STATISTICAL METHODS FOR VARYING THE DEGREE OF ARTICULATION IN NEW HMM-BASED VOICES

Benjamin Picart, Thomas Drugman, Thierry Dutoit

TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Belgium

ABSTRACT

This paper focuses on the automatic modification of the degree of articulation (hypo/hyperarticulation) of an existing standard neutral voice in the framework of HMM-based speech synthesis. Starting from a source speaker for which neutral, hypo and hyperarticulated speech data are available, two sets of transformations are computed during the adaptation of the neutral speech synthesizer. These transformations are then applied to a new target speaker for which no hypo/hyperarticulated recordings are available. Four statistical methods are investigated, differing in the speaking style adaptation technique (MLLR vs. CMLLR) and in the speaking style transposition approach (phonetic vs. acoustic correspondence) they use. This study focuses on the prosody model although such techniques can be applied to any stream of parameters exhibiting suited interpolability properties. Two subjective evaluations are performed in order to determine which statistical transformation method achieves the better segmental quality and reproduction of the articulation degree.

Index Terms— Speech Synthesis, HTS, Expressive Speech, Speaking Style Adaptation, Voice Quality

1. INTRODUCTION

The "H and H" theory [1] proposes two degrees of articulation of speech: hyperarticulated speech, for which speech clarity tends to be maximized, and hypoarticulated speech, where the speech signal is produced with minimal efforts. Therefore the degree of articulation provides information on the motivation/personality of the speaker vs. the listeners [2]. This paper is in line with our previous works on expressive speech synthesis [3] [4] [5] [6]. We here focus on the synthesis of different speaking styles, with a varying degree of articulation: neutral speech, hypoarticulated (or casual) and hyperarticulated (or clear) speech. "Hyperarticulated speech" refers to the situation of a teacher/speaker talking in front of a large audience (important articulation efforts have to be made to be understood by everybody). "Hypoarticulated speech" refers to the situation of a person talking in a narrow environment or very close to someone (few articulation efforts have to be made to be understood). It is worth noting that these three modes of expressivity are neutral on the emotional point of view, but can vary amongst speakers, as reported in [2]. The influence of emotion on the articulation degree has been studied in [7] and is out of the scope of this work.

Hypo/hyperarticulated speech synthesis has many applications: expressive voice conversion (e.g. for embedded systems and video games), "reading speed" control for visually impaired people (i.e. fast speech synthesizers, more easily produced using hypoarticulation), improving intelligibility performance in adverse environments (e.g. GPS voice inside a moving car, train/flight information in stations/halls), etc.

Starting from an existing standard neutral voice with no hypo/hyperarticulated recordings available, the ultimate goal of our research is to allow for a continuous control of its articulation degree. One way to achieve this goal is by using voice adaptation techniques in the spirit of [8], but applied here to intra-speaker voice adaptation [9]. Eigenvoice conversion techniques [10] [11] realizes the Hidden Markov Models (HMMs) speaker adaptation using a small amount of adaptation data by reducing the number of free parameters for controlling speaker dependencies of HMMs. Another way to achieve this is to use voice conversion techniques. One of the most popular voice conversion method is the probabilistic conversion based on a Gaussian Mixture Model (GMM) [12]. Voice morphing is also a technique for continuously modifying a source speaker's speech to sound as pronounced by another speaker [13].

All these methods cannot be applied to our case because we do not have target data (i.e. hypo/hyperarticulated speech data) for the existing standard neutral voice. Instead we propose to model the prosody transforms on a voice for which neutral, hypo and hyperarticulated speech data are available (Voice A), and to apply these transforms to an existing standard neutral voice (Voice B) with no hypo/hyperarticulated recordings.

This paper is therefore devoted to finding new methods for applying a prosody (pitch and phone duration) model, estimated on one voice, on another voice. These new methods can be generalizable, meaning that they could be used following the same idea to the cepstrum model. Unfortunately, our attempts to apply these transformation techniques to the Mel Generalized Cepstrum (MGC [14]) parameters, standard fil-

ter coefficients in HMM-based speech synthesis, did not currently provide convincing results as the segmental quality was poor. This is probably due to the poor interpolation properties of these features, and our future works encompass the study of the most appropriate filter parameterization for this purpose.

This paper is structured as follows. After a brief description of the contents of our database in Section 2, the implementation of our HMM-based speech synthesizer and the creation of the prosody model are detailed in Section 3. Section 4 presents different methods investigated for the application of the prosody model to an existing standard neutral voice. As no reference speech data are available, a subjective evaluation is performed in Section 5 to assess both the speech quality and the effectiveness in degree of articulation transposition. Finally Section 6 concludes the paper.

2. DATABASE WITH VARIOUS DEGREES OF ARTICULATION

For the purpose of our research, a new French database was recorded in [3] by a professional male speaker, aged 25 and native French (Belgium) speaking. The database contains three separate sets, each set corresponding to one degree of articulation (neutral, hypo and hyperarticulated). For each set, the speaker was asked to pronounce the same 1359 phonetically balanced sentences (around 75, 50 and 100 minutes of neutral, hypo and hyperarticulated speech respectively), as neutrally as possible from the emotional point of view. A headset was provided to the speaker for both hypo and hyperarticulated recordings, in order to induce him to speak naturally while modifying his articulation degree (see [3] for details on how this was induced). The 1359 sentences are split into 1220 for training the models while the rest is used for the evaluation.

3. CREATION OF THE ARTICULATION MODEL

An HMM-based speech synthesizer [15] was built for Voice A, relying on the implementation of the HMM-based Speech Synthesis System (version 2.1) HTS ("H-Triple-S" - a toolkit publicly available in [16]). 1220 neutral sentences sampled at 16 kHz were used for the training, leaving around 10% of the database for the test set. For the filter, we extracted the traditional Mel Generalized Cepstral (MGC) coefficients (with α = 0.42, γ = 0 and MGC analysis order = 24). For the excitation, we used the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in [17], since it was shown to significantly improve the naturalness of the produced speech. More precisely, both deterministic and stochastic components of DSM were estimated on the training dataset for each degree of articulation. In this study, we used 75-dimensional MGC parameters (including Δ and Δ^2). Moreover, each covariance matrix of the state output and state duration distributions were diagonal.

This neutral HMM-based speech synthesizer was adapted using the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [18] in the framework of the Hidden Semi Markov Model (HSMM) [19], using hypo/hyperarticulated speech data, to produce a Voice A hypo/hyperarticulated speech synthesizer. HSMM is an HMM having explicit state duration distributions (advantage during the adaptation process of phone duration). CMLLR is a feature adaptation technique which estimates a set of linear transformations for the features so that each state in the HMM system is more likely to generate the adaptation data. As spectrum, pitch and state duration are modeled simultaneously in a unified framework [20] [21], speaker adaptation techniques are applied simultaneously to spectrum, pitch and state duration. The linearly transformed models are further updated using a Maximum A Posteriori (MAP) adaptation [8].

In the following, the neutral full data model refer to the model trained on the entire training set (1220 neutral sentences), and the adapted models are the models adapted from the neutral full data model, using hypo/hyperarticulated speech data. We used the entire training hypo/hyperarticulated databases for the adaptation process (the effect of the amount of adaptation sentences on the synthesized speech quality was studied in [4]).

4. TRANSPOSITION OF THE ARTICULATION MODEL TO A NEW SPEAKER

In this section, the adaptation transforms computed on Voice A (see Section 3) are applied to an existing standard neutral voice (Voice B) with no hypo/hyperarticulated recordings available, in order to control its articulation degree. Voice B corresponds to a native French male. This Text-to-Speech (TTS) voice was kindly provided by Acapela Group S.A.. As stated in the introduction, we only apply the adaptation transforms to modify the prosody of Voice B. Voice B was trained using 2400 neutral sentences sampled at 16 kHz, following the same procedure as for Voice A. HTS was forced to use the same decision trees as Voice A in order to have a one-to-one mapping between the probability density functions (pdfs) of Voices A and B. Imposing decision trees has an impact on the generated speech quality, as the training process is no more allowed to construct the best trees considering the actual data, thus leading to a non-optimal clustering of the observations. However, informal listening tests showed no significant degradation in the output speech quality. We used the same settings for the filter and for the excitation as for Voice A.

Four methods are investigated to apply the prosody adaptation transforms from Voice A to Voice B. These methods differ in the speaking style adaptation technique and in the speaking style transposition approach they rely on.

Speaking style adaptation, from neutral to hypo/hyperarticulated speech, is performed on Voice A in two ways: model adaptation technique (MLLR) and feature adaptation

technique (CMLLR) [22]. In a model adaptation technique, a set of linear transformations is estimated to shift the means and alter the covariances in the source speaker's model so that each state in the HMM system is more likely to generate the adaptation data. In a feature adaptation technique, a set of linear transformations is estimated to modify the feature vectors in the source speaker's model so that each state in the HMM system is more likely to generate the adaptation data.

Speaking style transposition is performed on Voice B, by applying the prosody adaptation transforms obtained on Voice A during the speaking style adaptation step. Since we know the mapping between each pdf and each transformation matrix on Voice A, only the mapping information between each pdf of Voice A and each pdf of Voice B is missing. Again, two techniques are investigated in this work: phonetic mapping and acoustic mapping. The phonetic mapping represents the mapping between each pdf of Voice A and Voice B using decision trees only (as each pdf is associated with a full context label). The acoustic mapping is inspired by the cross-lingual speaker adaptation domain [23]. Here, the matching between each pdf of Voice A and Voice B is computed by finding a leaf correspondence using the Kullback-Leibler divergence between two distributions.

The four methods chosen to apply the prosody adaptation transforms of Voice A to Voice B results from a mix of the two speaking style adaptation techniques and the two speaking style transposition techniques. Table 1 summarizes these four methods.

Table 1. *Methods for applying the adaptation transforms from Voice A to Voice B.*

		Decision trees mapping	
		Phonetic	Kullback-Leibler
Transforms	Model adaptation	Mod_Phn	Mod_KL
	Feature adaptation	Feat_Phn	Feat_KL

5. EXPERIMENTS

As no target data (hypo/hyperarticulated) are available for Voice B, objective measurements cannot be used. Instead, a subjective assessment is performed: a Comparative Mean Opinion Score (CMOS) test, to assess the quality of the prosody model transposition (Section 5.1). The CMOS evaluation is then complemented with a Comparative Perception of the Degree of Articulation (CPDA) test, to quantify the (positive or negative) effects of prosody model transposition on the perception of the degree of articulation (Section 5.2).

The baseline system was chosen to be the neutral full data model of Voice B, where a straightforward phone-independent constant ratio is applied to decrease/increase pitch and phone durations to sound like hypo/hyperarticulated speech. This ratio is computed once for all over the Voice A hypo/hyperarticulated databases (see Section 2) by adapting the mean values of the pitch and phone duration from the neutral style. The same acoustic model for spectral features as for our proposed synthesizers (see Section 4) has been used for the baseline system. Finally, the phonetic transcription is manually adjusted to fit the real hypo/hyperarticulated transcription.

It is important to note that the baseline is here assumed to be a reference with a high segmental quality as it is based on a full data model without any statistical post-processing. Nonetheless, although it makes sense to apply this baseline technique on prosody features, such a straightforward ratio approach is obviously not generalizable to the filter coefficients. This then motivates the need in finding what is the most appropriate statistical adaptation method amongst the techniques presented in Section 4.

5.1. Speech Quality of the Prosody Model Transposition

For this evaluation, listeners were asked to compare two sentences of several pairs (X, Y) from the overall speech quality point of view: i) the sentence synthesized by one of the four methods described above; ii) the corresponding sentence synthesized by the baseline system. These two sentences were randomly presented as either X or Y all along the test. The CMOS values range on a 7-point gradual scale varying from 3 (meaning that X is much better than Y) to -3 (meaning the opposite). A score of 0 is given if both versions are found to be about the same.

Each listener was presented with 24 pairs, randomly chosen from the test set, corresponding to each degree of articulation and to each method. During the test, listeners were allowed to listen to each pair of sentences as many times as wanted, in the order they preferred. However they were not allowed to come back to previous sentences after validating their decision. 29 people, mainly naive listeners, participated to this evaluation. The mean CMOS score, for each method and each degree of articulation, is displayed in Figure 1. A positive (negative) score means that the considered method gives better (worse) results than the baseline. A score of around 0 means that the considered method and the baseline are found to provide an equivalent quality.

Figure 1 shows that method *Mod_Phn* interestingly achieves the best performance, for both hypo and hyperarticulated speech. It provides results significantly similar to the baseline which was considered as a golden reference in terms of segmental quality.

For hyperarticulated speech, the use of Kullback-Leibler divergence (instead of the phonetic mapping) for speaking style transposition (method *Mod_KL* vs. method *Mod_Phn*) leads to a dramatic drop in performance (the mean value of the distributions drops from -0.01 to -1.77). The major prob-

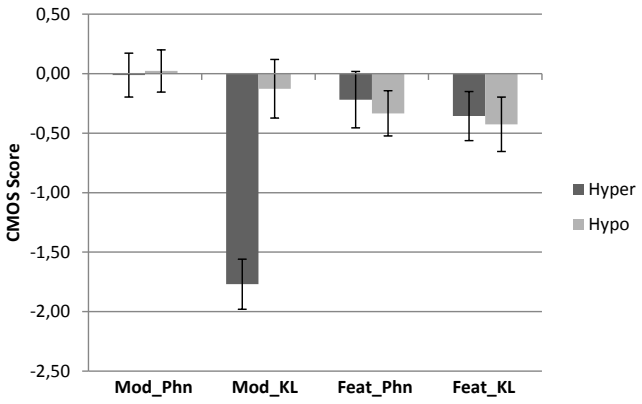

Fig. 1. CMOS Test - Mean CMOS score for each method and each degree of articulation, together with their 95% confidence intervals.

lem felt by the listeners comes from the poor supra-segmental quality (mainly due to pitch errors). This implies that the knowledge of the phonetic environment is essential for estimating the transforms, and that the acoustic information is not sufficient. Compared to MLLR adaptation (methods *Mod_Phn* and *Mod_KL*), CMLLR adaptation (methods *Feat_Phn* and *Feat_KL*) gives intermediate results, further to the baseline than method *Mod_Phn* but closer to the baseline than method *Mod_KL*. Again, using the Kullback-Leibler divergence (method *Feat_KL* vs. method *Feat_Phn*) leads here to a decrease in performance, minor in this case: mean values decrease from -0.22 (method *Feat_Phn*) to -0.36 (method *Feat_KL*).

For hypoarticulated speech, the overall performance of MLLR adaptation (methods *Mod_Phn* and *Mod_KL*, with mean values of 0.02 and -0.13 respectively) is higher than the CMLLR adaptation one (methods *Feat_Phn* and *Feat_KL*, with mean values of -0.33 and -0.43 respectively). The same conclusion as for hyperarticulated speech can be drawn for hypoarticulated speech, i.e. that the use of Kullback-Leibler divergence (instead of the phonetic mapping) for speaking style transposition leads to a slight reduction in performance.

We can also see on Figure 1 that the confidence intervals overlap between methods *Mod_Phn* and *Mod_KL* (hypoarticulated) and between methods *Mod_Phn* and *Feat_Phn* (hyperarticulated). It is however not the case between methods *Mod_Phn* and *Mod_KL* (hyperarticulated) and between methods *Mod_Phn* and *Feat_Phn* (hypoarticulated), confirming the best performance is achieved by method *Mod_Phn*.

Figure 2 displays the detailed preference scores for each method and each degree of articulation. It is again noticed that method *Mod_Phn* is the best technique, leading to preference scores equivalent (or even slighlty better in hyperarticulated speech) to the baseline which was considered to be a golden reference at this level.

5.2. Perception of the Degree of Articulation

The CMOS test performed in Section 5.1 provided useful information about the synthetic speech quality that could be obtained when applying the adaptation transforms obtained on Voice A to modify the prosody of Voice B. However it does not bring anything about the effective production of the desired degree of articulation. This is why we here complement the results of the CMOS evaluation with a CPDA test.

Listeners were given two pairs of sentences. The first pair was composed of: i) the neutral sentence synthesized by the Voice A full data model; ii) the hypo or hyperarticulated sentence (randomly shuffled) synthesized by the Voice A adapted hypo or hyperarticulated model. The second pair was composed of: i) the neutral sentence synthesized by the Voice B full data model; ii) the hypo or hyperarticulated sentence (same degree of articulation as the second sentence of the first pair) synthesized by one of the four methods or the baseline investigated in this work (see Section 4).

Listeners were also given a continuous scale ranging from 0 to 1. This scale was extended one point further on both sides (ranging therefore from -0.2 to 1.2) in order to prevent border effects. The neutral sentence synthesized by the Voice A full data model was set to 0, while the Voice A hypo or hyperarticulated sentence was set to 1. Given the distance between the sentences composing the first pair, listeners were asked to estimate perceptually the distance between the two sentences of the second pair (the neutral sentence synthesized by the Voice B full data model, and the hypo or hyperarticulated sentence synthesized using one of the four methods or the baseline). This corresponds to the extent to which the degree of articulation of Voice A is reproduced on Voice B.

The test consisted of 10 quadruplets. For each degree of articulation and for each method to be tested, 30 sentences were randomly chosen from the test set. The same listening protocol as in Section 5.1 was implemented. 24 people, mainly naive listeners, participated to this evaluation. Figure 3 shows the mean score, corresponding to the distance between the neutral sentence synthesized by the Voice B full data model and the hypo/hyperarticulated sentence synthesized by one of the four methods or the baseline, 1 being the target value corresponding to the reference degree of articulation (defined on Voice A).

It can be observed from Figure 3 that all methods outperform the baseline regarding the reproduction of hyperarticulated speech (with significant differences for methods *Mod_Phn* and *Feat_Phn*). On the contrary, a slight advantage is noticed in favor of the baseline for the rendering of hypoarticulated speech although there are no statistically significant differences. For this latter speaking style, methods *Mod_Phn*, *Mod_KL* and *Feat_KL* turn out to provide equivalent results, while method *Feat_Phn* performs slightly worse.

The size of 95% confidence intervals in this test is rather important, particularly for hypoarticulated speech. This could

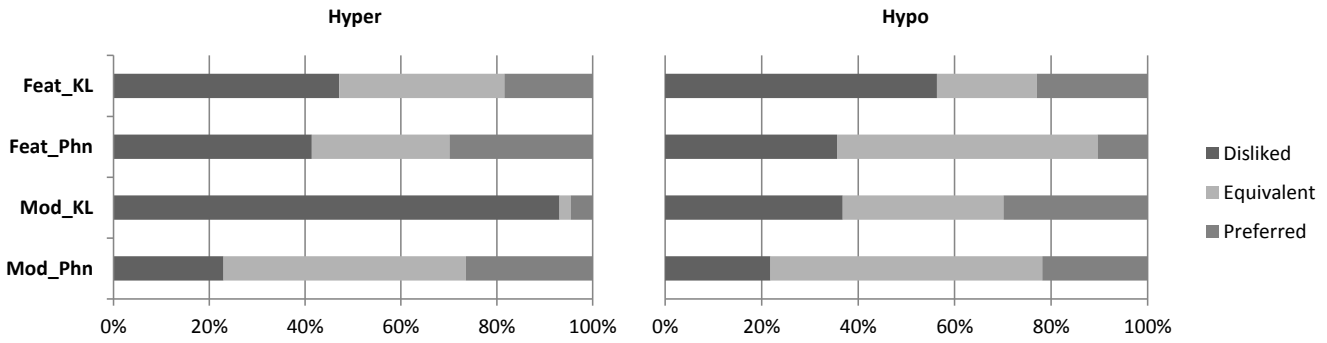

Fig. 2. CMOS Test - Detailed CMOS scores, expressed in [%], for hyperarticulated speech (left) and for hypoarticulated speech (right).

be explained by the difficulty of this evaluation. Indeed it is not easy to compare the two perceptual distances between the two sentence pairs, as the mean pitch of voice A and voice B is different. However, this test should be taken as an indication that overall the prosody modification is well reproduced after the application on Voice B of the adaptation transforms learned on Voice A.

The fact that the proposed methods are observed to outperform the baseline for the production of hyperarticulation is of interest in several applications where it is aimed at increasing the intelligibility of the synthetic voice (while keeping an equivalent naturalness). For example, hyperarticulated speech has been shown in [6] to enhance the comprehension of synthetic speech in a degraded environment (car noise and reverberant conditions).

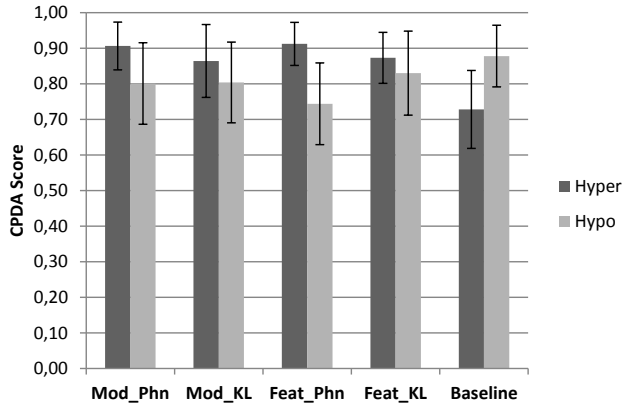

Fig. 3. CPDA Test - Mean score corresponding to the distance between the neutral sentence synthesized by the Voice B full data model and the hypo/hyperarticulated sentence synthesized by one of the four methods or the baseline (1 being the reference degree of articulation, defined on Voice A), together with their 95% confidence intervals.

6. CONCLUSIONS AND FUTURE WORKS

This paper addressed the automatic modification of the degree of articulation of an existing standard neutral voice (Voice B) for which no hypo/hyperarticulated recordings were available. We focused on prosody modifications, although the statistical techniques considered in this work can be applied to any stream of parameters exhibiting appropriate interpolability properties. In a first step, we performed on a voice for which data with different degrees of articulation are available (Voice A), the adaptation of a neutral synthesizer to generate hypo and hyperarticulated speech. Then we investigated several methods for the application of the adaptation transforms computed on Voice A to Voice B. Four methods were considered, differing in the speaking style adaptation and transposition techniques they rely on. A subjective evaluation showed that the method combining Maximum Likelihood Linear Regression (MLLR) adaptation and phonetic mapping between the decision trees of Voice A and Voice B clearly led to the most interesting results. This technique indeed outperformed other presented approaches and achieved a segmental quality comparable to the golden reference baseline. It also proved to be efficient for the reproduction of hyperarticulation, which is of interest for speech synthesis applications where an intelligibility enhancement is required.

As future works, we plan to investigate which filter coefficients are the most suited for statistical transformations: Line Spectral Pairs coefficients (LSP), PARtial CORrelation coefficients (PARCOR) and Log Area Ratio coefficients (LAR). The conclusions drawn in this paper could then be applied for integrating the adaptation of the filter in complement to the prosody and extending in this way the present study.

7. ACKNOWLEDGEMENTS

Benjamin Picart is supported by the “Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (FRIA). Thomas Drugman is supported by the Walloon Re-

gion, SPORTIC project #1017095.

8. REFERENCES

[1] B. Lindblom, *Economy of Speech Gestures*, Spinger-Verlag, New-York, 1983.

[2] G. Beller, *Analyse et Modèle Génératif de l'Expressivité - Application à la Parole et à l'Interprétation Musicale*, Ph.D. thesis, Université Paris VI - Pierre et Marie Curie, IRCAM, 2009.

[3] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo and hyperarticulated speech," in *Speech Synthesis Workshop 7 (SSW7)*, Kyoto, Japan, 2010, pp. 270–275.

[4] B. Picart, T. Drugman, and T. Dutoit, "Continuous control of the degree of articulation in hmm-based speech synthesis," in *Interspeech*, Firenze, Italy, 2011, pp. 1797–1800.

[5] B. Picart, T. Drugman, and T. Dutoit, "Perceptual effects of the degree of articulation in hmm-based speech synthesis," in *NOLISP Workshop*, Las Palmas, Gran Canaria, 2011, pp. 177–182.

[6] B. Picart, T. Drugman, and T. Dutoit, "Assessing the intelligibility and quality of hmm-based speech synthesis with a variable degree of articulation," in *The Listening Talker (LISTA) workshop*, Edinburgh, Scotland, May 2-3 2012.

[7] G. Beller, N. Obin, and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech," in *Fourth International Conference on Speech Prosody*, Campinas, Brazil, 2008.

[8] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Audio, Speech, & Language Processing*, vol. 17, no. 6, pp. 1208–1230, August 2009.

[9] T. Nose, M. Tachibana, and T. Kobayashi, "Hmm-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Transactions on Information and Systems*, vol. 92, no. 3, pp. 489–497, 2009.

[10] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Interspeech*, Pittsburgh, USA, September 2006, pp. 2446–2449.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Transactions on Information and Systems*, vol. 93, no. 6, pp. 1589–1598, June 2010.

[12] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[13] H. Ye and S. Young, "High quality voice morphing," in *Proc. IEEE ICASSP, Montreal*, Montreal, Canada, 2004.

[14] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proceedings of International Conference on Spoken Language Processing*, September 1994, vol. 3, pp. 1043–1046.

[15] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[16] http://hts.sp.nitech.ac.jp/, "[online] hmm-based speech synthesis system (hts)," .

[17] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 968–981, March 2012.

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

[19] J. Ferguson, "Variable duration models for speech," in *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[20] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Eurospeech*, September 1999, pp. 2347–2350.

[21] J. Yamagishi, *Average-Voice-Based Speech Synthesis*, Ph.D. thesis, Tokyo Institute of Technology, March 2006.

[22] 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 Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, January 2009.

[23] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for hmm-based speech synthesis," in *IEEE ICASSP*, Dallas, Texas, 2010.