Context-dependent additive log F_0 model for HMM-based speech synthesis

Heiga Zen, Norbert Braunschweiler

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

Abstract

This paper proposes a context-dependent additive acoustic modelling technique and its application to logarithmic fundamental frequency (log F_0) modelling for HMM-based speech synthesis. In the proposed technique, mean vectors of state-output distributions are composed as the weighted sum of decision tree-clustered context-dependent bias terms. Its model parameters and decision trees are estimated and built based on the maximum likelihood (ML) criterion. The proposed technique has the potential to capture the additive structure of log F_0 contours. A preliminary experiment using a small database showed that the proposed technique yielded encouraging results.
Index Terms: speech synthesis, HMMs, log F_0 modelling

1. Introduction

Hidden Markov model (HMM)-based speech synthesis [1] has grown in popularity in recent years. In this framework, the spectrum, excitation, and durations of speech are modelled simultaneously in a unified framework of HMMs. For a given text to be synthesized, speech parameter trajectories that maximise their output probabilities are generated from estimated HMMs under constraints between static and dynamic features [2]. Typical instances of this framework use mel-cepstral coefficients or line spectral pairs for their spectral parameters and log F_0 values including unvoiced regions for their excitation parameters. However, as far as we know, these parameters are the final results of speech production and only represent its surface; there are unobservable features and structures in their generation process. For spectral parameters, articulatory features seem to be unobservable features. The articulators determine the resonance characteristics of the vocal tract during the production of speech. Therefore, speech can be characterised by both the acoustics and vocal-apparatus properties. Ling *et al.* recently proposed an HMM-based acoustic and articulatory joint modelling and synthesis technique to construct HMM-based articulatory speech synthesis systems [3]. The state-output vector of HMMs used in this technique includes both acoustic and articulatory features. Acoustic and articulatory features were modelled in individual HMM streams and clustered separately by phonetic decision trees. A piece-wise linear transform was used to represent the dependency between these two feature streams.

It is known that the generation process of log F_0 contours has an additive nature [4]. To capture this additive nature and produce more natural log F_0 trajectories, various additive log F_0 models have been proposed. Some of them were applied to predict log F_0 contours for unit-selection synthesis systems, *e.g.*, [5]. However, the basic HMM cannot directly capture this additive nature due to a lack of capability to extract the underlying additive structure of observations. To add an additive nature to the HMM-based speech synthesis framework, Yao *et al.* proposed the boosting-style additive log F_0 model [6]. Although the additive acoustic model proposed in [7] was also motivated by log F_0 modelling for HMM-based speech synthesis, it has not been available due to a huge amount of computations.

This paper proposes a context-dependent additive acoustic modelling technique and its application to log F_0 modelling for HMM-based speech synthesis. In the proposed technique, mean vectors of state-output distributions are composed as the weighted sum of decision tree-clustered context-dependent additive bias terms. Its model parameters and decision trees are estimated and built based on the ML criterion. The proposed technique has the potential to capture the additive structure of log F_0 contours. A preliminary experiment showed that the proposed technique yielded encouraging results.

The rest of this paper is organised as follows. Section 2 describes the proposed context-dependent additive acoustic modelling technique. Section 3 discusses the application of the proposed technique to log F_0 modelling. Section 4 shows the results of a preliminary experiment. Finally, concluding remarks and future plans are presented in Section 5.

2. Additive acoustic model

2.1. Definition

Like the additive acoustic model described in [7], acoustic features[1] are assumed to be generated as the sum of additive components as

$$o_t = \sum_{i=1}^{P} o_t^{(i)} \tag{1}$$

where o_t is an acoustic feature at time t and $o_t^{(i)}$ denotes its i-th additive component. Unlike [7], here $1, \ldots, P-1$-th additive components are assumed to be bias terms (zero variance Gaussian), and the P-th additive component is generated according to a zero-mean Gaussian distribution as

$$o_t^{(i)} \sim \mathcal{N}(\lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}, 0) = \lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}, \quad 1 \leq i < P \tag{2}$$

$$o_t^{(P)} \sim \mathcal{N}(0, \sigma_{m_t}^2), \tag{3}$$

where m_t denotes a context associated to the t-th frame, $\mu_{m_t}^{(i)}$ and $\lambda_{m_t}^{(i)}$ correspond to the context-dependent bias term and scaling factor of the i-th additive component, and $\sigma_{m_t}^2$ is a variance of Gaussian distribution for the P-th additive component. The output distribution of $\boldsymbol{o}_t$ for the context m_t is given by

$$p(o_t \mid m_t, \Lambda) = \mathcal{N}\left(o_t \,;\, v_{m_t}, \sigma_{m_t}^2\right), \tag{4}$$

where

$$v_{m_t} = \sum_{i=1}^{P-1} \lambda_{m_t}^{(i)} \mu_{m_t}^{(i)}. \tag{5}$$

[1] For notational simplicity, acoustic features are assumed to be scalar values. Extension for vectors is straightforward.

Equation (4) is almost identical to the state-output distribution of cluster-adaptive training (CAT) [8],[2] *i.e.*, additive components in the proposed model correspond to clusters in CAT. In both speech recognition and synthesis, there are a vast number of possible contexts. Unfortunately, it is almost impossible to cover all possible contexts with a finite set of training data. To address this problem, the decision tree-based context clustering technique [9] has been widely used. This technique clusters similar contexts into the same class and ties model parameters among contexts associated to the same class. It also offers generation of unseen contexts by assigning them to a certain class by traversing decision trees. The derivation of CAT in [8] assumes that all clusters have the same parameter tying structure. However, there is no such explicit restriction in CAT, *i.e.*, each cluster and variance can have its own decision tree-clustered parameter tying structure.[3] The following section describes how to build an individual decision tree for each cluster and derives parameter reestimation formulae for CAT with cluster-dependent decision trees.

2.2. Parameter estimation

This section derives the parameter reestimation formulae for CAT with a cluster-dependent parameter sharing structure. The derivation here assumes that the parameter sharing structure (decision trees) of additive bias terms, scaling factors, and variances have already been constructed. From Eq. (4), the auxiliary function of the EM algorithm can be derived as

$$\mathcal{Q}(\Lambda, \Lambda') = -\frac{1}{2}\sum_{m,t}\gamma_{m,t}\left(\log\left|\sigma^2_{v(m)}\right| + \frac{(o_t - \nu_m)^2}{\sigma^2_{v(m)}}\right) + C, \tag{6}$$

where $\gamma_{m,t}$ is the posterior probability of a context m generating the acoustic feature o_t given the current model parameters Λ', $v(m) \in \{1 \ldots V\}$ and V correspond to the leaf node which m belongs to and the total number of leaf nodes in the decision trees for variances, and C is a constant independent of $\{\sigma_i^2\}$ and $\{\nu_m\}$. In Eq. (6), ν_m is given by

$$\nu_m = \sum_{i=1}^{P-1}\lambda^{(i)}_{w(m)}\mu_{c_i(m)}, \tag{7}$$

where $c_i(m) \in \{1 \ldots N\}$ and N correspond to the leaf node which the i-th additive bias term of m belongs to and the total number of leaf nodes in the decision trees for additive bias terms, and $w(m) \in \{1 \ldots W\}$ and W correspond to the scaling factor class of m and the total number of the scaling factor classes. By equating the first partial derivative of Eq. (6) with respect to $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^\top$ to $\mathbf{0}$, a set of linear equations to determine $\boldsymbol{\mu}$ is obtained as

$$\boldsymbol{G}\boldsymbol{\mu} = \boldsymbol{k}, \tag{8}$$

where

$$\boldsymbol{G} = \begin{bmatrix} g_{11} & \cdots & g_{1N} \\ \vdots & \ddots & \vdots \\ g_{N1} & \cdots & g_{NN} \end{bmatrix}, \quad \boldsymbol{k} = \begin{bmatrix} k_1 \\ \vdots \\ k_N \end{bmatrix}, \tag{9}$$

$$g_{n_1 n_2} = \sum_{\substack{m,t,i,j \\ c_i(m)=n_1 \\ c_j(m)=n_2}} \gamma_{m,t}\frac{1}{\sigma^2_{v(m)}}\lambda^{(i)}_{w(m)}\lambda^{(j)}_{w(m)} = g_{n_2 n_1}, \tag{10}$$

$$k_{n_1} = \sum_{\substack{m,t,i \\ c_i(m)=n_1}} \gamma_{m,t}\frac{1}{\sigma^2_{v(m)}}\lambda^{(i)}_{w(m)}o_t. \tag{11}$$

The matrix $\boldsymbol{G}$ can be viewed as a (weighted) co-occurrence matrix of the additive bias terms. When $\boldsymbol{G}$ cannot have full rank, the least-square solution is calculated instead of solving Eq. (8) as

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \|\boldsymbol{G}\boldsymbol{\mu} - \boldsymbol{k}\|_2. \tag{12}$$

It is difficult to store and solve Eqs. (8) and (12) in a straightforward way because the size of $\boldsymbol{G}$ becomes tens of thousands on large data. However, the use of a sophisticated matrix storage format and a sparse linear solver can avoid this problem since $\boldsymbol{G}$ is sparse and symmetric. The authors used the compressed sparse row (CSR) format to store $\boldsymbol{G}$ and the minimum residual (MINRES) method [13] to solve Eq. (12). The update formulae of the variances $\{\sigma_m^2\}$ and the scaling factors $\{\lambda_m^{(i)}\}$ are the same as those derived in [8].

2.3. Decision tree-based context clustering

The derivation of CAT in [8] assumes that all clusters have the same parameter tying structure (decision trees). However, there is no such explicit restriction in CAT, *i.e.*, each cluster and variance can have its own parameter tying structure. The idea of building cluster-dependent decision trees was first proposed by Saino for factor analysis-based eigenvoice models [14].[4] This allows us to build an individual decision tree for each cluster. In [14], a successive tree re-construction strategy was adopted to build decision trees, *i.e.*, while building a decision tree for a cluster, decision trees and parameters of another clusters are fixed. However, as mentioned in [7], the simultaneous clustering strategy seems essential to extract the additive structure because the additive components interact with one another to compose state-output distributions. The critical drawback of the simultaneous clustering strategy is its computational complexity because all split candidates should be re-evaluated at every split [7]. To reduce its computational complexity, this paper introduces constraints and heuristics.

Equations (10)–(12) indicate that $\hat{\boldsymbol{\mu}}$ depends on $\{\sigma_i^2\}$. Likewise, the ML estimates of $\{\sigma_i^2\}$ also depend on $\hat{\boldsymbol{\mu}}$. Therefore, the reestimation of both $\{\sigma_i^2\}$ and $\boldsymbol{\mu}$ should be repeated until it converges. Although $\{\sigma_i^2\}$ values can be fixed during the clustering process of the additive bias terms, it is not a good approximation because $\{\sigma_i^2\}$ strongly depends on $\boldsymbol{\mu}$. To avoid this problem, as investigated in [15], variances of all contexts are tied while clustering additive bias terms as

$$\forall_i \sigma_i^2 = \sigma^2, \tag{13}$$

where σ^2 is a globally-tied variance. By tying all variances, $\{\sigma_i^2\}$ can be removed from Eqs. (10) and (11) thus $\hat{\boldsymbol{\mu}}$ is independent of $\{\sigma_i^2\}$. This eliminates the requirement of repeatedly updating $\boldsymbol{\mu}$ and $\{\sigma_i^2\}$.

When the decision trees for the additive bias terms have N terminal nodes in total, the current clustering structure can be

[2] This technique has originally been developed for rapid speaker adaptation for speech recognition. It may be viewed as an extension to the speaker clustering technique.

[3] Note that the proposed model is closely related to multiple additive regression tree [10], Bayesian additive regression tree [11], and constrained tree regression [12].

[4] This model can be viewed as a generalised version of CAT.

expressed as

$$\boldsymbol{\tau} = \{c_1(1), \ldots, c_{P-1}(1), \ldots, c_1(M), \ldots, c_{P-1}(M)\}, \quad (14)$$

where M is the total number of contexts that appeared in the training data. The ML estimate of globally-tied variance for the current clustering structure can be determined as

$$\hat{\sigma}^2_{\boldsymbol{\tau}} = \frac{1}{T}\left(\sum_t o_t^2 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\tau}}^\top (2\boldsymbol{k}_{\boldsymbol{\tau}} - \boldsymbol{G}_{\boldsymbol{\tau}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\tau}})\right), \quad (15)$$

where T is the total number of frames in the training data, $\hat{\sigma}^2_{\boldsymbol{\tau}}$, $\boldsymbol{G}_{\boldsymbol{\tau}}$, $\boldsymbol{k}_{\boldsymbol{\tau}}$, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\tau}}$ correspond to $\hat{\sigma}^2$, $\boldsymbol{G}$, $\boldsymbol{k}$, and $\hat{\boldsymbol{\mu}}$ for $\boldsymbol{\tau}$. The total log likelihood, $\mathcal{L}(\boldsymbol{\tau})$, can be calculated as

$$\mathcal{L}(\boldsymbol{\tau}) = -\frac{T}{2}\left(1 + \log\left(2\pi\left|\hat{\sigma}^2_{\boldsymbol{\tau}}\right|\right)\right). \quad (16)$$

For all split candidates, Eq. (16) is evaluated, and then an optimal split that maximises the increase in log likelihood is chosen. If this increase exceeds a threshold, the node is divided using the optimal question and two new terminal nodes are created. This procedure is repeated until none of the terminal nodes can be split. The computationally expensive part of the simultaneous clustering strategy is that all split candidates should be re-evaluated at every split because the additive components interact with one another to compose state-output distributions. To reduce the number of split candidates, a limited number of split candidates are re-evaluated in this study.

The variance terms can also be clustered by decision trees while fixing the additive bias terms. The authors omit the details of variance clustering due to space limitations.

3. Application to log F_0 modelling

The effectiveness of the additive acoustic model depends on whether acoustic features really have a context-dependent additive nature. Here the proposed additive acoustic model is applied to log F_0 modelling. It is known that log F_0 has an additive nature, and various additive log F_0 models have been proposed, *e.g.*, [4, 5]. In HMM-based speech synthesis, Yao *et al.* proposed the boosting-style additive log F_0 model [6]. The additive acoustic model proposed in [7] was also originally motivated by log F_0 modelling for HMM-based speech synthesis. In HMM-based speech synthesis, log F_0 contours are modelled by multi-space probability distribution HMMs (MSD-HMMs) [16]. These can model variable-dimensional observations such as log F_0 with unvoiced regions without any heuristic assumptions. An MSD for log F_0 consists of a Gaussian distribution (for modelling voiced frames), a discrete distribution (for modelling unvoiced frames), and their weights (for modelling voiced/unvoiced ratio). The proposed technique is integrated into MSD-HMMs for modelling voiced frames.

Based on the linguistics knowledge, here voiced frames of log F_0 contours are assumed to have the following additive structure:

$$o_t = o_t^{(\text{spkr})} + o_t^{(\text{utt})} + o_t^{(\text{phr})} + o_t^{(\text{wrd})} + o_t^{(\text{syl})} + o_t^{(\text{phn})} + o_t^{(\text{state})} + o_t^{(\text{var})}, \quad (17)$$

where $o_t^{(\text{spkr})}$ is a context-independent speaker-dependent additive bias term, $o_t^{(\text{utt})}$, $o_t^{(\text{phr})}$, $o_t^{(\text{wrd})}$, $o_t^{(\text{syl})}$, and $o_t^{(\text{phn})}$ are utterance-, phrase-, word-, syllable-, and phoneme-level context-dependent additive bias terms, $o_t^{(\text{state})}$ is an HMM-state-level context-dependent additive bias term, and $o_t^{(\text{var})}$ is the remaining additive component generated according to the zero-mean Gaussian distribution (Eq. (3)). A decision tree is built for each of $o_t^{(\text{utt})}$, $o_t^{(\text{phr})}$, $o_t^{(\text{wrd})}$, $o_t^{(\text{syl})}$, and $o_t^{(\text{phn})}$ using questions for utterance, utterance+phrase, utterance+phrase+word, utterance+phrase+word+syllable, and all contexts, respectively. For $o_t^{(\text{state})}$, a decision tree is built for each state position with questions for all contexts. These decision trees are simultaneously constructed in the way described in Section 2.3. For $o_t^{(\text{var})}$, a decision tree is built for each state position with questions for all contexts. At the synthesis stage, a sentence HMM is composed by concatenating context-dependent HMMs according to the context-dependent label sequence converted from a text to be synthesized and then a smooth log F_0 contour is generated as well as spectral and excitation parameters using the speech parameter generation algorithm [2]. Finally, a speech waveform is reconstructed from the generated speech parameters using a speech synthesis filter.

4. Experiments

Because the clustering part of the proposed technique was still computationally very expensive, the proposed technique was evaluated with a small database. Two hundred and fifty English sentences uttered by a female professional speaker were used for training. The speech analysis conditions and model topologies of Nitech-HTS 2005 [17] were used in this experiment.[5] Note that phoneme segmentations, prosodic labels, and log F_0 values were manually corrected. After training the baseline system, the proposed additive log F_0 modelling technique was applied to the log F_0 streams of the baseline system. Because parameters in spectral and aperiodicity streams and state-duration distributions were unchanged, the baseline and proposed systems generated exactly the same spectral sequences and durations at the synthesis stage. The additive structure of Eq. (17) and its simplified version ($o_t = o_t^{(\text{spkr})} + o_t^{(\text{syl})} + o_t^{(\text{state})} + o_t^{(\text{var})}$) were investigated. The former and latter are referred to as `All` and `Syl+State` in the following.

In this experiment, all scaling factors $\{\lambda_m^{(i)}\}$ were fixed to 1.0. Tree growth was stopped when the total number of leaf nodes in the decision trees for the bias terms was reached to the same as that of the baseline system. The number of re-evaluations of split candidates for each node was limited to the top 64 candidates. The standard Jacobi pre-conditioner was used in association with the MINRES method to improve the convergence rate. The convergence tolerance of the MINRES method was set to 10^{-6}. Training of the proposed additive log F_0 model took approximately three days on a machine with 2 × quad-core Intel Xeon CPU 2.66 GHz (8 CPU cores in total). Note that clustering was parallelised using OpenMP thus 8 CPU cores were fully used.

Table 1 shows the numbers of leaf nodes in the cluster-dependent decision trees. It can be seen from the table that splits in higher levels occurred in the proposed technique and the additive acoustic models had the slightly larger number of parameters than `Baseline`.

To evaluate the performance of the proposed additive log F_0 model, a preference listening test was conducted. This test compared the baseline system and proposed additive log F_0 mod-

[5] The speech parameter generation algorithm considering global variance [18] was not used for log F_0 generation.

Table 1: The numbers of leaf nodes in cluster-dependent decision trees. Note that the decision trees for $o_t^{(\text{state})}$ and $o_t^{(\text{var})}$ of `Baseline` are identical.

Additive components	# of leaf nodes		
	`Baseline`	`Syl+State`	`All`
$o_t^{(\text{spkr})}$	–	1	1
$o_t^{(\text{utt})}$	–	–	11
$o_t^{(\text{phr})}$	–	–	48
$o_t^{(\text{wrd})}$	–	–	52
$o_t^{(\text{syl})}$	–	1,230	553
$o_t^{(\text{phn})}$	–	–	1,314
$o_t^{(\text{state})}$	2,607	1,376	628
$o_t^{(\text{var})}$	2,607	3,075	2,765
Total	5,214	5,682	5,372

Table 2: Preference scores (%) between the conventional HMM-based (`Baseline`) and proposed additive log F_0 models (`Syl+State` and `All`).

`Baseline`	`Syl+State`	No preference
32.6	43.5	23.9

`Baseline`	`All`	No preference
41.8	32.7	25.4

elling techniques over 60 sentences (10 domains) not included in the training sentences. The subjects were six speech researchers consisted of two native US English speakers and four native UK English speakers. Forty sentences were randomly chosen from the evaluation sentences for each subject. Samples were presented in a random order for each test sentence. Before starting the test, the subjects listened to speech samples of one sentence to become familiar with the task. This sentence was randomly chosen for each subject and excluded from the actual test. After listening to each test sample, the subjects were asked to choose their preferred one. Note that the subjects could select "No preference" if they had no preference.

Table 2 shows the preference test result. It can be seen from the table that `Syl+State` achieved a better score but `All` was worse than `Baseline`. The authors found that the proposed technique tended to overfit to the training data. Although the trained additive acoustic model gives much higher likelihood to the training data than the baseline HMM with the same number of parameters, it gives significantly lower likelihood to some held-out data. The proposed technique has more representation ability than the baseline HMM even with the same number of parameters. We expect that this caused the overfitting problem and significant difference between the preference scores of `Syl+State` and `All` against `Baseline` because `All` has more representation ability to fit to training data.

5. Conclusion

This paper proposed a context-dependent additive acoustic modelling technique and its application to logarithmic fundamental frequency (log F_0) modelling for HMM-based speech synthesis. In the proposed additive acoustic model, mean vectors of state-output distributions are composed as the weighted sum of decision tree-clustered context-dependent additive bias terms. The proposed technique has the potential to capture the additive structure of log F_0 contours. A preliminary experiment using small speech data showed that the proposed technique yielded encouraging results.

The clustering part of the proposed additive acoustic modelling technique is still computationally expensive even with the constraints and approximations described in this paper. The authors expect that GPU-based parallelisation is useful to solve this problem because evaluations of split candidates in the clustering process, which dominate the computational cost, are independent of one another. The authors are also planning to perform evaluations with a large database and apply this technique to other speech parameters.

6. References

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

[2] K. Tokuda, *et al.*, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[3] Z.-H. Ling, *et al.*, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. Interspeech*, 2008, pp. 573–576.

[4] H. Fujisaki, "In search of models in speech communication research," in *Proc. Interspeech*, 2008, pp. 1–10.

[5] S. Sakai, "F0 modeling with multi-layer additive modeling based on a statistical learning technique," in *Proc. ISCA SSW5*, 2004, pp. 151–154.

[6] Y. Qian, *et al.*, "Generating natural F0 trajectory with additive trees," in *Proc. Interspeech*, 2008, pp. 2126–2129.

[7] Y. Nankaku, *et al.*, "Acoustic modeling with contextual additive structure for HMM-based speech recognition," in *Proc. ICASSP*, 2008, pp. 4469–4472.

[8] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 417–428, 2000.

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

[10] J. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, pp. 367–378, 1999.

[11] H. Chipman, *et al.*, "Bayesian ensemble learning," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, *et al.*, Eds. Cambridge, MA: MIT Press, 2007, pp. 265–272.

[12] N. Iwahashi and Y. Sagisaka, "Statistical modelling of speech segment duration by constrained tree regression," *IEICE Trans. Inf. Syst.*, vol. E83-D, no. 7, pp. 1550–1559, 2000.

[13] C. Paige and M. Saunders, "Solution of sparse indefinite systems of linear equations," *SIAM J. Numerical Analysis*, vol. 12, pp. 617–629, 1975.

[14] K. Saino, "A clustering technique for factor analysis-based eigenvoice models," Master thesis, Nagoya Institute of Technology, 2008, (in Japanese).

[15] K. Oura, *et al.*, "Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems," in *Proc. Interspeech*, 2009.

[16] K. Tokuda, *et al.*, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[17] H. Zen, *et al.*, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

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