Formant-Controlled Speech Synthesis Using Hidden Trajectory Model

Ming-Qi Cai, Zhen-Hua Ling, Li-Rong Dai

National Engineering Laboratory for Speech and Language Processing,
University of Science and Technology of China, P.R.China
mqcai@mail.ustc.edu.cn, {zhling,lrdai}@ustc.edu.cn

Abstract

This paper presents a statistical parametric speech synthesis method using hidden trajectory model (HTM) for flexibly controlling the formant positions and bandwidths of synthetic speech. In an HTM, hidden formant trajectories are generated by a bidirectional filtering process on the time-aligned and phone-dependent formant targets. The observed cepstral features are constituted by a formant-related component, which is predicted from the hidden formant trajectories using a nonlinear and analytical function, and a residual component, which is modeled by context-dependent Gaussians. In this paper, we apply HTM-based acoustic modeling to speech synthesis. The distribution parameters of the formant targets are manipulated at synthesis time to control the characteristics of synthetic speech. In our implementation, the distributions of residual cepstra are estimated for each quinphone and the question set used in the decision-tree-based model clustering is tailored so as to acquire high controllability for vowels. Experimental results shows that this proposed method can achieve effective controllability on the formant positions and bandwidths while keeping almost the same naturalness as the conventional HMM-based approach.
Index Terms: speech synthesis, hidden trajectory model, hidden Markove model, formant targets

1. Introduction

Hidden Markov model (HMM) based speech synthesis has become a mainstream speech synthesis method in recent years [1]. In this method, spectrum, F0 and state duration are modeled simultaneously in a unified framework of HMM. At synthesis time, a sentence HMM is first constructed by concatenating the HMMs of context-dependent phones according to text analysis results. Acoustic parameters are generated from the sentence HMM using the maximum likelihood parameter generation (MLPG) algorithm [2], and then sent to a vocoder to construct waveforms. This method is able to synthesize highly intelligible and smooth speech sounds. However, its flexibility to control the characteristics of synthetic speech is constrained by the nature of the training data available [3].

A method of improving the controllability of HMM-based speech synthesis by integrating articulatory features has been proposed in [3, 4, 5], where the articulatory movements captured using electromagnetic articulography (EMA) were treated as auxiliary features to decide the distribution of acoustic features at each HMM state. This method achieved controllability on synthetic speech by manipulating the articulatory features according to phonetic knowledge at synthesis time. However, recording EMA data needs precise and expensive equipments and the speaker's pronunciation is always influenced by the sensors pasted in his mouth. Formant features, including the central frequencies and the bandwidths of formants which can be conveniently calculated from speech waveforms, also have a straightforward relationship to the shape of vocal tract and the movement of articulators. They have been used as intermediate representations to achieve flexible speech synthesis in [6], where the joint distribution of the acoustic features and formant features were modeled by multi-stream HMMs and the dependency between these two kinds of features was described by a piecewise linear transform. Although this formant-controlled HMM-based speech synthesis method can manipulate the predicted formant features to control the pronunciation of vowels effectively, it has several limitations. First, the piecewise linear transform is inconsistent with the nonlinear transform relationship between the formant features and the observed acoustic features, such as cepstra and LSPs. Second, the formant features are constrained to be generated from the HMMs at synthesis time, which makes the integration of phonetic knowledge into formant prediction somewhat inconvenient.

On the other hand, a multi-stage statistical generative model named hidden trajectory model (HTM) has been proposed to describe the speech dynamics and the hierarchical speech production process for speech recognition application [7, 8, 9]. In this model, vocal tract resonance (VTR, i.e. formant) trajectories which are generated by a bidirectional filtering process on VTR targets are treated as hidden variables between phonological descriptors and acoustic observations. The mapping from VTRs to acoustic features is achieved by a nonlinear prediction with context-dependent residuals. In this paper, we apply this model structure to the formant-controlled speech synthesis so as to solve the limitations of the existing approach [6] mentioned above. First, the nonlinear and analytical mapping relationship between formant features and cepstra is adopted, which is more accuracy and robust than the piecewise linear transform used in [6]. Second, the hidden formant trajectories are generated from stochastic formant targets. The distributions of the formant targets are estimated for each monophone and the coarticulation effects are described by the dynamic filtering process. Therefore, the statistical model of formant features is much more compact and easier for manipulation than the HMMs.

This paper is organized as follows. Section 2 briefly reviews the model structure of HTM and describes the proposed formant-controllable speech synthesis method using HTM. The experimental results and the conclusions are given in Section 3 and 4 respectively.

2. Methods

2.1. Acoustic modeling using HTM with formant targets

An HTM is a structured generative model, in which the hidden formant trajectories generated by target filtering are used as a intermediate level between phonetic specifications and acoustic

observations [7]. In an HTM, each phone is associated with a set of formant targets, whose distribution is assumed to be a Gaussian

$$p(\boldsymbol{t}(k)|s(k)) = \mathcal{N}(\boldsymbol{t}(k); \boldsymbol{\mu}_{T_{s(k)}}, \boldsymbol{\Sigma}_{T_{s(k)}}), \quad (1)$$

where $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $s(k)$ is the monophone index at frame k; $\boldsymbol{t}(k) = [t_{f,1}(k), t_{b,1}(k), ..., t_{f,P}(k), t_{b,P}(k)]^\top$ is the vector of formant targets at frame k; $(\cdot)^\top$ means the matrix transpose; P is the order of formants; $t_{f,p}(k)$ and $t_{b,p}(k)$ denote the central frequency and bandwidth of the p-th formant target respectively. Then, the distribution of the observed cepstra can be derived by two stages.

At the first stage, a low-pass finite impulse response (FIR) filter is used to obtain formant features $\boldsymbol{z}(k)$ from the time-aligned formant target sequence $\boldsymbol{t}(k)$ by convolution as

$$\boldsymbol{z}(k) = h(k) * \boldsymbol{t}(k), \quad (2)$$

where $h(k)$ is the impulse response of the FIR filter; $\boldsymbol{z}(k) = [f_1(k), b_1(k), ..., f_P(k), b_P(k)]^\top$; $f_p(k)$ and $b_p(k)$ denote the central frequency and bandwidth of the p-th formant at frame k respectively. $h(k)$ can be simply defined as

$$h(k) = \begin{cases} C\gamma^{-k} & -D \le k \le 0, \\ C & k = 0, \\ C\gamma^{k} & 0 \le k \le D, \end{cases} \quad (3)$$

where γ is the stiffness parameter specifying the degree of articulation, which is positive and real-valued, ranging from zero to one; D is the unidirectional length of the impulse response, which represents the temporal extent of coarticulation; C is a normalization constant, ensuring that $h(k)$ sums to one over all time frames. This filtering process describes coarticulation effects because the generation of the current phone's formant trajectories is influenced by the adjacent phones' targets. Considering the linearity between $\boldsymbol{z}(k)$ and $\boldsymbol{t}(k)$ as shown in (2), the distribution of the formant features $\boldsymbol{z}(k)$ is a Gaussian

$$p(\boldsymbol{z}(k)|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{z}(k); \boldsymbol{\mu}_{z(k)}, \boldsymbol{\Sigma}_{z(k)}), \quad (4)$$

where $\boldsymbol{s}$ denotes the sequence of phone indices;

$$\boldsymbol{\mu}_{z(k)} = \sum_{\tau=k-D}^{k+D} C\gamma^{|k-\tau|}\boldsymbol{\mu}_{T_{s(\tau)}}, \quad (5)$$

$$\boldsymbol{\Sigma}_{z(k)} = \sum_{\tau=k-D}^{k+D} C^2\gamma^{2|k-\tau|}\boldsymbol{\Sigma}_{T_{s(\tau)}}. \quad (6)$$

At the second stage, the formant features are mapped to cepstra by

$$\boldsymbol{o}(k) = \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{r}(k), \quad (7)$$

where $\boldsymbol{o}(k)$ is the vector of cepstra at frame k; $\mathcal{F}$ is a nonlinear function predicting cepstra from formant features; $\boldsymbol{r}(k)$ is the residual of the prediction. The q-th order cepstrum predicted from $\boldsymbol{z}(k)$ can be written analytically as

$$\mathcal{F}[\boldsymbol{z}(k)]_q = \frac{2}{q}\sum_{p=1}^{P} e^{-\pi q \frac{b_p(k)}{f_{samp}}} \cos(2\pi q \frac{f_p(k)}{f_{samp}}), \quad (8)$$

where f_{samp} is the sampling frequency. The distribution of the residual cepstra $\boldsymbol{r}(k)$ is described using a context-dependent Gaussian $\mathcal{N}(\boldsymbol{r}(k); \boldsymbol{\mu}_{r_{q(k)}}, \boldsymbol{\Sigma}_{r_{q(k)}})$, where $q(k)$ denotes the context description of the segment which frame k belongs to. One choice for implementation is to use HMM states as the segments and use monophone labels as the context description [9].

When the phone sequences and context descriptions of an utterance are known, the distribution of $\boldsymbol{o}(k)$ can be derived by marginalizing the hidden formant features $\boldsymbol{z}(k)$ as

$$p(\boldsymbol{o}(k)|\boldsymbol{s}, \boldsymbol{q}) = \int p(\boldsymbol{o}(k)|\boldsymbol{z}(k), \boldsymbol{q}(k))p(\boldsymbol{z}(k)|\boldsymbol{s})d\boldsymbol{z}, \quad (9)$$

where $p(\boldsymbol{z}(k)|\boldsymbol{s})$ is defined in (4-6); $p(\boldsymbol{o}(k)|\boldsymbol{z}(k), \boldsymbol{q}(k))$ can be derived from (7). In order to simplify the calculation of $p(\boldsymbol{o}(k)|\boldsymbol{z}(k), q(k))$, an approximation using first-order Taylor series is applied to the nonlinear prediction function

$$\mathcal{F}[\boldsymbol{z}(k)] \approx \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)](\boldsymbol{z}(k) - \boldsymbol{z}_0(k)) \quad (10)$$

where $\boldsymbol{z}_0(k)$ is the expansion point of Taylor series. Finally, we can get

$$p(\boldsymbol{o}(k)|\boldsymbol{s}, \boldsymbol{q}) \approx \mathcal{N}\{\boldsymbol{o}(k); \boldsymbol{\mu}_{\boldsymbol{o}(k)}, \boldsymbol{\Sigma}_{\boldsymbol{o}(k)}\} \quad (11)$$

where

$$\boldsymbol{\mu}_{\boldsymbol{o}(k)} = \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)][\boldsymbol{\mu}_{\boldsymbol{z}(k)} - \boldsymbol{z}_0(k)] + \boldsymbol{\mu}_{\boldsymbol{r}_q(k)}, \quad (12)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{o}(k)} = \boldsymbol{\Sigma}_{\boldsymbol{r}_q(k)} + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_{\boldsymbol{z}(k)}(\mathcal{F}'[\boldsymbol{z}_0(k)])^\top. \quad (13)$$

The model parameters of an HTM are composed of $\{\boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}\}$ which describe the distribution of formant targets for each phone and $\{\boldsymbol{\mu}_{r_q}, \boldsymbol{\Sigma}_{r_q}\}$ which describe the distribution of residual cepstra for each possible context description. Once the phone and segment boundaries of training data are available, these model parameter can be estimated under maximum likelihood criterion by gradient descent. The detailed formulae of model estimation can be found in [7, 9].

2.2. HTM-based speech synthesis

The HTM-based acoustic modeling method presented above has been applied to speech recognition successfully [8, 9]. In this paper, we apply HTM to statistical parametric speech synthesis. HTM simulates the hierarchical speech production process and consists of a compact model for formant generation. Therefore, it is suitable for formant-controlled speech synthesis. Fig. 1 shows the flowchart of the proposed HTM-based speech synthesis method, which consists of a training stage and a synthesis stage. The details are as follows.

2.2.1. Model training

As shown in Fig. 1, context-dependent HMMs with decision-tree-based model clustering are first trained in a way similar to the conventional HMM-based speech synthesis. Component phones, i.e. affricates and diphthongs, are decomposed into two sub-phones before model training because each phone is assumed to have static formant targets in an HTM as shown in (1). The trained HMMs can provide discrete state boundaries for HTM estimation and make F0 and duration prediction at synthesis time.

In an HTM, the distributions of formant targets $\{\boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}\}$ are specified by monophone labels and the distributions of residual cepstra $\{\boldsymbol{\mu}_{r_q}, \boldsymbol{\Sigma}_{r_q}\}$ are context-dependent. In our implementation, HMM states are used as the segments and quinphone descriptions are used as the context features for modeling residual cepstra. Because the possible

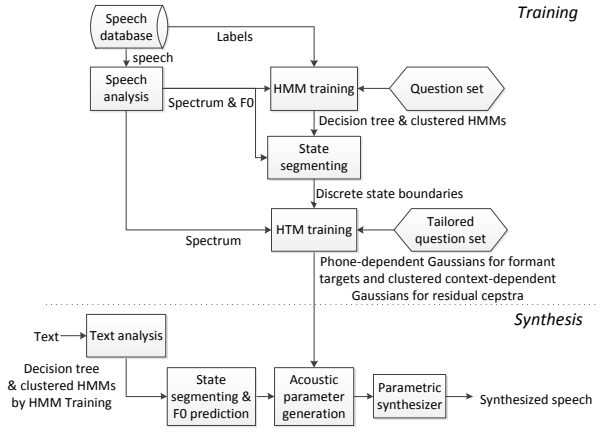

Figure 1: *The flowchart of HTM-based speech synthesis method.*

quinphone combinations are so many, a decision-tree-based model clustering technique [10] is applied to $\{\boldsymbol{\mu}_{r_q}, \boldsymbol{\Sigma}_{r_q}\}$ so as to solve the data-sparsity problem. Furthermore, the vowel related questions are removed from the question set for decision tree construction in order to achieve better formant controllability over vowels and to avoid the potential conflict between the manipulated formant features and the context features at synthesis time [11].

During HTM training, the formant target distributions $\{\boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}\}$ and the residual cepstrum distributions $\{\boldsymbol{\mu}_{r_q}, \boldsymbol{\Sigma}_{r_q}\}$ are updated iteratively to maximize the likelihood of the acoustic observations. Thus initial parameters are necessary in this iterative updating. The formant features of the training database are first extracted using Snack Sound Toolkit [12]. The means and variances of the extracted formant features are calculated for each monophone to initialize $\{\boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}\}$. Then, the formant features recovered from the initial $\boldsymbol{\mu}_{T_s}$ by FIR filtering are mapped to cepstra using (8). These cepstra are subtracted from the observations to get initial residual cepstra. Given discrete-state boundaries and the initial residual cepestra, context-dependent Gaussians are estimated and clustered using the tailored question set to initialize $\{\boldsymbol{\mu}_{r_q}, \boldsymbol{\Sigma}_{r_q}\}$. As mentioned in (10), an approximation using first-order Taylor series is applied to simply the calculation of the acoustic likelihood. The expansion point is set at $\boldsymbol{z}_0(k) = \boldsymbol{\mu}_z(k)$ in our implementation. Because the mean vectors of the formant target distributions are the most important parameters in formant-controlled speech synthesis, they are first updated while keeping the other model parameters constant after the initialization. Then the variances of formant targets and the residual cepstrum distributions are updated in turn within one iteration. The iterative updating stops when the increase of likelihood on training data becomes insignificant.

2.2.2. Parameter generation and formant control

To perform synthesis, the result of text analysis is first used to determine the discrete-state boundaries and predict the F0 features from the standard HMM-based system. Then the distribution of cepstra at each frame can be derived from the estimated HTM using (11). In our implementation, cepstral features are composed of static, velocity and acceleration components. Therefore, MLPG algorithm considering the constraints of dynamic features is applied to generate the cepstral sequence [2]. Then the generated F0s and cepstra are sent to a vocoder to reconstruct waveforms [13]. Because the formant targets are low-dimensional and phonetically meaningful, it is convenient to manipulate their distribution parameters according to phonetic knowledge. In the experiments of this paper, only the mean vectors of the formant target distributions are manipulated. Such manipulation can be reflected in the distribution of acoustic features effectively according to (9) and can finally affect the formant characteristics of the synthetic speech. It is more straightforward than the approach proposed in [6], in which to predict explicit formant features from HMMs is necessary and to manipulate the formant sequences directly is tricky.

Table 1: *Cepstral distortions(dB) of cepstral prediction on the test set for HMM-S and HTM-S .*

	frames	HMM-S	HTM-S
vowels+consonants	37921	2.9865	3.0415
vowels	16146	2.9199	3.1138
consonants	21775	3.0349	2.9867

3. Experiments

3.1. Experimental conditions

A male British English speech database was used in our experiments [14], including 1,199 sentences for training and 63 sentences for test. The waveforms were in 16kHz PCM format with 16-bit precision. The acoustic features used for standard HMM-based system training included F0 and spectral parameters, which were 23-order cepstras and an extra gain dimension derived by STRAIGHT analysis [13]. A total of 62 monophones were used in our system. The formant features were extracted by the Snack Sound Toolkit with the frame length of 5ms which matched the acoustic features. The formant features consisted of the first three central frequencies and corresponding bandwidths considering that they can characterize vowels well [15]. Then, a conventional HMM-based speech synthesis system and an HTM-based speech synthesis system using the methods described in Section 2.2 were constructed. They are denoted as HMM-S and HTM-S respectively.

3.2. Naturalness evaluation

Our proposed method aims at improving the flexibility of conventional HMM-based speech synthesis without degrading naturalness. First, the cepstral distortion of the cepstra generated by the two systems were calculated. The results on the test set are compared in Table 1. HMM-S and HTM-S achieve similar overall cepstral distortions. For consonants, the cepstral distortion of HTM-S is smaller than that of HMM-S. This is because the vowel related questions were removed from the question set used to cluster the distributions of residual cesptra in HTM-S and there are more consonant related distributions in HTM-S than that in HMM-S. Furthermore, a subjective listening test was conducted to compare the naturalness of speech synthesized by HTM-S and HMM-S. Ten Chinese students well educated in English were asked to take part in a preference test which contained twenty pairs of sentences synthesized by these two systems using parameter generation considering global variance [16]. Fig. 2 shows the average preference scores. There is no significant preference between the naturalness of these two systems.

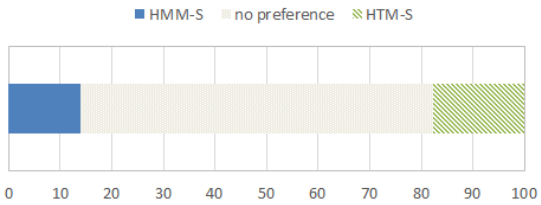

Figure 2: *Naturalness preference scores between HMM-S and HTM-S.*

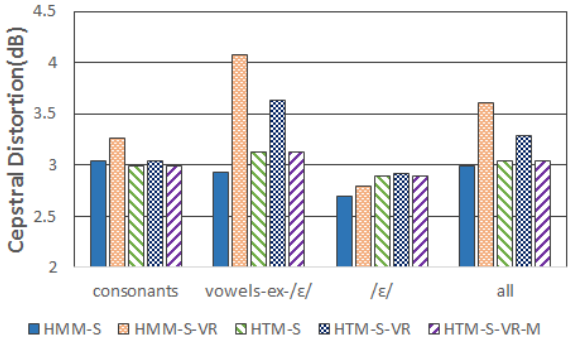

Figure 3: *Cepstral distortions for different systems and different types of phones in the vowel identity modification experiment, where "vowels-ex-/ɛ/" indicates all vowels excluding the vowel /ɛ/. HMM-S and HTM-S refer to the results without vowel replacement. HMM-S-VR and HTM-S-VR denotes the results after vowel replacement. HTM-S-VR-M indicates the results using HTM-S system with vowel replacement and formant position manipulation.*

3.3. Formant control on central frequencies

A vowel modification experiment similar to the one conducted in [11] was carried out to evaluate the effectiveness of controlling formant central frequencies using our proposed method. In this experiment, each test sentence was first subjected to standard front-end text analysis, then all vowels were replaced with the vowel /ɛ/. Obviously, the sentences synthesized by HMM-S contained no vowels other than /ɛ/. When synthesizing these sentences using HTM-S, we modified the instances of vowel /ɛ/ to different target vowels by manipulating the dimensions of $\boldsymbol{\mu}_{T_s}$ which corresponded to formant central frequencies of /ɛ/ to those of target vowel for each instance. The resulting cepstra distortions after vowel replacement and formant position modification for different types of phones are shown in Fig. 3. The cepstral distortion of HMM-S for all phones increases from 2.99dB to 3.60dB after vowel replacement due to a significant distortion increment for the vowels excluding /ɛ/. For HTM-S, the overall cepstral distortion is 3.04dB after vowel replacement and formant modification, which is comparable with HMM-S using correct vowel identifiers. These results show that our proposed method can achieve effective control on vowel identity by modifying the central frequency of formants.

3.4. Formant control on resonant bandwidths

From the nonlinear function denoted by (8), we can see that the generation of cepstra are also influenced by formant bandwidths. Bandwidths describe the sharpness of each formant and consequently influence the quality of synthetic speech. An experiment was conducted to multiply each dimension of estimated $\boldsymbol{\mu}_{T_s}$ which corresponds to formant bandwidths with a factor of 0.5 or 2. Then, two groups of preference test were carried out to compared the naturalness of speech synthesized after bandwidth modification with the original ones to investigate the effects of bandwidth modification. Twenty sentence pairs were used for each preference test. Ten Chinese students were asked to take part in this subjective evaluation. As shown in Fig. 4, the naturalness of synthetic speech increases with half bandwidths and decreases with double bandwidths. This is consistent with the knowledge that increasing the sharpness of formants may alleviate the over-smoothing effect caused by statistical parametric speech synthesis. Samples of the synthetic speech used in the experiments can be found at http://home.ustc.edu.cn/ mqcai/Demos4Interspeech2014.htm.

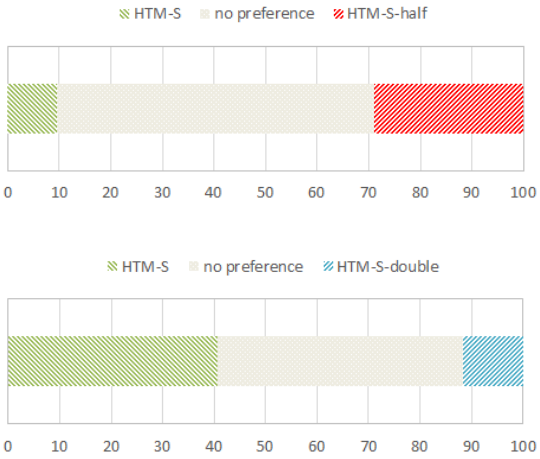

Figure 4: *Naturalness preference scores between original HTM-S, HTM-S with half formant bandwidths (HTM-S-half), and HTM-S with double formant bandwidths (HTM-S-double).*

4. Conclusions

This paper has proposed a novel framework that a hidden trajectory model with the distributions for phone-dependent formant targets is utilized for formant-controlled speech synthesis. The experimental results presented in this paper have shown that the proposed method can achieve similar performance with the conventional HMM-based synthesis method in naturalness. The vowel identity modification experiment and the formant bandwidth modification experiment have demonstrated the effectiveness of our proposed method in controlling the characteristic of synthetic speech by modifying the distribution parameters of the low-dimensional formant targets. The experimental results shown in this paper are still preliminary. To conduct more subjective evaluations and to apply this model to speaker adaptation application will be the tasks of our future work.

5. Acknowledgements

This work was partially funded by the National Natural Science Foundation of China (Grant No. 61273032).

6. References

[1] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.,"Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. Eurospeech, pp.2347-2350, 1999.

[2] Tokuda, K., Kobayashi, T. and Imai, S., "Speech parameter generation from HMM using dynamic features", in Proc. of ICASSP, vol. 1, pp. 660-663, 1995.

[3] Ling, Z.-H., Richmond, K., Yamagishi, J. and Wang, R.-H., "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis", Trans. Audio, Speech and Lang. Proc., vol. 17, pp. 1171-1185, 2009.

[4] Ling, Z.-H., Richmond, K. and Yamagishi, J., "Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis", Proc. of Interspeech, 2012.

[5] Cai, M.-Q., Ling, Z.-H. and Dai, L.-R., "Target-filtering model based articulatory movement prediction for articulatory control of HMM-based speech synthesis", in Proc. 11^{th} Int. Conf. Signal Process., pp. 605-608, 2012.

[6] Lei, M., Yamagishi, J., Richmond K., Ling, Z.-H., King, S. and Dai, L.-R., "Formant-controlled HMM-based Speech Synthesis", Proc. of Interspeech, pp. 2777-2780, 2011.

[7] Deng, L., Yu, D. and Acero, A., "Learning statistically characterized resonance targets in a hidden trajectory model of speech coarticulation and reduction", Proc. of Interspeech, pp. 1097-1100, 2005.

[8] Deng, L., Yu, D. and Acero, A., "A bi-directional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition", Trans. Audio, Speech and Lang. Proc., vol. 14, , pp. 256-265, 2006.

[9] Deng, L., "Dynamic Speech Models, Theory, Algorithms, and Applications", Chapt. 5, 2006.

[10] Odell, J., "The use of context in large vocabulary speech recognition", PhD dissertation, Cambridge University, 1995.

[11] Ling, Z.-H., Richmond, K. and Yamagishi, J., "Articulatory Control of HMM-Based Parametric Speech Synthesis Using Feature-Space-Switched Multiple Regression", Trans. Audio, Speech and Lang. Proc., vol. 21, pp. 205-217, 2013.

[12] The Snack Sound Toolkit, Department of Speech, Music and Hearing home, Online: http://www.speech.kth.se/snack/

[13] Kawahara, H., Masuda-Katsuse, I. and Cheveigne, A. de, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, vol. 27, pp. 187-207, 1999.

[14] Richmond, K., Hoole, P. and King, S., "Announcing the electromagnetic articulography(day 1) subset of the mngu0 articulatory corpus", Proc. Interspeech, pp. 1505-1508, 2005.

[15] Acero, A., "Formant analysis and synthesis using hidden Markov models", in Proc. of Eurospeech, 1999.

[16] Toda, T. and Tokuda, K., "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis", IEICE TRANS. INF. & SYST., vol. E90-D, 2007.