Cross-variety speaker transformation in HSMM-based speech synthesis

Markus Toman, Michael Pucher, Dietmar Schabus

Telecommunications Research Center (FTW), Vienna, Austria

{toman,pucher,schabus}@ftw.at

Abstract

We present and compare different approaches for cross-variety speaker transformation in Hidden Semi-Markov Model (HSMM) based speech synthesis that allow for a transformation of an arbitrary speaker's voice from one variety to another one. The methods developed are applied to three different varieties, namely standard Austrian German, one Middle Bavarian (Upper Austria, Bad Goisern) and one South Bavarian (East Tyrol, Innervillgraten) dialect. For data mapping of HSMM-states we use Kullback-Leibler divergence, transfer probability density functions to the decision tree of the other variety and perform speaker adaptation. We investigate an existing data mapping method and a method that constrains the mappings for common phones and show that both methods can retain speaker similarity and variety similarity. Furthermore we show that in some cases the constrained mapping method gives better results than the standard method.

Index Terms: speech synthesis, dialect, transformation, language variety

1. Introduction

Acoustic language transformation has received much interest in the last years [1, 2, 3]. In general, it is the problem of transforming a speaker's voice into another language retaining speaker identity. Language transformation has applications in speech-to-speech translation and education. In this paper we consider a restricted version of the problem where we want to transform a speaker's acoustic model into a model of the same speaker in a different variety of the same language. A variety can be a dialect, sociolect, or accent. Here we apply variety transformation to standard and dialect.

Although being a simpler problem, it still has an interesting range of possible applications like language learning. If a user wants to learn a certain variety V_{learn} (dialect, sociolect, accent), a variety transformation system can transform his or her variety V_{user} into the target variety V_{learn}. One can then use samples of this voice for learning and comparison with speech that is produced by the user. The main application of such a system would consist in teaching the standard variety to speakers with non-standard varieties but of course it is also possible to use the transformation in the other direction. In this paper we consider transformations between dialects, from the standard to the dialect and from dialect to standard.

We have previously shown how to achieve a one-way transformation between standard and dialect [4]. In this paper, we consider one standard variety and two different dialects and perform all possible transformations. The modeling techniques developed here can further be applied to accented speech. As a first method, we implemented a data mapping approach that is described below and in [4]. We extend this data mapping approach by a constraint-based approach where we map only between models that are in an overlapping phone set. The basic idea of the second method is to exclude mappings where the phones are common in the phone set but the representative phones of the mapped models are different.

2. Data

In previous and current projects we recorded and annotated phonetically balanced speech data in two dialectal Austrian varieties from Innervillgraten in East Tyrol (IVG) and from Bad Goisern in Upper Austria (GOI), as well as standard Austrian German (AT). The data acquisition process is described in [5, 6]. The main problems of recording dialect speech are the missing orthographic standard that define how speech is produced from a written form, the missing linguistic resources, and the speaker selection. In this paper we focus on transformations between varieties.

Table 1: *Non-existent phones in mapping.*

	Target variety		
	AT	IVG	GOI
AT	-	49/90 (0.54)	58/100 (0.58)
IVG	35/78 (0.45)	-	23/100 (0.23)
GOI	36/78 (0.46)	23/90 (0.26)	-

Table 1 describes the phone set relations. If we want to tranform AT to IVG, for example, we have to model 49 of the 90 phones in IVG that are nonexistent in AT. The higher the ratio in the table, the more difficult we expect the corresponding variety transformation pair. We can also see that the ratio of missing phones is larger when we have the standard (AT) as the source variety because the phone overlap between the two dialects (GOI, IVG) is larger than the overlap between the standard and either of the two dialects.

3. Speaker-adaptive acoustic modeling

For speaker-adaptive acoustic modeling we used a version of the HSMM-based speech synthesis system (HTS) as published by the EMIME project [7] for our experiments. The input to the system in the training phase is a training set of speech signal waveforms and corresponding full-context label files. These labels contain symbolic representations (phones) of the speech signal as well as contextual information like phonetic and linguistic features. Using this input, speaker-adaptive Hidden semi-Markov average voice Models (HSMMs) are trained for all varieties. In the synthesis phase, labels from a test set are used to generate a synthesized speech signal from the trained models. Methods from text analysis can be used to generate new labels. Multiple speakers can be combined in an average

voice model. Speaker adaptation can then be used to derive a speaker-specific model from this average voice model [8].

Five-state HSMMs are employed in our experiments. We extract 40 mel-frequency cepstral coefficients, fundamental frequency F_0 (modeled as multi-space probability distribution [9]), and a set of 25 band-limited aperiodicity measures from the speech signal. Dynamic features were used to improve continuity of the generated speech spectra [10]. The decision-tree based context clustering technique as described in [11] 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.

4. Speaker-adaptive cross-variety transformation

Here we present a cross-variety transformation system that is based on a speaker-adaptive HSMM-based speech synthesis system [7]. The system uses average voice models of the source and target varieties for cross-variety transformation.

Based on the state-level transformation described in [2], we integrated a state mapping mechanism into our cross-variety adaptation system. Using 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 train average voice models [8], denoted as AVG_1 and AVG_2, respectively. The decision trees for those models will then be denoted as DT_1 and DT_2, respectively. DT_1 and DT_2 actually consist of multiple trees for mel-frequency cepstral coefficients, $F0$, aperiodicity and duration for each of the five HSMM states. Since AVG_1 and AVG_2 were trained on different data with different (albeit overlapping) phone sets, they also have a different decision tree structure.

4.1. Data mapping

For every probability density function (pdf) $A \in AVG_1$, we find a pdf $B \in AVG_2$ which minimizes the Kullback-Leibler divergence (KLD),

$$M(A) = \underset{B \in AVG_2}{\arg\min}\ \mathrm{KLD}(A, B), \tag{1}$$

which defines a mapping function M from AVG_1 to AVG_2.

In Figure 1, an illustration of the relation between decision tree, pdf and KLD-mapping can be seen. For example, "mcep_s2_12" refers to the 40-dimensional pdf number 12 for the mel-frequency cepstral coefficients in HSMM 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 "t" as in "hat". 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.

In the actual system, multiple trees for the feature streams for each HSMM state have been used, resulting in 15 decision trees and 5 additional decision trees for duration modeling. Also, a typical set of questions for the decision tree in our case consists of 1,700 different questions. For example, we calculated the mapping between an Austrian German (AT) average voice and an Innervillgraten (IVG) average voice. This mapping consisted of 13,808 pdf pairs. Therefore, the Austrian German decision trees have 13,808 leaf nodes, making vivid visualizations and manual analysis difficult.

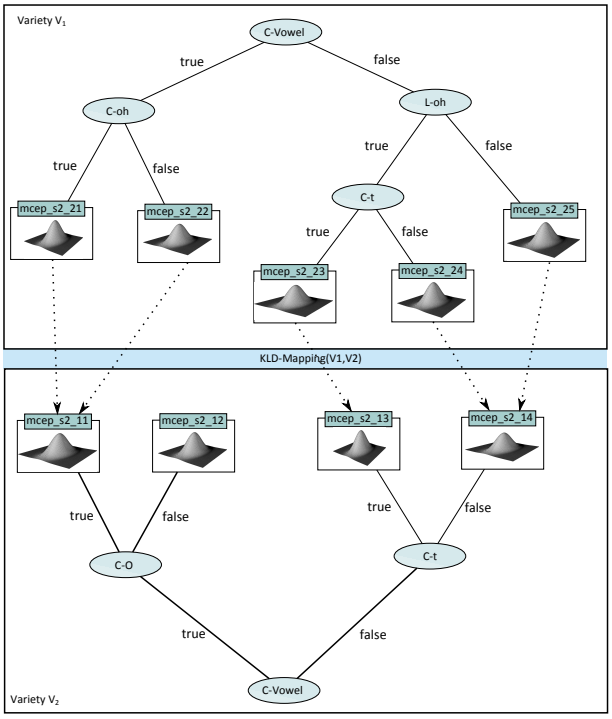

Figure 1: *KLD-Mapping between probability density functions clustered by decision tree.*

Using mapping function M, we map the pdfs of the speaker to be adapted from V_1 to V_2. This is implemented as described in [2, 4].

4.2. Constrained mapping

In addition to the data mapping approach described above, we also investigated a constrained mapping approach. This is based on the idea that mappings should only be made between the same phones, if existent in both varieties. We can however not fulfill this constraint directly, since the mapping is not defined on the phoneme level but only on the model (pdf) level.

To implement the constraint, we apply the following algorithm:
For each pdf $A \in AVG_1$ of a certain variety V_1, we define the representative phone $r(A)$ as the center phone that is most common in the associated full-context labels. Furthermore, we define $P(A)$ as the list of all center phones in all labels used to train A. Figure 2 illustrates how $r(A)$ and $P(A)$ are defined. First we find all full-context labels that have been used to train the corresponding pdf. $P(A)$ is then the set of all center phones from this list of labels and $r(A)$ is the center phone occurring most often. Another possible method to calculate $r(A)$ would be to weigh each label with the number of associated samples, as these have greater influence on the pdf estimation. Next we constrain the mappings on the set of common phones. If the representative phone $r(A)$ occurs not only in phone set P_1 of variety V_1 (as it does per definition) but also in phone set P_2 of variety V_2, then we only map from $A \in AVG_1$ to $B \in AVG_2$, if $r(A)$ is in $P(B)$.

So the common phone data mapping $M(A) = B$ from

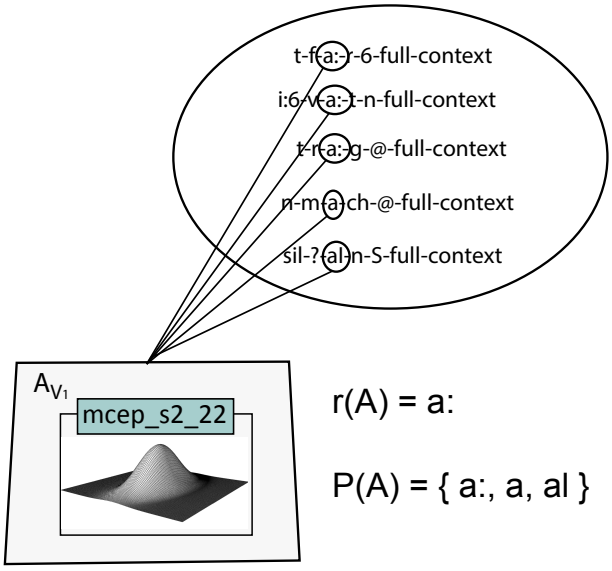

Figure 2: *Definition of representative phone $p(A_{V_1})$ and list of all center phones $pa(A_{V_1})$.*

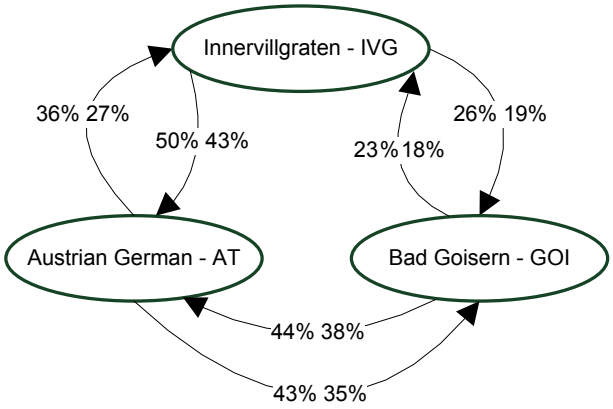

Figure 3: *Transformations between varieties.*

Table 2: Results of the variety similarity judgment part of the evaluation.

Compared methods	wins	ties
DM : CPDM	31 : 27	82

AVG_1 to AVG_2 must fulfill the conditional constraint given in Equation 2:

$$(r(A) \in P_2) \Rightarrow (r(A) \in P(B)) \quad (2)$$

In other words, if the representative phone $r(A)$ occurs in both varieties, we discard all potential mappings $M(A) = B$ for which $r(A)$ is not in the training data of B. If $r(A)$ does not occur in both varieties, we keep all potential mappings (A, B). Of the remaining potential mappings, the mapping with the lowest KLD value as in Equation 1 is selected for further processing.

A possibility to make the constraint stronger would be to require that $r(A)$ is also the most common phone in P_2, so $r(A) = r(B)$. However, we did not evaluate this stronger constrained variant.

Figure 3 shows the percentage of 1-best and 200-best mappings between the different varieties that fulfill the constraint given by Equation 2. n-best means the n mappings with the lowest KLD score. To map from AT to IVG, for example, there are 36% in the 1-best lists and 27% in the 200-best list where the mapping is between equal phones if there is an overlap. The relations here are similar to the nonexistent phones given in Table 1. The smaller the phone overlap, the fewer mappings fulfill the constraint.

4.3. Regression tree generation

For generating the regression tree, we use the algorithm as described previously [4]. 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 place labels from the data set of V_1 into the leaf nodes of DT_2 not according to their decision tree questions but to their associated mapped pdfs. Using Figure 1 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.

We modified the regression tree building method of HTS to reflect this strategy.

5. Evaluation

We conducted a subjective and an objective evaluation as well as an analysis of specific cases. These will be described in the following sections.

5.1. Subjective evaluation

To evaluate the two methods *Data Mapping* (DM) and *Common Phone Data Mapping* (CPDM), we have carried out a subjective listening evaluation with 27 listeners participating (17 males and 10 females, aged 20 to 55, mean age 28.95), all native German speakers from different regions in Austria, including 9 listeners from our target regions (East Tyrol or Upper Austria). The evaluation consisted of two parts. In the first part we compared synthesized samples from the two methods with a reference signal, and asked the listeners which synthesized sample they found to be more similar to the reference signal in terms of variety. The reference signal was a recording of the same sentence spoken by a (different) speaker of the target variety. We assume that this experiment design allows that listeners who are not themselves speakers of the target variety can still judge the variety similarity. The results in Table 2 show that method DM was considered more similar 31 times, CPDM was considered more similar 27 times and 82 times they were considered equally similar to the reference. The difference in the number of "wins" (31 vs. 27) is not statistically significant according to a Bonferroni-corrected Pearson's χ^2-test of independence ($p > 0.58$). This and the large number of "ties" suggest that none of the two methods is superior to the other.

Additionally, we asked the listeners to specify the degree of similarity concerning variety for the "winning" method (or both methods, in case of a tie), by choosing one of the five options

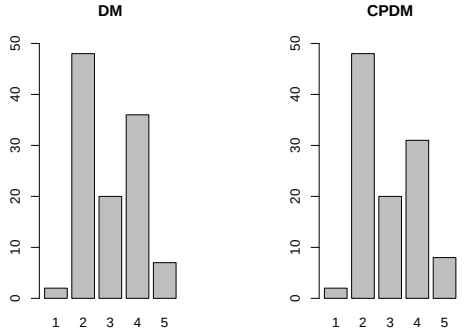

Figure 4: *Frequencies of variety similarity votes for the two methods. 1 means "very similar" and 5 means "very different".*

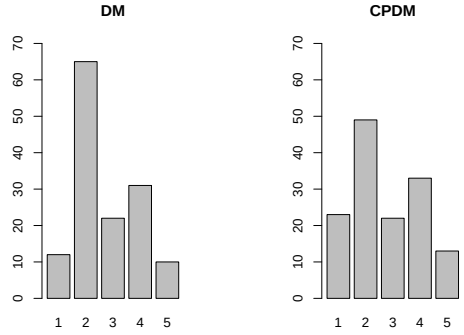

Figure 5: *Frequencies of speaker similarity votes for the two methods. 1 means "very similar" and 5 means "very different".*

Table 3: *Results of the speaker identification part of the evaluation.*

Method	correct	wrong	undecided	sig.
DM	91	35	14	*
CPDM	91	28	21	*

"very similar", "similar", "no opinion", "different" and "very different". The results are given in Figure 4 as frequency bar plots, where 1 means "very similar" and 5 means "very different". We can see that "similar" was the most frequently chosen option. The number of votes for "different" can be explained by the difficulty of the concept variety similarity, which often includes a factor of authenticity. Authenticity can be affected negatively by the overall quality of synthetic speech. Furthermore, there are listeners who are not native speakers of the target varieties.

In the second part of the evaluation, the goal was to assess speaker similarity. We showed the listeners one synthesized sample from one of the two methods and two recorded reference samples. The two references were both the same utterance in a variety different from the target variety of the synthesized sample, one from the target speaker and one from a randomly selected different speaker. The listeners were asked to decide to which of the two references the synthesized sample sounded more similar in terms of speaker identity. The results are given in Table 3, where for each of the two methods, the number of correct, wrong and undecided judgments are presented. For both methods, the number of correct speaker identifications is statistically significantly higher than the number of wrong speaker identifications (Bonferroni-corrected Pearson's χ^2-test of independence with $p < 0.001$).

Again, the listeners were also asked to specify the degree of similarity by choosing one of the five options "very similar", "similar", "no opinion", "different" and "very different". The results are given in Figure 5 as frequency bar plots, where 1 means "very similar" and 5 means "very different". It can be seen that while the number of votes for "similar" decreased for CPDM compared to DM, the number of votes for "very similar" increased.

Table 4: *Results of the objective evaluation.*

	DM		CPDM		SD	
Speaker	μ	σ	μ	σ	μ	σ
IVG 1	6.33	0.43	6.29	0.36	5.42	0.31
IVG 2	7.01	0.72	6.68	0.66	5.46	0.29
GOI 1	6.85	0.43	6.86	0.44	5.90	0.47
GOI 2	7.03	0.74	7.12	0.73	6.10	0.81

5.2. Objective evaluation

We also conducted an objective evaluation by calculating mel-cepstral distortion between the trajectories resulting from the presented methods and trajectories extracted from original recordings. This was possible as we had recordings of standard as well as dialect from some of the speakers. For the analysis we used our AT test set consisting of 23 utterances. We transformed two IVG and two GOI speakers to AT and calculated mel-cepstral distortion between the synthesized results and the AT recordings of the same speaker for the same utterance. The samples were synthesized using the phone durations obtained by automatic alignment of the test recordings.

Table 4 shows the result of this analysis for the four speakers. It can be seen that the mean mel-cepstral distortion is lower for CPDM than for DM for the IVG speakers while it is higher for the GOI speakers. However, only the difference for speaker IVG 2 is significant according to a Bonferroni-corrected Paired t-test ($p < 4 \times 10^{-8}$). This shows that CPDM improves the model for one speaker and does not corrupt the model for the others.

We also trained speaker-dependent (SD) models (using 223 utterances) in AT for every speaker for reference. The mean values and standard deviations for the speaker dependent models compared to the recordings can also be seen in Table 4. As expected, all speaker-dependent AT models have significantly ($p < 2 \times 10^{-14}$) lower mel-cepstral distortion compared to the cross-variety transformation models. This shows that the mel-cepstral distortion metric covers aspects of speaker similarity.

5.3. Analysis of specific cases

When manually inspecting the synthesized waveforms, we noticed structures that remarkably differed for the DM and

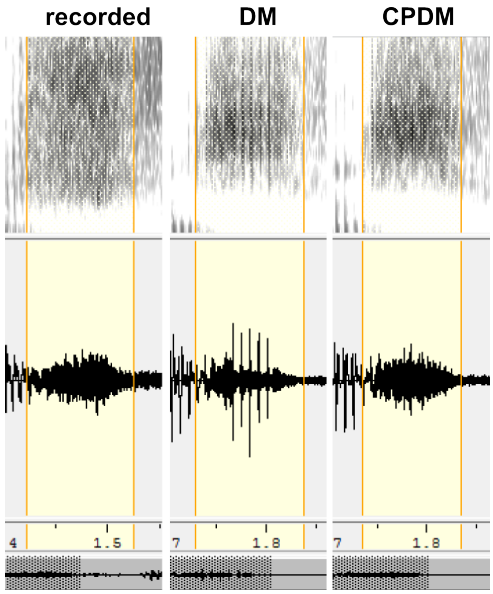

Figure 6: *Waveforms for "s" as recorded and synthesized using DM and CPDM.*

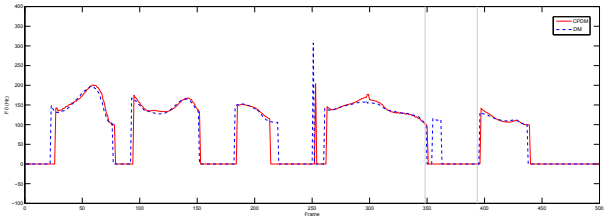

Figure 7: *Different F0 trajectory for DM and CPDM.*

CPDM methods. When listening to these parts, we could find glitches in DM that were absent in CPDM. While these glitches were quite distinct, they did not seem to be the predominant factor to influence scoring in the subjective listening test.

For an example, consider Figure 6. The highlighted section of the waveform corresponds to the main part of the phone "s" and is presented for the recording of a GOI speaker and an AT speaker transformed to GOI using DM and CPDM method.

It can be seen that the waveform generated by CPDM is smoother and more closely resembles the waveform of the natural "s". In the DM version, the "s" has a crackling effect that is absent in the CPDM synthesis. Analyzing the different trajectories for this sample resulted in Figure 7. The "s"-sound is highlighted and it can be seen that the f0 values differ in this region. DM produces a voiced region compared to the correct, unvoiced region produced by CPDM. Reasons for this behavior remain subject of further investigation.

Listening samples for some specific cases can be found on the demopage[1].

6. Conclusion and future work

We compared different approaches for cross-variety speaker transformation in HSMM-based speech synthesis. The developed methods were applied to three different varieties. We investigated a standard data mapping method and a mapping method that constrains the mappings for common phones. In the subjective evaluation, we saw that both data mapping methods can retain speaker similarity to a high degree and variety similarity to a smaller degree. In the pairwise comparison we did not see significant differences between the two methods. This conforms with prior work where different data mapping approaches also led only to subtle changes in the results [3].

We performed an objective evaluation for the bi-lingual data of speakers in two varieties. One of four speakers showed a significant improvement in mel-spectral distortion for CPDM over DM.

We also analyzed specific cases and reported one of them in this article. Further analyses would be necessary to gain a deeper understanding of these effects.

[1] Synthesis samples on http://userver.ftw.at/~mtoman/ssw2013/t

7. Acknowledgements

This research was funded by the Austrian Science Fund (FWF): P23821-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

8. References

[1] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation", in Proc. ICASSP, Prague, Czech Republic, 2011, pp. 5120-5123.

[2] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis", in Proc. INTERSPEECH, Brighton, United Kingdom, 2009, pp. 528-531.

[3] H. Liang and J. Dines. "Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation", in Proc. INTERSPEECH, Florence, Italy, 2011, pp. 1825–1828.

[4] M. Toman, M. Pucher, "Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis", in Proc. SPPRA, Innsbruck, Austria, 2013.

[5] 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", Speech Communication, 52(2):164–179, 2010.

[6] M. Pucher, N. Kerschhofer-Puhalo, D. Schabus, S. Moosmüller, G. Hofer, "Language resources for the adaptive speech synthesis of dialects", in Proc. SIDG, Vienna, Austria, 2012.

[7] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard challenge 2010", in Blizzard Challenge Workshop, Kansai Science City, Japan, 2010.

[8] 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", IEEE Transactions on Audio, Speech, and Language Processing 17.1 (Jan. 2009): 66-83, 2009.

[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. ICASSP, Phoenix, AR, USA, 1999, pp. 229–232.

[10] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features". in Proc. ICASSP, Atlanta, GA, USA, 1996, pp. 389-392.

[11] 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. EUROSPEECH, Budapest, Hungary, 1999, pp. 2347–2350.