An HMM-Based Mandarin Chinese Text-To-Speech System

Conference Paper · January 2006

DOI: 10.1007/11939993_26 · Source: DBLP

Yao Qian
Microsoft
75 PUBLICATIONS **1,729** CITATIONS
SEE PROFILE

Frank K. Soong
Microsoft
346 PUBLICATIONS **7,782** CITATIONS
SEE PROFILE

Yining Chen
Victoria University of Wellington
66 PUBLICATIONS **1,270** CITATIONS
SEE PROFILE

An HMM-Based Mandarin Chinese Text-to-Speech System

Yao Qian, Frank Soong, Yining Chen and Min Chu

Microsoft Research Asia, Beijing
{yaoqian, frankkps, ynchen, minchu}@microsoft.com

Abstract In this paper we present our Hidden Markov Model (HMM)-based, Mandarin Chinese Text-to-Speech (TTS) system. Mandarin Chinese or Putonghua, “the common spoken language”, is a tone language where each of the 400 plus base syllables can have up to 5 different lexical tone patterns. Their segmental and supra-segmental information is first modeled by 3 corresponding HMMs, including: (1) spectral envelop and gain; (2) voiced/unvoiced and fundamental frequency; and (3) segment duration. The corresponding HMMs are trained from a read speech database of 1,000 sentences recorded by a female speaker. Specifically, the spectral information is derived from short-time LPC spectral analysis. Among all LPC parameters, Line Spectrum Pair (LSP) has the closest relevance to the natural resonances or the “formants” of a speech sound and it is selected to parameterize the spectral information. Furthermore, the property of clustered LSPs around a spectral peak justify augmenting LSPs with their dynamic counterparts, both in time and frequency, in both HMM modeling and parameter trajectory synthesis. One hundred sentences synthesized by 4 LSP-based systems have been subjectively evaluated with an AB comparison test. The listening test results show that LSP and its dynamic counterpart, both in time and frequency, are preferred for the resultant higher synthesized speech quality.

Keywords: Speech synthesis, Trainable TTS, corpus-based TTS, statistics-based TTS, LSP

1 Introduction

HMM-based speech synthesis has been successfully applied to TTS synthesis of many different languages, e.g. Japanese and English [1-3]. In this framework, the spectral envelop, fundamental frequency, and duration are modeled simultaneously by the corresponding HMMs. For a given text sequence, speech parameter trajectories and corresponding signals are then generated from the trained HMMs in the Maximum Likelihood (ML) sense. HMM is very effective to model the evolution of speech signals as a stochastic sequence of acoustic feature vectors. Many techniques have been developed for HMM-based speech recognition, e.g. context-dependent modeling, state-tying based on decision tree clustering, and speaker adaptation. They can be applied equally well to HMM-based speech synthesis in the sense of parameter trajectory generation.

The current performance of HMM-based speech synthesis has been further improved by using dynamic feature constraint in trajectory generation [3] and global variance for parameter generation [4], a high quality vocoder called STRAIGHT [5], and Hidden Semi-Markov Model duration model [6], , and trajectory model [16] or minimum generation error training [17]. Compared with the large corpus based concatenative speech synthesis, HMM-based speech synthesis is statistics based and vocoded. The speech generated from it is fairly smooth. Characteristics of the synthetic speech can be easily controlled by transforming HMM parameters in a statistically tractable metric like likelihood function. Furthermore, the small footprint of the HMM synthesizer has made it an ideal choice for an embedded system.

In this paper, we apply HMM-based speech synthesis to Mandarin, a syllabically paced tonal language. A tone-dependent phone set and corresponding phonetic and prosodic question set of decision tree are designed for HMM training. Line Spectrum Pair (LSP) [7], an alternative linear prediction parametric representation, is investigated as feature parameter to HMM-based speech instead of mel-ceptral features [8]. According to the properties of LSP [9], the speech generation module is revised correspondingly. The performances of four systems based on LSPs are tested in an AB comparison test. It shows that the S*ystem III*, which uses LSP and the dynamic features of adjacent LSP differences, achieves the better performance than the S*ystem II*, using the conventional method.

The rest of paper is organized as follows. In Section 2, HMM-based speech synthesis system is briefly illustrated; the representation and properties of LSP are introduced in Section 3; the speech parameter generation algorithm based on LSP is proposed in Section 4; Section 5 shows the experimental evaluation; and the conclusions are given in Section 6.

2 HMM-based Speech Synthesis System

The schematic diagram of HMM-based Speech Synthesis system is shown in Figure 1 where both training and synthesis are shown.

In the training phase, the speech signal is converted to a sequence of observed feature vectors through the module of feature extraction and modeled by a corresponding sequence of HMMs. The observed feature vector consists of spectral parameters and excitation parameters, which are separated into different streams. The spectral feature comprises line spectrum pair (LSP) and log gain, and the excitation feature is log fundamental frequency. LSPs are modeled by continuous HMMs and F0s are modeled by multi-space probability distribution HMM (MSD-HMM) [10], which provides a cogent modeling of F0 without any heuristic assumptions or interpolations. Context-dependent phone models are used to capture the phonetic and prosody co-articulation phenomena. State typing based on decision-tree and minimum description length (MDL) [11] criterion is applied to overcome the problem of data sparseness in training. Stream-dependent models are built to cluster the spectral, prosodic and duration features into separated decision trees.

In the synthesis phase, input text is converted first into a sequence of contextual labels through the text analysis. The corresponding contextual HMMs are retrieved by traversing the trees of spectral and pitch information and the duration of each state is also obtained by traversing the duration tree, then the LSP, gain and F0 trajectories are generated by using the parameter generation algorithm based on maximum likelihood criterion with dynamic feature and global variance constraints. Finally, speech waveform is synthesized from the generated spectral and excitation parameters by LPC synthesis.

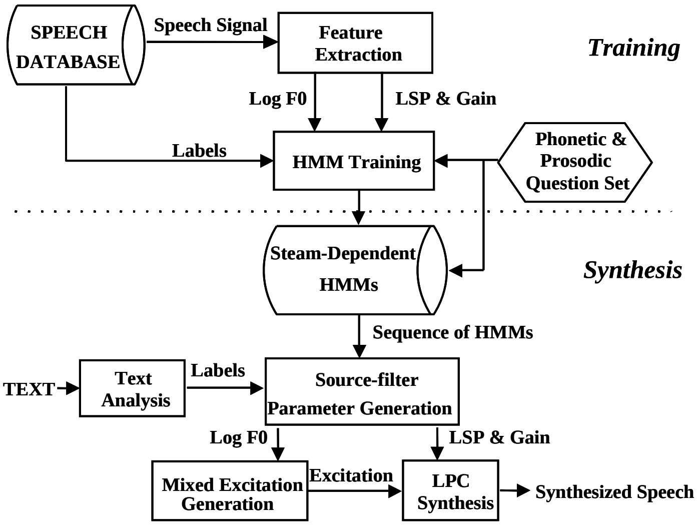

Fig. 1. HMM-based speech synthesis

3 The properties of LSP

Line Spectrum Pair (LSP) [7] is an alternative linear prediction parametric representation. In LPC analysis, the speech signal is modeled as the output of an all-pole filter *H(z)* defined as

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{M} a_i z^{-i}} \tag{1}$$

where M is the order of LPC analysis and $\{a_i\}_{i=1}^{M}$ are the corresponding LPC coefficients. The LPC coefficients can be represented by the LSP parameters, which are mathematically equivalent (one-to-one) and more amenable to quantization. LSP are calculated as follows:

$$P(z) = A(z) + z^{-(M+1)} A(z^{-1}) \tag{2}$$

$$Q(z) = A(z) - z^{-(M+1)} A(z^{-1}) \quad \textbf{(3)}$$

The symmetric polynomial *P(z)* and anti-symmetric polynomial *Q(z)* have the following two properties [9] : 1) All zeros of *P(z)* and *Q(z)* are on the unit circle; 2) zeros of *P(z)* and *Q(z)* are interlaced with each other. These properties are useful for finding the LSPs $\{\omega_i\}_{i=1}^{M}$, i.e., the roots the polynomial *P(z)* and *Q(z)*, which are ordered and bounded,

$$0 < \omega_1 < \omega_2 < \ldots < \omega_M < \pi \quad \textbf{(4)}$$

LSP has many advantages for speech representation [9,12,13]:

1) LSP parameters correlate well to “formant” or spectral peak location and bandwidth. The LPC power spectrum and the associated LSPs for vowel /a/ are shown in Figure 2, where clustered (two or three) LSPs depict a formant peak, in terms of both the center frequency and bandwidth.

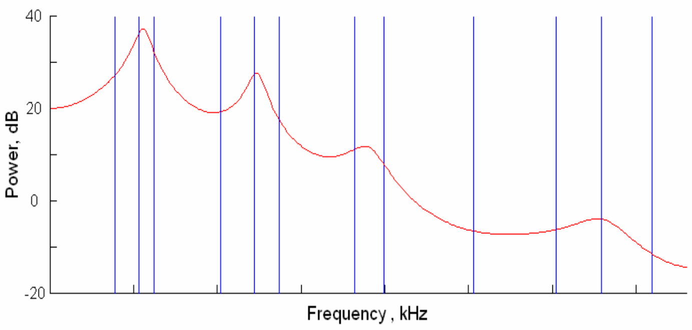

Fig. 2. LPC power spectrum and the associated LSPs for vowel /a/

2) Perturbation of an LSP parameter has a localized effect, i.e., a perturbation in a given LSP frequency only introduces a perturbation of LPC power spectrum in its neighborhood.
3) LSP parameter has a good interpolation property.

4 LSP Parameter generation

In the HMM-based speech synthesis shown in Section 2, the speech parameter generation from given HMM state sequence is based on maximum likelihood criterion. In order to generate a smoother parameter trajectory, dynamic features are used as a constraint in the generation algorithm [3]. For a given HMM λ, it determines a speech parameter vector sequence $O = [C, \Delta C, \Delta^2 C]^T$, $C = [c_1^T, c_2^T, ..., c_T^T]^T$, $\Delta C = [\Delta c_1^T, \Delta c_2^T, ..., \Delta c_T^T]^T$, $\Delta^2 C = [\Delta^2 c_1^T, \Delta^2 c_2^T, ..., \Delta^2 c_T^T]^T$, which maximizes:

$$\begin{aligned} P(O \mid \lambda) &= \sum_{all\ Q} P(O, Q \mid \lambda) \\ &\approx \max_{Q} P(O \mid Q, \lambda) P(Q \mid \lambda) \end{aligned} \qquad (5)$$

If given state sequence $Q = \{q_1, q_2, q_3, ..., q_T\}$, Eq. 5 only need consider maximizing the logarithm of $P(O \mid Q, \lambda)$ with respect to $O = WC$, i.e.,

$$\frac{\partial LogP(WC \mid Q, \lambda)}{\partial C} = 0 \qquad (6)$$

We obtain

$$W^T U^{-1} WC = W^T U^{-1} M \qquad (7)$$

where

$$W = \begin{bmatrix} I_F \\ W_{\Delta F} \\ W_{\Delta\Delta F} \end{bmatrix} = \underbrace{\left[\begin{array}{cccc} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \cdots \\ \hline 0.5 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \cdots \\ \hline 1 & & & \\ & \ddots & & \\ & & -2 & \\ & & & \ddots \\ & & & \cdots \end{array}\right]}_{DT} \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} DT \\ \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} DT \\ \left.\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right\} DT \end{array} \qquad (8)$$

$$M = [m_{q_1}^T, m_{q_2}^T, ..., m_{q_T}^T]^T \qquad (9)$$

$$U^{-1} = diag[U_{q_1}^{-1}, U_{q_2}^{-1}, ..., U_{q_T}^{-1}] \qquad (10)$$

D is the dimension of feature vector and T is the total number of frame in the sentence. W is a block matrix which composes of three $DT \times DT$ matrices: Identity matrix (I_F), delta coefficient matrix ($W_{\Delta F}$) and delta-delta coefficient matrix ($W_{\Delta\Delta F}$). M and U are the $3DT \times 1$ mean vector and the $3DT \times 3DT$ covariance matrix, respectively.

As mentioned in Section 3, a gathering of (two or three) LSPs depicts a formant frequency and the closeness of the corresponding LSPs indicates the bandwidth of a given formant. Therefore, the distance between the adjacent LSPs is more critical than the absolute value of individual LSP. On the other hand, all LSP frequencies are ordered and bounded, i.e. any two adjacent LSP trajectories do not cross each other. Using static and dynamic LSPs in modeling and generation can not ensure the stability of LSPs. Consequently, we add the difference of adjacent LSP frequencies directly into spectral parameter modeling and generation. The *W*, which is used to transform the observation feature vector, is modified as

$$W = \left[I_F \ , \quad W_{DF} \ , \quad W_{\Delta F} \ , \quad W_{\Delta F} W_{DF} \ , \quad W_{\Delta\Delta F} \ , \quad W_{\Delta\Delta F} W_{DF} \right]^T \qquad \textbf{(11)}$$

where F is static LSP; DF is the difference between adjacent LSP frequencies; ΔF and $\Delta\Delta F$ are dynamic LSPs, i.e., first and second order time derivatives; and W_{DF} is $(D-1)T \times DT$ matrix and constructed as

$$W_{DF} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \cdots & \\ & & & & \ddots \\ & & & & \\ & & & & \end{bmatrix} \qquad \textbf{(11)}$$

In this way, the correlation of adjacent LSPs can be modeled and diagonal covariance structure is still kept the same.

5 Experimental Evaluations

5.1 Experimental Setup

A broadcast news style speech corpus recorded by a female speaker is used in this study. The training data composes of 1,000 phonetically and prosodically rich sentences [14]; while the testing data consists of 100 sentences. Speech signal are sampled at 16 kHz, windowed by 25-ms window with a 5-ms shift, and transformed into 24th-order LSPs and their dynamic features in both frequency and time.

5-state,left-to-right HMMs with single, diagonal Gaussian distribution is adopted for phone model training. The phone set used is Ph97 [15], which achieved better performance than the other phone set in Mandarin tonal syllable recognition task. In Ph97, each Chinese tonal syllable is divided into a consonant followed by two consecutive tonal sonorant segments, e.g. /huang4/ is decomposed into /hu/, /aaH/ and

/ngH/. Here, the glides like /i/ and /u/ are assigned to the Initial part and 2-scales (High/Low) pitch label is used instead of five numerical scales. The phone set designed in this way can carry tone information in the modeling at a little extra cost of the phone inventory size.

The phonetic and prosodic factors, which are used as question set in decision tree growing for contextual state tying, are listed as following

1) {preceding, current, succeeding} phone
2) Break index after the current word, three indices: minor, medium and major breaks, are used.
3) tone label (in 5-categories) of {preceding, current, succeeding} syllable
4) Position of a phone in a syllable
5) Position of a syllable in a "prosodic word" which are sandwiched by minor breaks
6) Position of a syllable in a breath group phrase which are limited by major breaks
7) Length of the current breath group phrase in terms of number of syllables

5.2 Experiments and Results

Four synthesis systems based on LSP features are built for comparison.

System I:

The *W* generating the observation feature in $O = WC$ is constructed as

$$W = \left[I_F \ , \quad W_{DF} \ , \quad W_{\Delta F} \ , \quad W_{\Delta F} W_{DF} \ , \quad W_{\Delta\Delta F} \ , \quad W_{\Delta\Delta F} W_{DF} \right]^T$$

where W is a $6DT \times DT$ matrix. Considering the higher orders of LSP are almost evenly spaced and most of speech formants are located below 4kHz, only the lower 16 out of 24 LSPs are used to compute *DF* (the distance between the adjacent LSPs). Therefore, the total dimension of observation feature vector is 126 (120 for LSP, 3 for gain and 3 for F0).

System II:

In this system, the observation feature vector is computed in conventional method. It consists of static, first and second order time derivatives. The corresponding *W* is defined as

$$W = \left[I_F \ , \quad W_{\Delta F} \ , \quad W_{\Delta\Delta F} \right]^T$$

The total dimension of observation feature vector is 78.

System III:

In order to make the results comparable with *system II* in feature dimensions, we only use the static and dynamic features of LSP difference (in frequency) as observation vector. The *W* is modified as

$$W = [I_F, \quad W_{\Delta F} W_{DF}, \quad W_{\Delta\Delta F} W_{DF}]^T$$

Here, the total dimension of observation feature vector is equal to that of *system II*.

System IV

40th -order LSP are used instead of 24th -order LSP in *System II*.

One hundred sentences are synthesized by the above four systems and evaluated in a subjective test. Fifty out of the one hundred sentences are randomly selected for an AB comparison preference test. Eight subjects are forced to choose one which sounds more natural from each pair. The results of the preference test are given in Figure 3, where shows:

a) *System I* achieves a better performance than *System II*. Modeling the difference of adjacent LSP frequency (*DF*) is very critical in reproducing the salient features of speech spectrum in HMM-based speech synthesis.
b) *System III* gives almost the same performance as *System I*. But the dimensionality of feature vector in *System III* is much less than that of *System I*. It indicates that with/without dynamic features of LSP frequency difference is critical to the performance of system.
c) S*ystem IV* slightly improve the performance comparing with *system II*, i.e., the performance improvement by using a higher order LSP is marginal.

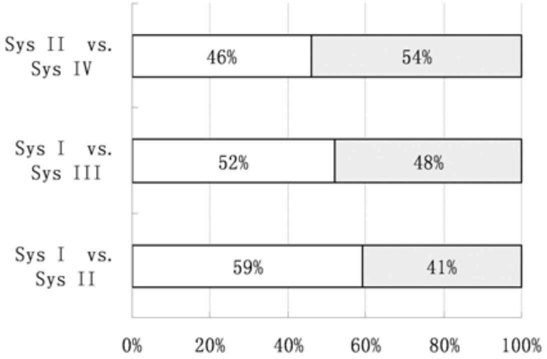

Fig. 3. The results of AB Test for four systems

5.3 Analysis

To analyze experimental results, we plot the spectra of synthesized and original speech signals for comparison. However, the duration of generated utterance can be different from that of the original since only means of state duration models are used in speech generation. An oracle experiment is designed to compare the spectra by isolating the effect of duration difference. A sequence of states, which are obtained by force-aligning the original feature observations with the spectral and pitch models, is used as *Q* in Eq. 6 for speech parameter generation. In this way, the spectra can be compared on a frame-by-frame basis between two different systems. An example of spectral comparison, LPC power spectra for vowel /u/ , is given in Fig. 4, where the bold dotted line, dotted line and solid line represent the spectra of the original, S*ystem II* and *System III*, respetively. The log power spectrum is plotted in dB scale and 25dB offset is used for separating adjacent frames. In Fig. 4 the formant structure of the generated spectra of S*ystem II* is sharper and closer to the original spectra than that of *System III*.

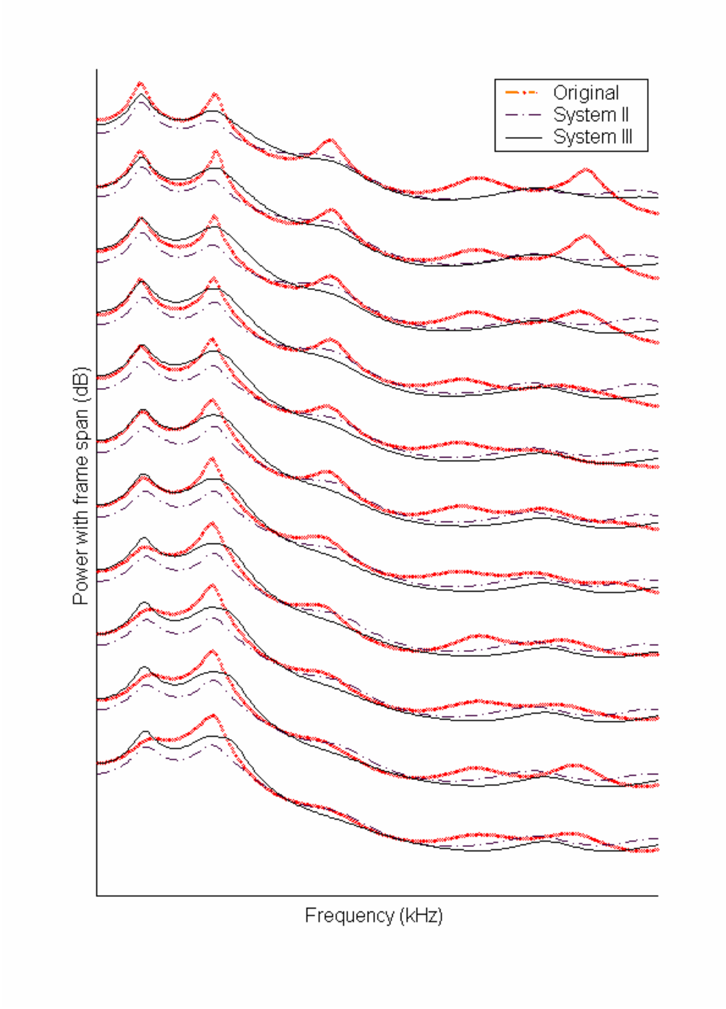

Fig. 4. LPC power spectra for vowel /u/ from original waveform, System II and System III.

6 Conclusions

We present our HMM-based Text-to-Speech system for Mandarin Chinese synthesis in this paper. A tone-dependent phone set, Ph97, is employed in training HMMs with phonetic and prosodic question set in corresponding decision trees. We adopted LSP frequencies as acoustic spectral features for training HMMs. Subjective AB comparison preference test show that using LSPs and the dynamic features of adjacent LSPs in frequency considerably improve the quality of synthetic speech, in comparing with the conventional method.

References

[1] Zen, H. and Toda, T., An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005, Proc. EuroSpeech, 2005.

[2] Tokuda, K., Zen, H. and Black, A.W., An HMM-based speech synthesis system applied to English,' 2002 IEEE Speech Synthesis Workshop, Santa Monica, California, Sep. 11-13, 2002.

[3] Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., and Kitamura, T., Speech Parameter generation algorithms for HMM-based speech synthesis. Proc. ICASSP, pp. 1315-1318, Istanbul, Turkey, June 2000.

[4] Tomoki, T. and Keiichi, T., Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis, Proc. Eurospeech 2005.

[5] Kawahara, H., Masuda-Katsuse, I. and Cheveigne, A., 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.

[6] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., Hidden semi-Markov model based speech synthesis, Proc. ICSLP, 2004, pp. 1185–1180.

[7] Itakura, F., Line spectrum representation of linear predictive coefficients of speech signals, J. Acoust. Soc. Am. 57 (Apr. 1975), S35.

[8] Fukada, T., Tokuda, K., Kobayashi, T, and Imai,S., An adaptive algorithm for mel-cepstral analysis of speech, Proc. ICASSP, 1992, pp. 137-140.

[9] Soong, F. K., and Juang, B. H. Line spectrum pair (LSP) and speech data compression. Proc. ICASSP, pp.1.10.1-1.10.4.,San Diego, CA, 1984.

[10] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T., Multi-space Probability Distribution HMM, IEICE Trans. Inf. & Syst., E85-D(3):pp. 455-464, 2002.

[11] Shinoda, K. and Watanabe, T., Acoustic Modeling Based on The MDL Principle for Speech Recognition, Proc. EuroSpeech 1997, pp. 99-102.

[12] Wakita, H, Linear prediction voice synthesizers: line spectrum pairs (LSP) is the newest of the several techniques. Speech Technol. 1 (1981), pp.17-22

[13] Paliwal K. K. On the use of line spectral frequency parameters for speech recognition, Digital Signal Processing 2, pp 80-87 (1992)

[14] Chu, M., Peng, H., Yang, H. and Chang, E., Selecting non-uniform units from a very large corpus for concatenative speech synthesizer, Proc. ICASSP 2001,Salt Lake City.

[15] Huang, C., Shi, Y., Zhou, J. L., Chu, M., Wang, T., and Chang, E., Segmental Tonal Modeling for Phone Set Design in Mandarin LVCSR, Proc. ICASSP 2004, pp.901-904, 2004.

[16] Zen,H., Tokuda, K. and T. Kitamura, A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features, Proc. of ICASSP 2004, pp. 837–840.

[17] Wu, Y.J. and Wang, R.H., Minimum generation error training for HMM-based speech synthesis. Proc. of ICAPP 2006, pp. 89-93