ADAPTATION OF PITCH AND SPECTRUM FOR HMM-BASED SPEECH SYNTHESIS USING MLLR

Masatsune Tamura†, *Takashi Masuko*†, *Keiichi Tokuda*††, *Takao Kobayashi*†

†Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan

††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan

Email: {mtamura,masuko,tkobayas}@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper describes a technique for synthesizing speech with an arbitrary speaker characteristics using speaker independent speech units, which we call "average voice" units. The technique is based on an HMM-based text-to-speech (TTS) system and MLLR adaptation algorithm. In the HMM-based TTS system, speech synthesis units are modeled by multi-space probability distribution (MSD) HMMs which can model spectrum and pitch simultaneously in a unified framework. We derive an extension of the MLLR algorithm to apply it to MSD-HMMs. We demonstrate that a few sentences uttered by a target speaker are sufficient to adapt not only voice characteristics but also prosodic features. Synthetic speech generated from adapted models using only four sentences is very close to that from speaker dependent models trained using 450 sentences.

1. INTRODUCTION

Text-to-Speech (TTS) synthesis which generate speech with arbitrary voice characteristics and speaking styles is one of the key technologies for realizing human computer interaction systems. There have been proposed a number of TTS techniques, and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech [1] [2]. However, it is not easy to make these systems have the ability of synthesizing with various voice characteristics and speaking styles, because it is impractical to prepare a large number of speech units of arbitrary speakers.

We have proposed an HMM-based TTS system in which each speech synthesis unit is modeled by HMM [3]. A distinctive feature of the system is that speech parameters used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm [4] [5]. The parameter generation algorithm takes account of dynamic features of speech parameters and this results in providing realistic speech parameter sequences. Since the HMM-based TTS system uses HMMs as the speech units in both modeling and synthesis, we can easily change voice characteristics of synthetic speech by transforming HMM parameters appropriately. In fact, we have shown that voice characteristics of synthetic speech are converted from one speaker to another using a small amount of target speaker's speech data by applying speaker adaptation techniques, such as MAP/VFS (Maximum *a Posteriori* / Vector Field Smoothing) algorithm [6], or MLLR (Maximum Likelihood Linear Regression) algorithm [7].

In this paper, we propose a technique which enables the HMM-based TTS system to change not only voice characteristics but also prosodic features. In the HMM-based TTS system, spectrum, pitch, and state duration are modeled simultaneously in a unified framework of HMM [8]. Specifically, spectrum and pitch are modeled by multi-space probability distribution (MSD) HMMs [9] which includes conventional continuous and discrete HMMs as special cases. We derive an extension of the MLLR algorithm [10] which can be applied to MSD-HMMs. To generate speech of an arbitrarily given target speaker, we first make speaker-independent speech units modeled by MSD-HMMs, which we call "average voice" models, then we adapt the average voice models to the target speaker using the extended MLLR algorithm.

2. HMM-BASED SPEECH SYNTHESIS SYSTEM

2.1. System overview

A blockdiagram of the HMM-based TTS system is shown in Fig.1. The system consists of three stages, the training stage, the adaptation stage, and the synthesis stage.

In the training stage, mel-cepstral coefficients [11] and logarithm of fundamental frequency are extracted as the static features from multi-speaker speech database. Then, the dynamic features, i.e., delta and delta-delta parameters, are calculated from the static features. Spectral parameters and pitch observations are combined into one observation vector frame by frame (Fig.2), and speaker independent phoneme HMMs, which we refer to as the average voice HMMs, are trained using the observation vectors. To model variations of spectrum and pitch, phonetic and linguistic contextual factors, such as phoneme identity factors and stress related factors, are taken into account [8]. Since pitch observations are composed of one-dimensional continuous values corresponding to pitch values and discrete symbols representing unvoiced, conventional discrete or continuous HMMs cannot be applied for pitch pattern modeling without any heuristic assumptions. To overcome this problem, we adopt multi-space probability distributions (MSDs) [9], and model spectral and pitch parameters simultaneously by multi-stream MSD-HMMs [8]. Then, a decision tree based context clustering technique [12] [13] is separately applied to the spectral and pitch parts of the context dependent phoneme HMMs. Finally, state durations are modeled by multi-dimensional Gaussian distributions, and the state clustering technique is also applied to the duration models [14].

In the adaptation stage, the average voice HMMs are adapted to a target speaker using speech from the target speaker. In this paper, we adopt the Maximum Likelihood Linear Regression (MLLR) algorithm [10] and extend it to MSD-HMM as shown in Section 3. The extended MLLR technique is applied to the mean vectors of the distributions in each stream of the average voice HMMs. As a result, distributions for spectral and pitch parameters are adapted simultaneously.

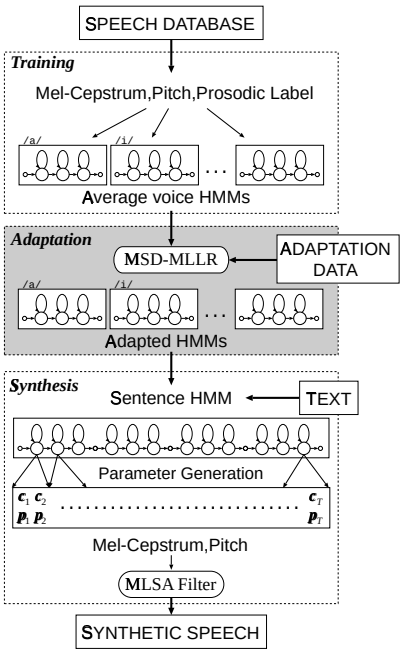

Fig. 1. A blockdiagram of HMM-based speech synthesis system.

In the synthesis stage, first, an arbitrarily given text to be synthesized is transformed into a context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating adapted phoneme HMMs. From the sentence HMM, spectral and pitch parameter sequences are generated using the algorithm for speech parameter generation from HMMs with dynamic features [4], in which phoneme durations are determined based on state duration distributions [14]. Finally, by using MLSA filter [11], speech is synthesized from the generated mel-cepstral and pitch parameter sequences.

2.2. Pitch modeling using MSD-HMM

We assume that pitch pattern is a sequence of outputs from a one-dimensional space Ω_1 and a zero-dimensional space Ω_2 which correspond to voiced and unvoiced regions of speech, respectively. Each space Ω_g has its probability w_g, i.e., probability for voiced observation w_1 and for unvoiced observation w_2, where $\sum_{g=1}^{2} w_g = 1$. The space Ω_1 has a one-dimensional probability density function $\mathcal{N}_1(\boldsymbol{x})$ where $\int_{\Omega_1} \mathcal{N}_1(\boldsymbol{x}) d\boldsymbol{x} = 1$, and Ω_2 has only one sample point. A pitch observation $\boldsymbol{o}$ consists of a continuous random variable $\boldsymbol{x}$ and a set of space indices X, that is,

$$\boldsymbol{o} = (X, \boldsymbol{x}) \tag{1}$$

where $X = \{1\}$ for voiced region and $X = \{2\}$ for unvoiced region. The observation probability of $\boldsymbol{o}$ is defined by

$$b(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{o})} w_g \mathcal{N}_g(V(\boldsymbol{o})) \tag{2}$$

where

$$V(\boldsymbol{o}) = \boldsymbol{x}, \quad S(\boldsymbol{o}) = X.$$

It is noted that, although $\mathcal{N}_2(\boldsymbol{x})$ does not exist for Ω_2, we define as $\mathcal{N}_2(\boldsymbol{x}) \equiv 1$ for simplicity of notation.

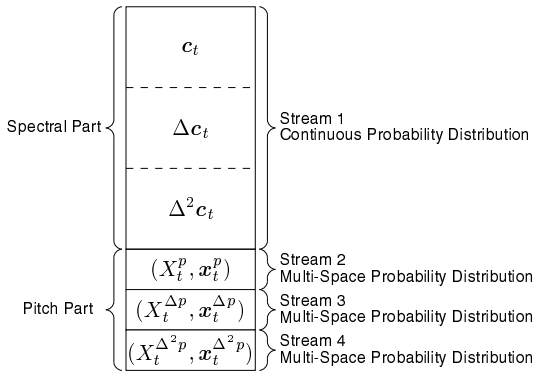

Fig. 2. Observation vector. In the figure, $\boldsymbol{c}_t$, $\boldsymbol{X}_t^p$, and $\boldsymbol{x}_t^p$ represent the spectral parameter vector, a set of space indices of pitch, and pitch parameter at time t, respectively, and Δ and Δ^2 represent the delta and delta-delta parameters, respectively.

Using an HMM whose output probability in each state is given by eq.(2), called MSD-HMM, we can model voiced and unvoiced observations of pitch in a unified model without any heuristic assumption [9] . Moreover, we can model spectrum and pitch simultaneously by multi-stream MSD-HMM, in which spectral part is modeled by continuous probability distribution (CD), and pitch part is modeled by MSD (see Fig. 2).

3. MULTI-SPACE PROBABILITY DISTRIBUTION MLLR

To generate speech with an arbitrarily given target speaker's voice, we adapt the speaker independent models, i.e., average voice models, to the target speaker. We use here a transformation-based model adaptation approach. Maximum likelihood estimation of the transformation matrices for MSD-HMM is derived in a manner similar to MLLR [10].

Mean adaptation of MLLR is based on affine transformation. Let $\boldsymbol{\mu}_{ig}, \boldsymbol{U}_{ig}$ be the mean vector and the covariance matrix of output probability $\mathcal{N}_{ig}(\boldsymbol{x})$ for the gth space of state i, respectively. For given adaptation samples $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$, the new mean vector $\hat{\boldsymbol{\mu}}_{ig}$ is estimated by

$$\hat{\boldsymbol{\mu}}_{ig} = \boldsymbol{W}_{ig} \boldsymbol{\xi}_{ig} \tag{4}$$

where $\boldsymbol{\xi}_{ig} = [1, \boldsymbol{\mu}_{ig}^T]^T$.

To derive a maximum likelihood estimation of transformation matrix $\boldsymbol{W}_{ig}$, we define an auxiliary function $Q(\lambda', \lambda)$ of the current parameters λ' and the newly estimated parameter λ as

$$Q(\lambda', \lambda) = \sum_{all\, \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda') \log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l} | \lambda) \tag{5}$$

where $\boldsymbol{q}$ and $\boldsymbol{l}$ are possible state and space sequences, respectively. It is shown that the auxiliary function of MSD-HMM increases monotonically in the likelihood unless λ is a critical point of the likelihood.

We next define the probability $\gamma_{ig}(t)$ of being in state i and space Ω_g at time t, given the model λ and the observation $\boldsymbol{O}$, as follows:

$$\begin{aligned} \gamma_{ig}(t) &= P(q_t = i, l_t = g | \boldsymbol{O}, \lambda) \\ &= \frac{1}{P(\boldsymbol{O}|\lambda)} \sum_{all\, \boldsymbol{q}, \boldsymbol{l}} P(\boldsymbol{O}, q_t = i, l_t = g | \lambda). \end{aligned} \tag{6}$$

We also define a set of time slots $T(\boldsymbol{O}, g)$ as

$$T(\boldsymbol{O}, g) = \{t | g \in S(\boldsymbol{o}_t)\} \tag{7}$$

and use the following notation

$$h(\boldsymbol{o}_t, i, g) = (V(\boldsymbol{o}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig})^T \boldsymbol{U}_{ig}^{-1}(V(\boldsymbol{o}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig}). \tag{8}$$

Then, the auxiliary function of (5) becomes

$$\begin{aligned} &Q(\lambda', \lambda) \\ &= K + \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} P(\boldsymbol{O}, q_t = i, l_t = g|\lambda') \\ &\quad (\log w_{ig} - \frac{n_g}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{U}_{ig}^{-1}| - \frac{1}{2}h(\boldsymbol{o}_t, i, g)) \\ &= K + P(\boldsymbol{O}|\lambda')\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t)(\log w_{ig} - \\ &\qquad \frac{n_g}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{U}_{ig}^{-1}| - \frac{1}{2}h(\boldsymbol{o}_t, i, g)) \end{aligned} \tag{9}$$

where K is a constant which is independent of $\boldsymbol{W}_{ig}$, and n_g is the dimensionality of space Ω_g. Differentiating eq.(9) with respect to $\boldsymbol{W}_{ig}$, we have

$$\begin{aligned} &\frac{d}{d\boldsymbol{W}_{ig}}Q(\lambda', \lambda) \\ &= -\frac{1}{2}P(\boldsymbol{O}|\lambda')\frac{d}{d\boldsymbol{W}_{ig}}\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t)h(\boldsymbol{o}_t, i, g) \\ &= P(\boldsymbol{O}|\lambda')\sum_{t\in T(\boldsymbol{o},g)} \gamma_{ig}(t)\boldsymbol{U}_{ig}^{-1}(V(\boldsymbol{O}_t) - \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig})\boldsymbol{\xi}_{ig}^T. \end{aligned} \tag{10}$$

By setting $dQ(\lambda', \lambda)/d\boldsymbol{W}_{ig} = 0$, we obtain a maximum likelihood estimation of $\boldsymbol{W}_{ig}$ as

$$\begin{aligned} &\sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t)\boldsymbol{U}_{ig}^{-1}V(\boldsymbol{o}_t)\boldsymbol{\xi}_{ig}^T \\ &\qquad = \sum_{t\in T(\boldsymbol{O},g)} \gamma_{ig}(t)\boldsymbol{U}_{ig}^{-1}\boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig}\boldsymbol{\xi}_{ig}^T. \end{aligned} \tag{11}$$

When the amount of adaptation data is limited, the transformation matrices $\boldsymbol{W}_{ig}$ should be tied across several Gaussian distributions. We construct a regression class tree to group the distributions. By doing this, we can estimate transformation matrices which is not observed in the adaptation data and estimate transformation matrices according to the amount of adaptation data.

Regression class tree is constructed by recursively applying a binary splitting algorithm based on LBG for all Gaussian distributions. In the binary tree, each leaf has a distribution, and all the distributions below the lowest node in which the amount of adaptation data is larger than the prescribed threshold are adapted using the same transformation matirx.

Here we apply the above adaptation algorithm to the MSD-HMMs which represent pitch information. In this case, the mean vectors of distribution $\mathcal{N}_1(\boldsymbol{x})$ of the space corresponding to voiced sounds are adapted. The adaptation of spectral parameters is accomplished by applying ordinary MLLR adaptation technique.

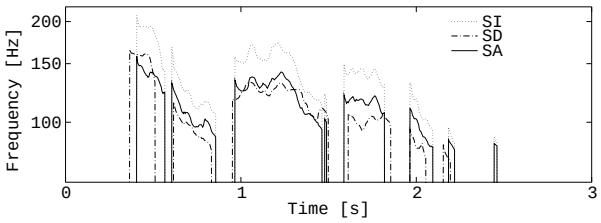

Fig. 3. Comparison of pitch contours generated from speaker independent models (SI), speaker dependent models (SD), and speaker adapted models (SA).

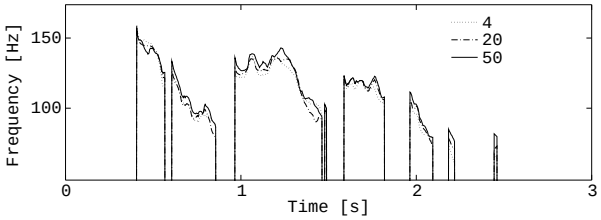

Fig. 4. Comparison of pitch contours generated from speaker adapted models with 4, 20, and 50 sentences.

4. EXPERIMENTS

4.1. Experimental conditions

We used phonetically balanced sentences from ATR Japanese speech database for training and adaptation. Based on phoneme labels and linguistic information included in the database, we made context dependent phoneme labels. We used 42 phonemes including silence and pause. The details of contextual factors are shown in [8].

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. Pitch values were obtained using ESPS `get_f0` program [15]. Delta and delta-delta pitch parameters were calculated only within voiced regions, and the frames where delta or delta-delta pitch parameters were not computable because of the boundaries of voiced and unvoiced regions were treated as unvoiced. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right models in which the spectral part of each state is modeled by single diagonal Gaussian output distributions.

The average voice HMMs (SI models) were trained using 5 male speakers' speech data, 400 sentences for each speaker. The states of the models were clustered using a decision tree based context clustering technique with MDL-criterion [13] . The total number of states in SI models is 3,765 for spectral part and 12761 for pitch part. We chose a male speaker MHT from the database as the target speaker, who was not included in the training speakers of average voice HMMs. We also trained HMMs using 450 sentences uttered by MHT (SD models) to compare with the adapted models. The total number of states in SD models is 906 for spectral part and 1894 for pitch part. Thresholds for traversing regression class tree were set to 1500 for spectral stream and 100 for pitch stream.

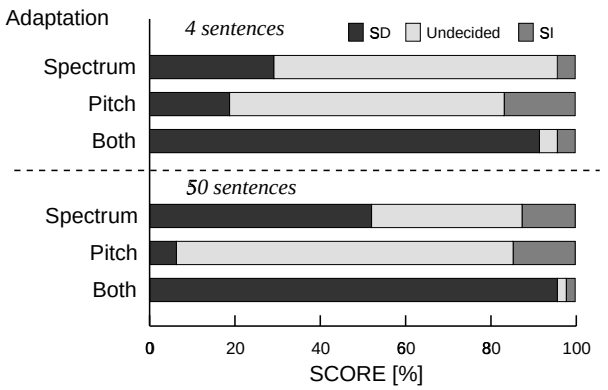

Fig. 5. Results of ABX-Listening tests.

4.2. Pitch contour generation

Figure 3 shows an example of the generated pitch contours from a Japanese sentence /fu-ko-o-he-e-no-so-N-za-i-wa-hi-ni-N-shi-na-ka-tta/ ("he didn't deny the existence of unfairness," in English) which is not included in the training sentences. In Fig. 3, dotted line, dash-dotted line, and solid line represent the pitch contours generated from average voice HMMs (SI), speaker dependent HMMs (SD), and speaker adapted HMMs (SA), respectively. In this example, the number of adaptation utterances is 50 sentences. From this figure, it can be seen that the pitch contour generated from SA HMMs becomes closer to that generated from SD HMMs. It is noted that phoneme durations generated from SI HMMs and SA HMMs are the same because duration models were not adapted.

Figure 4 shows the result using SA HMMs with a small amount of adaptation data. In the figure, dotted line, dash-dotted line, and solid line represent pitch contours generated from SA HMMs adapted using 4, 20, and 50 sentences, respectively. From this figure, we can see that only a few utterances are sufficient to adapt pitch HMMs.

4.3. Subjective evaluations

To evaluate the performance of the proposed technique, we conducted ABX listening tests. In the ABX test, A and B were either synthetic speech generated from SI and SD models (the order of assignment was randomized), and X was synthetic speech generated from SA models adapted using 4 or 50 sentences. Subjects were six males and asked to select A or B as being similar to X. Figure 5 shows the results of the listening tests. In these tests, four sentences, which were not included in the adaptation data, were synthesized and tested. In the figure, "Both" indicates the case for simultaneous adaptation of pitch and spectrum, while "Pitch" and "Spectrum" indicate the cases for pitch adaptation only and spectral adaptation only, respectively. The results show that adaptation of both spectrum and pitch is effective for synthesizing speech with the target speaker characteristics. It is also shown that synthetic speech generated from SA models with only four sentences is very close to that from the target speaker's models.

5. CONCLUSION

In this paper, we described a technique for adapting voice characteristics and prosodic features of HMM-based TTS system to an arbitrarily given target speaker. To convert the pitch model parameters, we extended the MLLR algorithm for MSD-HMMs. We have shown that synthetic speech generated from adapted models using average voice models with a few amount of adaptation data becomes closer to the target speaker's voice. Our future work is deriving an adaptation technique for the duration models in the same framework.

6. REFERENCES

[1] A. W. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," in *Proc. EUROSPEECH-95*, Sept. 1995, pp. 581–584.

[2] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Storm, K. Lee, and M. J. Makashay, "Corpus-based techniques in the AT&T NEXTGEN synthesis system," in *Proc. ICSLP-2000*, Oct. 2000, pp. 411–416.

[3] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP-96*, May 1996, pp. 389–392.

[4] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, May 1995, pp. 660–663.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorighms for HMM-based speech synthesis," in *Proc. ICASSP-2000*, June 2000, pp. 1315–1318.

[6] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP-97*, Apr. 1997, pp. 1611–1614.

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, Nov. 1998, pp. 273–276.

[8] 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-99*, Sept. 1999, pp. 2374–2350.

[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-99*, Mar. 1999, pp. 229–232.

[10] C.J.Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.

[12] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Mar. 1994, pp. 307–312.

[13] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EUROSPEECH-97*, Sept. 1997, pp. 99–102.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP-98*, Dec. 1998, pp. 29–32.

[15] Entropic Research Laboratory Inc, *ESPS Programs Version 5.0*, 1993.