placed in the same category by this decision tree. This could happen if the tree consists only of questions not fitting to V_1. For example, if the sets of phones of the two varieties are completely disjunct, all decision tree questions concerning phone symbols will yield false. Having all adaptation data in a single leaf node would on the other hand lead to a single, global transformation generated in the adaptation step.

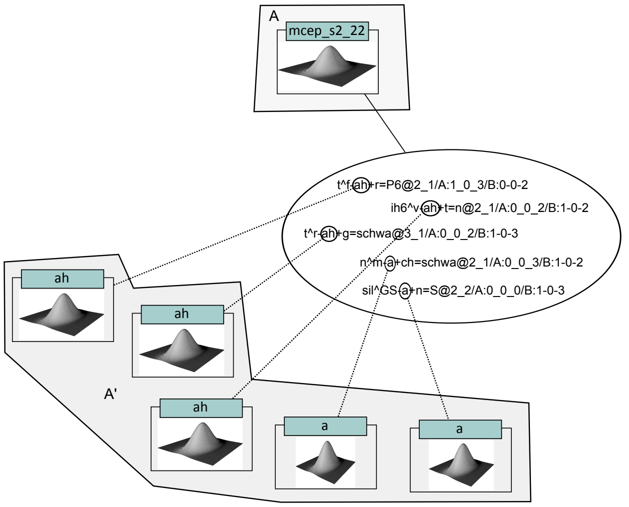
A
mcep_s2_22
t^f-ah+r=P6@2_1/A:1_0_3/B:0-0-2
ih6^v-ah+t=n@2_1/A:0_0_2/B:1-0-2
t^r-ah+g=schwa@3_1/A:0_0_2/B:1-0-3
n^m-a+ch=schwa@2_1/A:0_0_3/B:1-0-2
sil^GS-a+n=S@2_2/A:0_0_0/B:1-0-3
ah
ah
A'
ah
a
a

Figure 3. Relationship between full-context pdf A and monophone set of pdfs A'.

Our solution to this problem is to place the labels from S_1 into the leaf nodes of DT_2 not according to their decision tree questions but to their associated mapped pdfs. Using Figure 2 as an example, a label from V_1 that would be placed in “mcep_s2_22” in DT_1 will be part of the node “mcep_s2_11” in decision tree DT_2. Again note that each label has one pdf associated for each available decision tree, so this process is repeated on multiple trees.

To build the regression tree, we delete leaf nodes from DT_2 and move their associated labels to their parent node until the number of adaptation labels associated to every leaf node is above a certain threshold. As these leaf nodes then form the regression classes, this method assures that every regression class contains a certain amount of adaptation data for the calculation of the transformation. We modified the regression tree building method of HTS to reflect this strategy.

5 Integrating structural information

We also extended the previously described method to add weight to structural, phonetic information in the mapping process. Hence the mapping function M (Equation 1) is replaced by M' (Equation 2).

$$M'_\lambda(A) = \arg\min_B(\lambda \text{KLD}(A,B) + (1-\lambda)\text{KLD}_{mono}(A,B)) \tag{2}$$

With every pdf A we associate a set of monophone pdfs A' with A'_i being the i-th pdf in the set. Given a set of monophone pdfs A' for A and B' for B, we then calculate KLD$_{mono}$ as in Equation 3.

$$\text{KLD}_{mono}(A,B) = \frac{\sum_i \sum_j \text{KLD}(A'_i, B'_j)}{(|A'||B'|)} \tag{3}$$

The relationship between A and A' is illustrated in Figure 3. Each pdf $A \in AVG_1$ has a number of labels from the training data of AVG_1 associated with it. These are the labels which would be placed in the node of A when classified using the decision tree DT_1. We retrieve these labels, extract the center phones (encircled symbols "ah" and "a" in Figure 3) and find their associated monophone pdfs created during the average voice building.

Figure 3 also illustrates the fact that a single pdf often covers labels with different center phones. Often, these center phones are phonetically close. We repeat this process for B and then calculate the mean KLD of all combinations of pdfs from A' with all pdfs from B'. This results in our monophone KLD function KLD$_{mono}(A,B)$ which is then linearly interpolated with the regular KLD KLD(A,B) using an interpolation parameter λ. We then select the mapping with the lowest value for $M'(A,B)$.

As calculating the monophone KLD for all possible combinations of A and B is computationally expensive, we first calculate the n best regular KLD values and calculate the monophone KLD for those combinations only. In our experiments, we set $n = 50$.

For the target speaker used in our evaluation, we trained voice models for λ values from 0.5 to 1.0 in steps of 0.02. We then selected the model with the highest likelihood as calculated by the HTS system. See Figure 4 for the search space for the data used in our evaluation. For the interpretation of the likelihood values used in HTS, see [18]. Experiments with other adaptation speakers showed that the behavior of likelihood when varying λ is unpredictable with many local maxima, making efficient optimization difficult. But as average voice building and feature extraction of adaptation data has to be done only once, sampling the search space in even smaller steps for λ than 0.02 still seems reasonable.

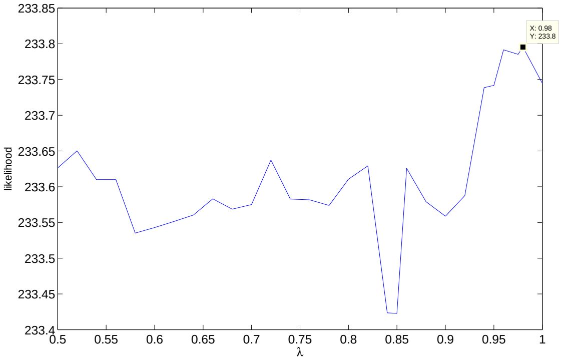

Figure 4. Effect of interpolation parameter λ on model likelihood.

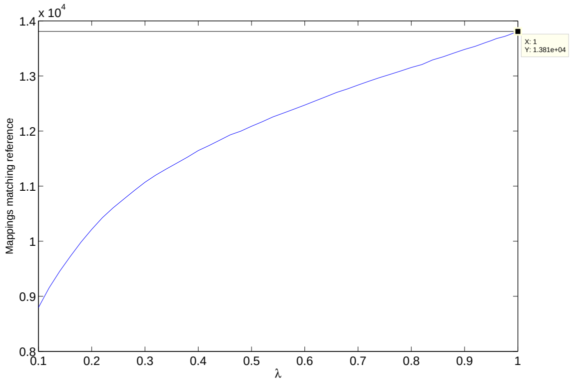

Figure 5. Mappings that match the reference (mapping with regular KLD) when varying λ.

To see the effect of integrating monophone KLD into the equation, we compared the resulting mappings for different values of λ with the mappings that result from using regular KLD only (which is equivalent to using $\lambda = 1.0$). Using our Austrian German and Innervillgraten average voice models, we generated optimal mappings using M and M'_λ for different values for λ. Figure 5 shows the number of mappings that exist in both result sets plotted against λ. As expected, the number of matching mappings increases with increasing values of λ. At the peak of model likelihood at $\lambda = 0.98$, only about 100 mappings are different from $\lambda = 1.0$. Further research is needed to analyze the influence of the pdfs involved in these mappings, as the importance of single pdfs to the final speech can differ tremendously. Also, pdfs that are rarely used during the synthesis process could still have an important impact when they amplify or produce a rare but strongly perceptible error in the output speech.

6 Evaluation

For our evaluation we used our male Austrian German average voice, our male Innervillgraten average voice and we used data from an Austrian German speaker to build an Innervillgraten voice model using the cross-variety transformation mechanism described previously. From this model, we synthesized a test set of 21 utterances that have been excluded from the training. As a baseline, we used the same speaker adaptation mechanism without the cross-variety extensions described previously.

	mean			median			std. deviation	
	regular	cross-variety	Δ	regular	cross-variety	Δ	regular	cross-variety
speaker similarity	2.494	2.940	+0.446	2	3	+1	0.929	1.0517
language similarity	2.282	3.270	+0.988	2	3	+1	0.923	1.063
overall quality	2.186	3.212	+1.025	2	3	+1	0.931	0.962

Table 1. Evaluation of sample scores.

We conducted a subjective listening test[2] with 5 expert listeners with a speech processing or linguistics background. Each expert listened to the results for all 21 utterances for both methods. The listeners were not given any information on the method used to synthesize each sample. Also, the positions of the samples on the evaluation interface were swapped randomly for each utterance. For each utterance, the listeners had to answer questions for language similarity, speaker similarity and overall quality. For language similarity, a sample of the same utterance from a native Innervillgraten speaker was provided as reference. For speaker similarity, an unrelated utterance from the original recordings of the adaptation speaker was provided as reference.

The questions to be answered were:

- Which sample sounds more similar to the reference in terms of speaker identity?
- Rate the speaker similarity with the reference for each sample (1 - very different, 5 - very similar).
- Which sample sounds more similar to the reference in terms of language variety?
- Rate the language variety similarity with the reference for each sample (1 - very different, 5 - very similar).
- Which sample has the better overall speech quality?
- Rate the quality of each sample (1 - very bad quality, 5 - very good quality).

When the listener would not prefer one sample over the other, selecting none was allowed.

Table 1 and Figure 6 show the results for the rating questions. It can be seen that the mean and median score was higher in all categories for cross-variety adaptation compared to regular speaker adaptation. Wilcoxon rank sum test and Welch two sample t-test both resulted in $p < 0.0001$ for both language similarity and overall quality as well as in $p < 0.005$ for speaker similarity.

The results for the questions where the listeners had to choose one sample over the other are shown in Table 2. It can be seen that the listeners were undecided on 25 samples when evaluating speaker similarity. This is consistent with the fact that the differences in speaker similarity scores were much lower than the other categories.

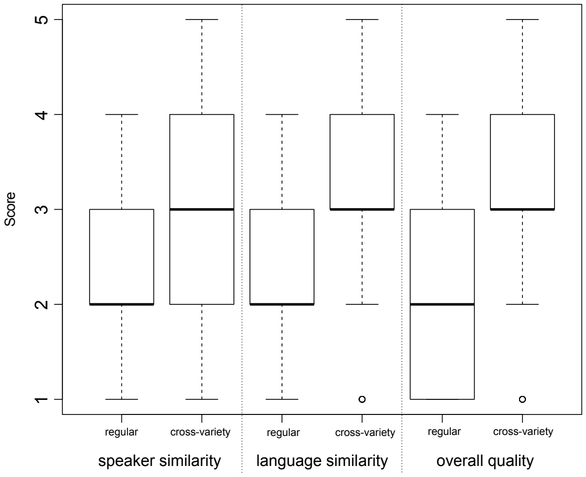

Figure 6. Boxplots for listening test scores.

	regular	cross-variety	undecided
speaker similarity	15	65	25
language similarity	19	84	2
overall quality	15	87	3

Table 2. Evaluation of preferred samples.

7 Conclusion

A method for cross-variety transformation based on state mapping [15] has been presented in this article. We described our approach to regression tree generation and the integration of structural information into the mapping process. During the subjective listening test it became evident that regular speaker adaptation is not sufficient for cross-variety adaptation. The method presented here greatly improves the quality of the generated voice. Unfortunately it is still very dependent on good quality average voices. An average voice biased to a very distinctive speaker will

[2] Samples: http://userver.ftw.at/~mtoman/sppra2013/

also introduce noticeable elements of this speaker into the adapted voice. Also, errors in the synthesized speech of the average voices will also be noticeable in synthesized speech of the adapted voice. This is especially relevant for varieties with few available speech data. While overall voice quality and language similarity increased significantly, speaker similarity is still an important topic for further improvement. To this end, we plan to integrate frame-level conversion mechanisms to make more efficient use of the adaptation data. Similarities in the varieties involved could also be exploited more efficiently by further research at the level of the regression tree and speaker adaptation itself.

References

[1] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom. "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis". In: *Speech Communication* 52.2 (Feb. 2010), pp. 164–179.

[2] M. Pucher, F. Neubarth, V. Strom, S. Moosmller, G. Hofer, C. Kranzler, G. Schuchmann, and D. Schabus. "Resources for speech synthesis of Viennese varieties". In: *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC)*. 2010.

[3] M. Pucher, N. Kerschhofer-Puhalo, D. Schabus, S. Moosmüller, and G. Hofer. "Language resources for the adaptive speech synthesis of dialects". In: *Proc. of the 7th Congress of the International Society for Dialectology and Geolinguistics*. Vienna, Austria, July 2012.

[4] J. Yamagishi and T. Kobayashi. "Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training". In: *IEICE Transactions on Information and Systems* E90-D.2 (Feb. 2007), pp. 533–543.

[5] H. Zen, K. Tokuda, and A. W. Black. "Statistical parametric speech synthesis". In: *Speech Communication* 51.11 (Nov. 2009), pp. 1039–1064.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai. "Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm". English. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.1 (Jan. 2009), pp. 66–83.

[7] S. Levinson. "Continuously variable duration hidden Markov models for automatic speech recognition". In: *Computer Speech & Language* 1.1 (Mar. 1986), pp. 29–45.

[8] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. "An adaptive algorithm for mel-cepstral analysis of speech". In: *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1992, 137–140 vol.1.

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling". In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1999, 229–232 vol.1.

[10] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds". In: *Speech Communication* 27.3-4 (Apr. 1999), pp. 187–207.

[11] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. "Speech synthesis using HMMs with dynamic features". In: *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, 1996, pp. 389–392.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". In: *Proc. of Eurospeech*. 1999.

[13] Y. Qian, J. Xu, and F. K. Soong. "A frame mapping based HMM approach to cross-lingual voice transformation". In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 5120–5123.

[14] L. Lee and R. Rose. "A frequency warping approach to speaker normalization". English. In: *IEEE Transactions on Speech and Audio Processing* 6.1 (1998), pp. 49–60.

[15] Y.-J. Wu, Y. Nankaku, and K. Tokuda. "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis". In: *INTERSPEECH*. ISCA, 2009, pp. 528–531.

[16] H. Liang and J. Dines. "Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation". In: *Proc. of Interspeech*. Idiap-RR-17-2011. Florence, Italy, Aug. 2011.

[17] Y.-J. Wu, S. King, and K. Tokuda. "Cross-Lingual Speaker Adaptation for HMM-Based Speech Synthesis". English. In: *6th International Symposium on Chinese Spoken Language Processing*. IEEE, Dec. 2008, pp. 1–4.

[18] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.