A STUDY OF F0 MODELLING AND GENERATION WITH LYRICS AND SHAPE CHARACTERIZATION FOR SINGING VOICE SYNTHESIS

S. W. Lee, Minghui Dong, and Haizhou Li

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632
{swylee, mhdong, hli}@i2r.a-star.edu.sg

ABSTRACT

Natural pitch fluctuation is essential to singing voice. Recently, we have proposed a generalized F0 modelling method which models the expected F0 fluctuation under various contexts with note HMMs. Knowing that having F0 contours close to human professional singing promotes perceived quality, we are confronted with two requirements: (1) accurate estimation on F0 and (2) precise voiced/unvoiced decisions. In this paper, we introduce two techniques in the above directions. Influence of lyrics phonetics on singing F0 is considered to capture the F0 and voicing behaviour brought from different note-lyrics combinations. The generalized F0 modelling method is further extended to frequency-domain to study if shape characterization in terms of sinusoids helps F0 estimation or not. Our experiments showed that the use of lyrics information leads to better F0 generation and improves naturalness of synthesized singing. While the frequency-domain representation is viable, its performance is less competitive than time-domain representation, which requires further study.

Index Terms— singing, synthesis, pitch, modelling, lyrics

1. INTRODUCTION

Singing voice synthesis has become popular recently, in both research and industry [1]–[5], which offers innovative services and applications, such as entertainment on mobile devices, music production and computer-assisted vocal training [6], [7].

Similar to speech synthesis, singing synthesis starts from concatenation-based approaches [6], to adaptive voice generation and parameter estimation [4], [5]. Synthesized voices are expected to be high-quality natural singing, possibly with pleasant singing styles. It is also desirable to generate singing voice for an individual with his/her characteristics, vocal timbre, for instance. Consequently, everyone has his/her high-quality singing, no matter he/she is good at singing or not.

This study of F0 modelling is aimed at generating appropriate fluctuations of fundamental frequency for natural singing synthesis. A smooth and natural F0 movement is required by the speech-to-singing approach [8] where, with a given lyrics-reading speech, we would like to manipulate the respective F0 contour and spectral envelope to resemble singing, while preserving the speaker's characteristics of voice, which is also called timbre.

High-quality singing can be made by properly adjusting the above two attributes. In vocal music studies, pitch variation in a singing voice is always essential [9]. Various types of fluctuations, such as, vibrato, overshoot, and preparation, are found to contribute to the perceived singing quality [10]. These fluctuations are dependent and vary under assorted contexts. There are many research works on F0 generation for singing voice. Some of them determine the current type of F0 fluctuation, generate the corresponding F0 segments and join together [8]. Another work models F0 dynamics with a Gaussian processes-based parametric system [11]. All of these works model F0 fluctuation in time domain.

We recently proposed a generalized F0 modelling and generation method [12]. Given a sequence of music notes (describing the target melody), an F0 contour is estimated by maximizing the likelihood. A set of full-context note models is statistically built by learning the possible F0 behaviour over time, under various scenarios. This modelling method is generalized that the F0 contour with various fluctuations is generated as a whole. Explicit decisions on the parametric forms of models and existence of certain types of F0 fluctuations are bypassed. This time-domain, generalized F0 modelling has been shown to achieve accurate pitch generation and better naturalness of synthesized outputs, over other alternatives.

This paper extends the above F0 modelling method, by examining two techniques on natural singing generation. Having pitch fluctuation in synthesized singing close to human singing promotes the perceived quality; whereas proper voiced/unvoiced decisions play an important role as well. Our previous experiments on F0 modelling [12] indicated that the perceived quality is greatly suffered from voicing errors, even if the F0 values during voiced portions are accurately estimated. There are several studies [13], [14] showing the importance of voiced/unvoiced decisions as well.

Our first technique is on the use of lyrics for voicing decision. Depending on the statistics of the F0 observations, individual voicing patterns may be assigned to note models with distinct lyrics context. In other words, the aforementioned time-domain, generalized F0 modelling is now further extended by phonetic considerations learnt from singing voice training data.

Frequency-domain representation has been widely used for speech coding and recognition [15]. Speech spectrum and singing pitch contour are in common that the associated shape contributes to perceived content on speech or melody, and quality. The generalized F0 modelling method will be reformulated in frequency domain below (as our second technique), by characterizing the F0 contour shape in terms of sinusoids. The global trajectory and local fluctuation of F0 are captured consequently.

Similar frequency-domain approach has been attempted in [16], where singing style parameters, in particular, the vibrato in sustained (steady) F0 portion, are modelled. In comparison, our generalized F0 modelling works on the entire F0 contour, regardless of the type of F0 fluctuation. It also enables the

generation of F0 contour for unseen melody, which can be applied to imitating systems, such as [5]. Last but not least, experiments using an identical set of data on the proposed modelling techniques and the others will be reported here. We will also compare the performance from time-domain and frequency-domain representations, which is not often available in previous publications.

2. REVIEW ON GENERALIZED F0 MODELLING

Before moving to the two techniques on the use of lyrics and shape characterization, let's briefly review the generalized F0 modelling method [12]. It is the platform where the two proposed techniques applied to generate singing F0 contours. In our speech-to-singing synthesis system [12], this generated F0 contour is combined with sequences of estimated singing spectral envelope and aperiodicity. Tandem-STRAIGHT [17] is finally used to output the synthesized singing.

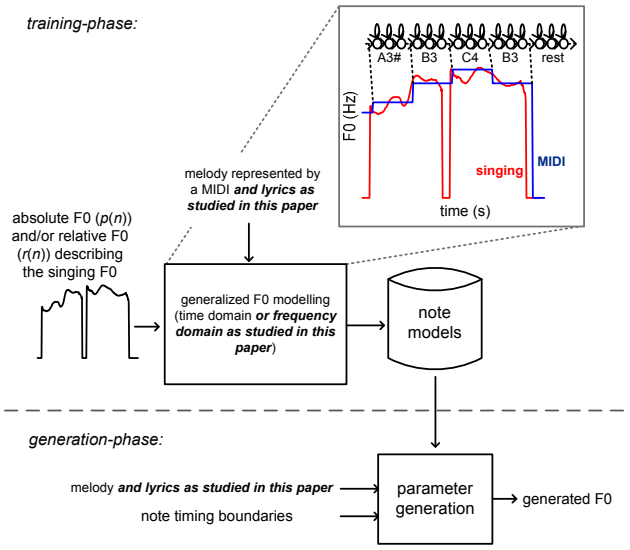

Fig. 1. Block-diagram of the generalized F0 modelling, where the introduced techniques applied to.

The generalized F0 modelling aims at learning the expected F0 behaviour with reference to the associated melody, as shown in Fig. 1. Models are built on a note basis, analogous with the phone models in speech recognition. The melody, as the input for F0 modelling, is often represented in terms of a MIDI file. By using the HMM-based Speech Synthesis System (HTS) [18], a set of context-dependent note and rest models, covering the desired range of singing F0 (note A1 to C6), is built. Specifically, these models are *M*-state single-mixture left-to-right hidden semi Markov models (HSMMs) [19]. They are trained in a way similar to the standard HTS process for speaker-dependent systems [18]. It is expected that within one note model, various F0 fluctuations may occur, depending on the training data. Overshoot and preparation are encoded in states at model boundaries; whereas vibrato is probably located at much latter states.

This modelling system has been implemented with two types of features, namely absolute F0 at time n ($p(n)$) and relative F0 ($r(n)$), leading to three systems (which will be described soon). For each type of feature, the feature vector consists of the static, delta and delta-delta features. As $p(n)$ and $r(n)$ are sometimes undefined, for example, during rest note and unvoiced singing, multi-space probability distribution HMM (MSD-HMM) [20] is used.

During training, we apply tree-based context clustering with the following context information: (1) note identity (previous, current and next note); (2) note interval relative to the current note in the unit of semitones (previous and next note); (3) note duration (previous, current and next note); and (4) tempo class of the song. They are extracted merely from the melody. Nevertheless, a MIDI file contains much information, besides the melody. Lyrics are some of the examples, which enable effective voicing decision, as shown in coming sections.

During generation, the F0 contour for a given segment is estimated. By a 'segment', we refer to a line of lyrics. Full-context note labels for this segment are first constituted. Model durations in the labels are specified by the time alignment of the note sequence. Then, the parameter generation algorithm for Case 3 in [21] is adopted to estimate the F0 contour. Based on the type of features used, there are three possible systems: (1) trained from $p(n)$; (2) trained from $r(n)$; and (3) fusing the generated F0 contours from the two systems above to form an integrated F0 contour [12].

3. USE OF LYRICS FOR VOICING DECISION

The voiced/unvoiced decision on generated F0 contours is influential in perceived singing quality. In Mandarin Chinese, the voicing pattern of a speech signal is governed by the content and the phonetic properties of the language [22]. These motivate the first technique introduced here. Since the voicing pattern of singing voice may not be the same as the one for speech, the model parameters and voiced/unvoiced weights in MSD-HMM note models will be learnt from singing observations. Phonetic considerations of lyrics, with reference to Mandarin Chinese, are made as follows.

Given a line of lyrics, a number of factors contribute to this voicing property observed over time. Take an example, the initial of a word of lyrics spanning two consecutive notes occurs only in the first note, rather than at the beginning of both notes. Hence, the voicing pattern of a word of lyrics spanning two notes is different from the case for two words of lyrics spanning identical notes. Another example is nasals and voiced fricatives of some Chinese initials.

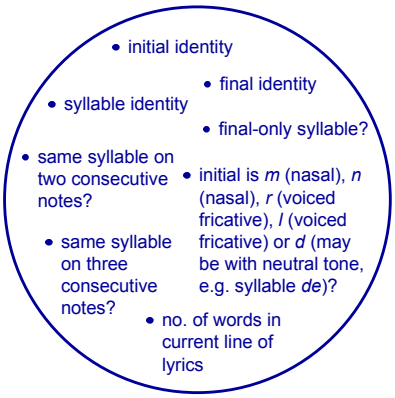

Fig. 2 Questions related to lyrics information used during context clustering.

The full-context label of a note sequence now consists of the following lyrics information, besides those listed in Section 2: (5) number of words in the current line of lyrics; (6) initial identity

(previous, current and next note); (7) final identity (previous, current and next note). Questions regarding the phonetic and other relevant properties of lyrics are used during tree-based context clustering. They are depicted in Fig. 2.

4. CHARACTERIZING F0 CONTOUR SHAPE IN FREQUENCY-DOMAIN

We now describe the proposed frequency-domain modelling and generation for F0 contours, which aims for capturing the natural F0 fluctuation. This representation can be applied to either the system trained on $p(n)$ or the system trained on $r(n)$. The following illustration will be on $p(n)$, while the same procedures can be applied to $r(n)$ as well.

Fig. 3 illustrates our frequency-domain F0 representation and reconstruction scheme. The overlap-add approach [23] is used. Given an F0 contour $p(n)$ for training, it is first interpolated across MIDI rest notes. This is to ensure the generated F0 contour during generation phase to be smooth and with negligible glitch. Overlapped segments of the interpolated F0 contour are then windowed (as frames) and converted to frequency domain by discrete cosine transform (DCT). DCT has been adopted here, based on its strong energy compaction property and efficient modelling of pitch contour for speech synthesis [24]. Very few numbers of DCT coefficients are usually needed. Likewise, these coefficients are real, which means at least half amount of memories is saved, compared to fast Fourier transform. Compact fluctuation modelling can thus be achieved, since a vector of DCT coefficients represents certain fluctuation over time, even consecutive singing frames are assigned to the same model state.

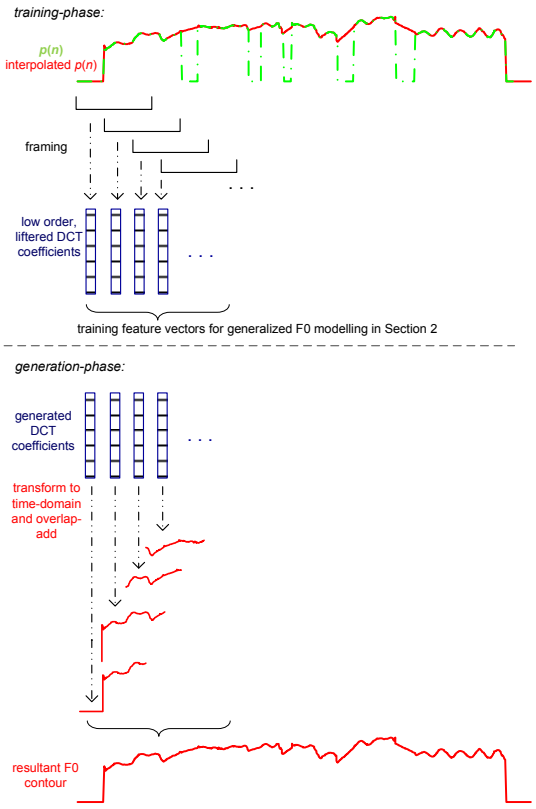

Fig. 3 Block-diagram of the proposed frequency-domain representation and generation.

Let $x(k, t)$ be the k-th order DCT coefficient of time t, $0 \le k \le N-1$, calculated as follows,

$$x(k,t) = w(k)\sum_{n=0}^{N-1} p_t(n)\cos\frac{k\pi(2n+1)}{2N}, \quad (1)$$

where $p_t(n)$ is the frame windowed at time t. N is the length of $p_t(n)$ and

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le k \le N-1 \end{cases}. \quad (2)$$

Only DCT coefficients with order up to L are kept. They essentially represent the global trajectory for the melody and local fluctuation in natural F0 contours. As the 0-th order DCT coefficient is numerically larger than the others, leading a wide range of variances when going from 0-th to high order coefficients, liftering [25] is performed. Feature vectors, comprising the resultant DCT coefficients, are now ready for training. Since the values in feature vectors are always defined, conventional HMM is used for this frequency-domain modelling, rather than MSD-HMM.

During generation, vectors of DCT coefficients $\hat{x}(k,t)$ are estimated. The coefficients are then unliftered and transformed back to time-domain by

$$\hat{p}_t(n) = \sum_{k=0}^{N-1} w(k)\hat{x}(k,t)\cos\frac{k\pi(2n+1)}{2N}. \quad (3)$$

A preliminary contour is obtained by (weighted-) summing these time-domain sequences $\hat{p}_t(n)$. The summing weights are determined based on the windowing function, N and the overlapping portion between frames [23]. Since this contour from frequency-domain representation is always defined, without unvoiced portions as the one generated from time-domain representation, the voicing decisions from time-domain representation is combined with the resultant contour to form the final F0 contour.

5. EXPERIMENTS

The performance of the proposed techniques on generating natural singing F0 is reported in the following. The experiments were performed using the same set of data, for ease of comparison across different systems. We evaluated our system performance from two perspectives: (1) objective measurement of the errors in generated F0 contours; and (2) subjective listening test on the naturalness of synthesized singing using the generated F0 contours. A medium size collection of solo singing recordings from a male professional singer was used for the studies below. There were altogether 50 Mandarin Chinese pop songs. Each song lasted about four minutes, totalling 194 min 33 sec. The MIDI files for these songs were used as the inputs for F0 modelling and generation. There were 1996 segments for training and 54 segments for testing in total. These testing segments were unseen from training. $p(n)$ and $r(n)$ were extracted for every 5 msec.

With this collection of singing recordings, a set of context-dependent models, covering the frequency range from D2# (77.82 Hz) to D5 (587.33 Hz) and rest notes, were trained. Each contained five emitting states. For systems with frequency-domain modelling, frame length was 50 msec with a shift of 5 msec.

Hamming window was used. L was set to four. These were empirically determined by a small data set that was unseen from the testing set.

5.1. Objective Measurements of F0 Generation Accuracy

The measurements on various types of error of F0 generation are given in Table 1. The root-mean-square-error (RMSE) measurement was applied on voiced singing frames with correct voicing in the generated F0. It is in the unit of cent, a logarithmic measure used for distance between two frequencies in music theory. If there is any voicing error, i.e. a voiced frame treated as unvoiced (E_{10}), or an unvoiced frame treated as voiced (E_{01}), the corresponding voicing error is counted. As the fusion system using $p(n)$ and $r(n)$ was found to achieve the best RMSE and E_{10} [12] , it was used for this study. There were four systems (A, B, C and D), comparing the effectiveness of the proposed two techniques.

system	error type		
	RMSE (cents)	E_{10} (%)	E_{01} (%)
A. fusion system, without lyrics, time-domain	139.67	10.09	**27.38**
B. fusion system, with lyrics, time-domain	**138.51**	9.52	27.40
C. fusion system, with lyrics, freq.-domain (voicing decisions were from System B)	139.23	-	-
D. 2nd order damping system [8]	146.57	**3.68**	59.48

Table 1. Measurements of F0 accuracy from various systems.

The best results under individual error types are bolded. It was found that System B achieved the best RMSE. System D has the lowest E_{10}. This is expected as it always generates voiced frames, except during rest notes. E_{01} from System A and B were very close to each other. Among the four systems, the worst RMSE and E_{01} were from System D. By comparing system A and B, the proposed first technique using the lyrics information helped to reduce RMSE and E_{10}, generating F0 contours closer to the reference singing. Comparison between System B and C, time-domain representation achieved lower RMSE.

The influence of lyrics information on context clustering can be seen in the trees as well. In the tree of the first model state trained by $p(n)$, questions regarding lyrics were found to appear starting from the seventh branch, where there were 26 branches for the lowest leaf node. The first of these questions was about the existence of nasal or voiced fricative on the previous note. The second one was whether the final on the next note is *a* or not. Similar situations were seen in the trees built in models of $r(n)$.

5.2. Subjective Listening Tests

Two subjective listening tests were conducted. The first one verified if the proposed use of lyrics information promotes the naturalness of synthesized singing or not. System A and B were under comparison. To focus on the influence of F0 generation on synthesized singing, the spectral envelopes and aperiodicity functions extracted from reference singing voice remained unchanged and used for speech-to-singing synthesis for all systems. Listeners were asked to choose the system with higher naturalness. This test consisted of 12 questions, taken from the testing set. Each contained one line of lyrics. Listeners could play the stimuli and many times as they wished. A total of 38 subjects participated, contributing 456 responses.

Fig. 4 (left figure) shows the naturalness result. About 72% of the responses indicated that the proposed use of lyrics information for F0 modelling generates higher naturalness. This showed that better performance could be achieved by the same amount of training data, but having additional context labels.

The second listening test studied the effectiveness of the proposed frequency-domain modelling, against alternative systems. Four systems were compared. They are: (O) recorded singing voice with Tandem-STRAIGHT analysis and synthesis (no modification on singing F0); and System B, C and D in Section 5.1. Listener were asked to compare and rate the naturalness of the four systems by mean opinion score (MOS). Possible MOS ranged from 1 (bad) to 5 (excellent). There were 20 questions in total. 33 subjects participated, contributing 660 responses.

The box plot of the results is depicted in Fig. 4 (right figure). On each box, the central mark is the median. The edges are the 25th and 75th percentiles. Outliers are indicated by plus symbols. The horizontal lines represent the most extreme points not considered outliers. The experimental results showed that the naturalness of System C (frequency-domain modelling) was significantly better generation using System D (second order damping system) with 95% confidence. Nevertheless, performance of System B (time-domain modelling) was found to be significantly better than System C (frequency-domain modelling). This indicates that frequency-domain modelling and generation is effective, but further extension is required to offer comparable performance as the time-domain approach.

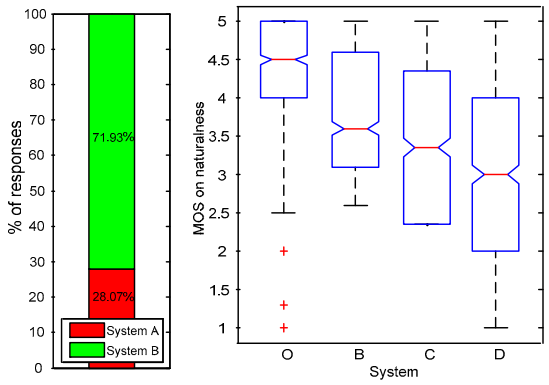

Fig. 4. (left) Results on naturalness from systems with and without using lyrics information. (right) Box plot of the MOS naturalness results from System O, B, C and D.

6. CONCLUSIONS

Synthesized singing has to be natural and high-quality. Among the 'sources of naturalness' in human professional singing, F0 fluctuation is one of the most dominant attributes, governing the perceived quality. This paper introduces two techniques on F0 modelling and studies their performance on a note HMM-based framework. Lyrics information is considered. Specifically, F0 and voiced/unvoiced behaviour under various contexts are now modelled with respect to phonetic attributes. Frequency-domain F0 modelling is another technique, which represents the F0 contour of a given melody in terms of a few sinusoids. We evaluated the above two techniques and found that: (1) The use of lyrics information is effective, leading to better F0 generation and synthesis; (2) The frequency-domain representation is feasible, but not as good as the time-domain counterpart.

7. REFERENCES

[1] *Synthesis of Singing Challenge* (*Special Session*), *Proc. Interspeech*, Aug. 2007.

[2] M. Akagi, "Rule-based voice conversion derived from expressive speech perception model: How do computers sing a song joyfully?" in *Proc. ISCSLP*, Tutorial 01, Nov. 2010.

[3] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. Interspeech, pp.* 2274-2277, Sep. 2006.

[4] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, "Pitch adaptive training for HMM-based singing voice synthesis," in *Proc. ICASSP*, pp. 5377-5380, Mar. 2012.

[5] M. Goto, T. Nakano, S. Kajita, Y. Matsusaka, S. Nakaoka, and K. Yokoi, "Vocalistener and vocawatcher: Imitating a human singer by using signal processing," in *Proc. ICASSP*, pp. 5393-5396, Mar. 2012.

[6] H. Kenmochi and H. Ohshita, "VOCALOID – Commercial singing synthesizer based on sample concatenation," in *Proc. Interspeech*, Aug. 2007.

[7] P. Kirn (2009, Oct. 6). iPhone Day: LaDiDa's Reverse Karaoke Composes Accompaniment to Singing [Online]. Available: http://createdigitalmusic.com/2009/10/iphone-day-ladidas-reverse-karaoke-composes-accompaniment-to-singing/

[8] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 215-218, Oct. 2007.

[9] M. B. Dayme, *Dynamics of the Singing Voice*, 5th ed. New York: Springer, 2009.

[10] T. Saitou and M. Goto, "Acoustic and perceptual effects of vocal training in amateur male singing," in *Proc. Interspeech*, pp. 832-835, Sep. 2009.

[11] Y. Ohishi, H. Kameoka, D. Mochihashi, H. Nagano, and K. Kashino, "Statistical modeling of F0 dynamics in singing voices based on Gaussian processes with multiple oscillation bases," in *Proc. Interspeech*, pp. 2598-2601, Sep. 2010.

[12] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modeling with absolute and relative pitch features for singing voice synthesis," in *Proc. ICASSP*, pp. 429-432, Mar. 2012.

[13] K. Yu, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, & Lang. Proc.*, vol. 19, pp. 1071-1079, Jul 2011.

[14] J. Latorre, M. J.F. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification," in *Proc. ICASSP*, pp. 4724-4727, May 2011.

[15] L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition*, New Jersey: Prentice Hall, 1993.

[16] K. Saino, M. Tachibana, and H. Kenmochi, "A singing style modeling system for singing voice synthesis," in *Proc. Interspeech*, pp.2894-2897, Sep. 2010.

[17] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. ICASSP*, pp. 3933-3936, Mar. 2008.

[18] H. Zen, K. Oura, T. Nose, j. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proc. APSIPA ASC*, pp. 121-130, Oct. 2009.

[19] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, pp. 533-543, May 2007.

[20] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst*., vol. E85-D, pp. 455-464, Mar. 2002.

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

[22] C.-C. Kuo, "Phonetic and phonological background of Chinese spoken languages," in *Advances in Chinese Language Processing*, C.-H. Lee, H. Li, L.-S. Lee, R.-H. Wang, and Q. Huo, Eds. World Scientific Publishing Co. Pte. Ltd., 2007, ch. 2, pp. 33-55.

[23] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, pp. 1558-1577, Nov. 1977.

[24] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Proc. Interspeech*, pp. 2274-2277, Sep. 2008.

[25] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, 2001.