PERSONALIZED NATURAL SPEECH SYNTHESIS BASED ON RETRIEVAL OF PITCH PATTERNS USING HIERARCHICAL FUJISAKI MODEL

Yi-Chin Huang, Chung-Hsien Wu, and Shih-Lun Lin

National Cheng-Kung University
Dept. of Computer Science and Information Engineering
Tainan, Taiwan

ABSTRACT

In recent years, speech synthesis based on Hidden Markov Model (HMM) has been developed, which can synthesize stable and intelligible speech with flexibility and small footprint. However, synthesized prosodic features are still incapable to convey personalization and natural property. Previous prosody models, mainly constructed from the clustered prosodic features, are unable to characterize personalized prosodic information as the linguistic cues of the input sentence are indistinguishable for all speakers. An approach to retrieval of personalized pitch patterns from the real speech corpus of the target speaker, is proposed, incorporating with the HMM-based speech synthesizer, to generate a personalized natural pitch contour. The modified Fujisaki model is adopted to depict the hierarchical pitch patterns, aiming to model local pitch contour variation and global intonation of utterances in the corpus. The codeword sequences of utterances in the training and the synthesized corpora are constructed and used to obtain the relationship of pitch patterns between the real and synthesized speech. Finally, a language model of pitch pattern is constructed to obtain an optimal pitch pattern sequence of the input sentence. The experimental results using subjective and objective evaluations demonstrated the proposed approach can substantially outperform the conventional statistical synthesis methods, in terms of naturalness and speaker similarity.

Index Terms— Fujisaki Model, Hierarchical Prosodic Structure, Pattern Retrieval, Personalized Speech Synthesis

1. INTRODUCTION

There are two research mainstreams for generating intelligible and decent quality speech, namely concatenation-based method [1] [2] and statistical methods [3] in the nowadays technologies of speech synthesis. Among the statistical methods, HMM-based speech synthesis method provides a unified framework, which consists of three different acoustic features: spectrum, duration, and fundamental frequency for speech synthesis. With high portability and flexibility, HMM-based synthesis models can be adapted to either speakers [4], emotions [5], or languages [6] [7] of interests. In addition, voice conversion methods have been proven useful to convert the speech signals for different speaking styles, speakers, or emotions using limited speech data [8]. In recent years, constructing a personalized speech synthesizer with similar speech timbre and prosodic characteristics is becoming an attractive research topic [9]. For personalized speech synthesis, an intuitive way is to collect speech data of the target speaker for personalized model construction. However, synthesized speech generated by HMM-based acoustic model has the over-smoothing problem of acoustic features. The resultant synthesized speech is steady but less dynamic in terms of perceived prosody comparing with synthesized speech generated by concatenation-based method, which preserves speech rate and pitch variation from the original speech. Thus, in this study, the proposed prosody model is aimed to adopt the prosodic information in the original speech data to preserve the personalized prosodic features. A novel concatenation-based method is employed to retrieve suitable pitch patterns by considering the relationship between the real and synthesized speech. Besides, a hierarchical prosodic-level clustering method is used to characterize local pitch variation and global intonation in a synthesized speech sentence. In order to maintain the dynamic range of prosodic information, the prosodic units are modeled by the modified Fujisaki commands. Finally a two-level dynamic programming algorithm is employed to select the optimal pitch pattern sequence for the synthesized speech considering not only linguistic information, but also the language model of pitch patterns and the continuity of concatenated pitch contour.

This rest of paper is organized as follows. The proposed system framework is introduced in Section 2. Section 3 introduces the two-level selection algorithm of pitch patterns and re-synthesis module of the HMM-based speech synthesis system. The subjective and objective evaluations of the proposed system comparing with the original HMM-based system are presented in Section 4. Lastly, concluding remarks are given in Section 5.

2. HIERARCHICAL PERSONALIZED PROSODIC MODEL

Previous studies on prosodic feature generation used the clustered prosodic features to construct a prosody model. A regression-based clustering method for GMM-based prosody conversion was proposed in [10]. The prosody hierarchy and dynamic pitch features were used in [11] for pitch modeling in HMM-based speech synthesis. Another research interest is to find the relationship between different prosodic units and their behaviors in terms of pitch contour to generate natural prosody of synthesized speech [12]. For Mandarin speech, the prosodic structure is generally characterized by a hierarchical prosodic model [13]. An ANN-based method for pitch contour generation of Mandarin speech was proposed in [14]. As discussed in these studies, the prosodic structure, representing the short stops or pauses in a sentence, is imperative for the naturalness and intelligibility of the synthesized speech, and the prosodic variations in a fluent utterance are also affected by the phrase and sentence types. So, it could be useful if the prosodic units, such as prosodic word (PW) and prosodic phrase (PP), are used as the basic unit to characterize local pitch variation and global intonation of an utterance, respectively. In Mandarin, a PW is a sequence of

syllables and a PP is a sequence of PWs. Distinct pauses can be observed between PPs and shorter pauses appear at the boundaries of every two PWs. This study combines the HMM-based Mandarin speech synthesis and a novel pitch pattern retrieval method to obtain a personalized natural pitch pattern sequence.

2.1. Conventional HMM-based Mandarin Speech Synthesis

For system construction, a conventional HMM-based speech synthesis system of a specific speaker is firstly constructed [15]. In this framework, spectrum, pitch, and state duration of speech signals are modeled simultaneously. In the constructed Mandarin speech synthesis system, each Mandarin syllable is composed of three sub-syllables which are used as the basic synthesis units.

$$C + V_1 + V_2, \tag{1}$$

The first sub-syllable, C, is an extended initial portion of a syllable, and is followed by two final tonal sub-syllables, V_1 and V_2. The initial sub-syllable C is generally a consonant, and the sub-syllables V_1 and V_2 compose the rhyme part of a syllable. The five lexical tones of a Mandarin syllable are encoded as the combination of high/middle/low (H/M/L) categories of sub-syllables in the rhyme part, according to the lexical tone of the final portion. The lexical tones can be represented as Tone 1 (high): HH, Tone 2 (rising): LH, Tone 3 (low): LL, Tone 4 (falling): HL, and the neutral tone: MM [16]. The training and synthesis procedures follows the framework proposed in [10].

2.2. Hierarchical Prosodic Structure

2.2.1. Hierarchical Prosodic Units

Because prosodic features are essential for perceiving the personalized speech characteristics, such as phone length, pitch, and prosodic word/phrase boundaries. In Mandarin speech, the prosody of an utterance can be modeled by a hierarchical structure, and the prosodic variations in the speech utterances of different speakers are resulted from the semantic meanings and the prosodic word/phrase boundaries in an utterance. Thus, it is crucial to model different prosodic units and the prosodic variations between and within these units. In corpus analysis of the collected speech corpora of Mandarin speech with manually transcribed prosodic boundaries by linguist experts, the length of an utterance is approximately 1 to 25 syllables, and an utterance usually consists of one to five PPs. This phenomenon is consistent with the description in previous studies [17]. Based on the observation, syllable is regarded as the basic and low-level unit, PW is regarded as the higher-level unit, and PP is the highest-level unit in the prosodic structure of an utterance. As described in previous research, each prosodic unit has its contribution to the observed pitch contour. Thus, in this study, each prosodic unit is modeled, and the original pitch contour can be regenerated by the superposition of prosodic units at different levels.

2.2.2. Hierarchical Prosodic Command Generation Based on Modified Fujisaki Model

To regenerate the pitch contour by the superposition of different-level prosodic units, a model which can depict the partial contribution of pitch contours at different prosodic levels is required. In this study, the well-known Fujisaki model [18] is adopted for prosodic unit modeling. As illustrated in Fig. 1, Fujisaki model is designed to generate pitch contour by combining the three components. The

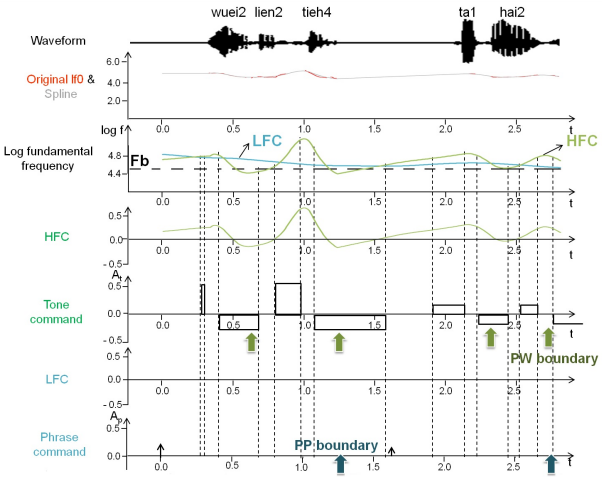

Fig. 1. An example of waveform and its HFC, LFC, and Fb with the corresponding tone command of PW and phrase command of PP generated by the modified Fujisaki model.

first one is Low Frequency Component (LFC), which represents the global movement of input pitch contour and is modeled by phrase commands. The second one is High Frequency Component (HFC), which represents the subtle, local variation of the input pitch contour and is modeled by accent/tone commands. The last one is Base Frequency (Fb), which represents the basis frequency value of pitch contours of all utterances and varies according to the speaker characteristics. In this study, the original Fujisaki model is modified in some aspects to fit the hierarchical prosodic units, as showed in Fig. 1. First, the positive and negative commands are adopted for the tone component to simulate the rising and falling tone more accurately, and the sequence of tone commands in one PW is calculated. Second, for each PP, instead of calculating phrase command of the entire utterance, the phrase command sequence in each PP of an utterance is calculated to model the intonation of each PP.

2.2.3. Hierarchical Prosodic Unit Clustering

For selecting a suitable sequence of prosodic units, prosodic unit clustering is useful for reducing search time and improving selection accuracy. Since the pitch contour of a Mandarin syllable can be modulated by its tone, lexical tone is considered as a clustering criterion. Besides, other essential features, such as the number of syllables and the length of a PW, are also considered. Other than these linguistic and prosodic features, the similarity between pitch contours of prosodic units is used for pitch contour clustering. In this study, a two-level prosodic unit clustering algorithm is applied. First, syllables and PWs are categorized by their linguistic and prosodic information. Subsequently, in the same linguistic category of the first level, an iterative K-means clustering algorithm based on Euclidean distance between length-normalized pitch contours is adopted. The splitting and stopping criterion is set to a variance threshold to control the size of each cluster. Finally, each resultant cluster is called a "PW-codeword" to represent a clustered prosodic unit. For PPs, in the first level clustering, the position and the length of a PP in an utterance are used as the clustering criterion. The same k-means clustering algorithm based on its shape is adopted. Finally, the "PP-codeword" can be obtained. For the second level clustering, spline

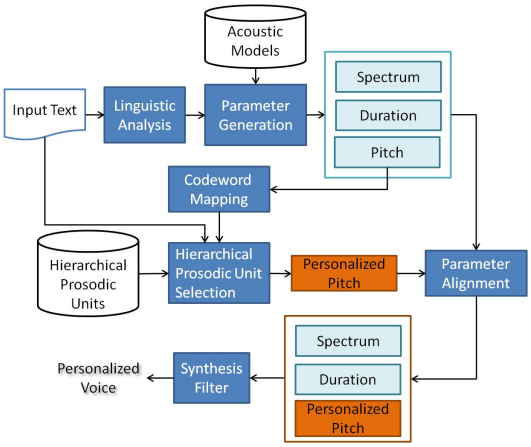

Fig. 2. The proposed method integrated with an HMM-based Mandarin TTS system.

interpolation is adopted to fill the unvoiced part of each syllable. The codeword sequence of each training utterances can also be generated. These PW and PP codewords are further adopted to calculate the statistical distribution in the collected corpus in order to find the relationship among codewords.

3. PROSODIC UNIT SELECTION ALGORITHM

The main idea in the proposed method compared to the conventional unit selection method is that the proposed method adopts the pitch contour of the speech synthesized by the HMM-based speech synthesis system as the information to retrieve suitable pitch patterns of the real speech. First, a pseudo corpus generated by the HMM-based synthesizer is constructed. Each utterance in the pseudo corpus is forcedly aligned to the corresponding utterance in the original training corpus, so the prosodic structure of each utterance is the same as the one in the training corpus. The codewords of the prosodic units in the pseudo corpus can be clustered in the same manner as described in Section 2.2.3. Codeword mapping between the original training corpus and the pseudo corpus can be obtained easily. For each input sentence, the linguistic information and synthesized speech are generated first. The PW and PP command sequences of the synthesized speech are obtained by finding the most suitable codeword clusters, respectively. Both command sequences are adopted as a query to search the related utterances in the synthesized pseudo corpus, and the query is formed as the vector representation for retrieving similar utterances in the pseudo corpus, which is defined as follows.

$$(c_1, c_2, \cdots, c_i, \cdots, c_n, c_1c_1, c_1c_2, \cdots, c_nc_n) \tag{2}$$

where c_i is the term frequency of the i-th codeword and n is the number of codewords. The term frequencies of unigram and bigram of the codewords are estimated. The same vector representation for the utterances in the pseudo corpus are also constructed. The cosine measure is applied to rank each utterance in the pseudo corpus given the input utterance. After ranking all utterances in the pseudo corpus, the candidate synthesized codewords of PWs and PPs are selected based on their ranking order, respectively. The codeword mapping procedure is then adopted to map each codeword of the pseudo corpus to its corresponding codeword in the same utterance of the original training corpus. In this manner, the candidate codewords of the training corpus are acquired for each pitch pattern of the input utterance. Besides, the number of each codewords of the training corpus for each pitch pattern of the input utterance is expanded because the mapping is one-to-many.

Furthermore, a codeword language model is used to find the most optimal codeword sequence for PWs and PPs. Note that the codeword language model is trained based on the distribution of PW and PP codewords in the training corpus, the optimal codeword sequence should be more similar to the pitch pattern sequence of the target speaker. The optimal codeword sequence is selected based on the following equation.

$$\hat{c}_1^L = \arg\max_{c_1^L} \left\{ Lan(c_1^L, u_1^L) \right\} \tag{3}$$

where

$$Lan(c_1^L, u_1^L) = \sum_{i=1}^{L} Uni\,(i_i, c_i) + \sum_{i=2}^{L} Bi\,(c_{i-1}, c_i) \tag{4}$$

and $Uni\,(u_i, c_i)$ is the unigram probability of codeword c and $Bi\,(c_{i-1}, c_i)$ is the bigram probability of (c_{i-1}, c_i). After the optimal codeword sequences is selected, pitch contour regeneration is achieved by selecting suitable pitch patterns in the selected codeword clusters. Each pitch pattern represented as a command stored in the codeword clusters, so the pitch contour can be regenerated by applying the synthesis filter function of the Fujisaki model. A suitable pitch pattern of each codeword is selected based on the continuity at the boundaries between the nearby pitch patterns, which is described in the following Section 3.1.

3.1. Continuity Measure

The continuity measure of the proposed method considers three different measures. The first one is the conventional distance measure of the connecting points between two pitch contours, which is trivial and may not be enough to measure the difference between the average ranges of pitch values of nearby pitch patterns. Thus, the distance of the mean of nearby pitch patterns is also considered. However, besides these two intuitive distance measure, the continuity of two pitch contours is also considered, which shows the smoothness of the concatenated pitch contour. For smoothness issue, the non-uniform rational B-spline (NURBS) method, which is commonly used to generate a continuous surface of the connected control points, is adopted. So, the contour generated by the NURBS between two pitch patterns is regarded as a smooth contour. If the concatenated pitch contour is similar to the NURBS contour, it means the continuity of the selected prosodic units can re-generate a smooth pitch contour. The continuity cost $S\,(a_i, b_{i+1})$ between pitch contour a_i at time i and pitch contour b_{i+1} at time $(i+1)$ is calculated as follows:

$$S\,(a_i, b_{i+1}) = \int_{k=0}^{1} \frac{\sum_{j=0}^{N^C} B_{j,n}^{(2)}\,(k)\, w_j^i p_j^i}{\sum_{j=0}^{N^C} B_{j,n}^{(2)}\,(k)\, w_j^i} \tag{5}$$

where $B_{j,n}^{(2)}\,(k)$ is the 2-nd derivative of the n-th degree B-spline basis function, $K = \{k_o, ..., k_{N^k}\}$ is the knot vector, p_i denoting a control point, and w_i denotes the corresponding weight. $N^k + 1$ and $N^C + 1$ are the number of knot points and control points, respectively. The steps of calculating the difference between the concatenated pitch contour and the NURBS contour are listed as follows:

1. Manually select two cut points for calculating the NURBS contour. The first cut point belongs to the tail region of the first pitch contour, while the second cut point is selected from the head region of the second pitch contour.
2. Generate the ideal pitch contour of the selected cut points of the two prosodic units using the NURBS curve.
3. Calculate the Euclidean distance between the original concatenated pitch contour and the ideal contour. This distance is regarded as the continuity cost for the two pitch patterns.

A linear combination of these three distance measures is adopted to find the optimal sequence of pitch patterns. Because the ranges of the three distance values are different, a min-max normalization is adopted to keep each distance value to lie between 0 to 1. The empirical weight for each distance is obtained by the inside test.

4. EVALUATION

4.1. Speech Data Collection and Experimental Setup

For the evaluation of the proposed method, the read speech of a female target speaker (speaker FR00) in the Tsing Hua Corpus of Speech Synthesis (TH-CoSS) [19] was used as the data set, which had been used for constructing an HMM-based Mandarin speech synthesis system [10]. This Mandarin data set consists of $5,406$ utterances with $98,749$ syllables. Note that a large size of data of the target speaker to cover the personalized pitch pattern variations is preferred. We used the last 406 utterances for outside objective test, and the other utterances for training the proposed model.

For pitch pattern clustering, the stoping criterion is that the number of pitch patterns should be greater than 10 (or 3) in each PW (or PP) cluster and the variance of the clustered pitch patterns should be smaller than a threshold. The resultant numbers (the average number in a codeword cluster) of the PW and PP codewords are $2,381$ (10.33) and $2,598$ (3.23) for the pseudo corpus, respectively. The number of codewords for the training corpus is $2,329$ (10.56) and $2,592$ (3.24) for PW and PP, respectively.

4.2. comparison with Conventional HMM-based method

In order to compare with the conventional HMM-based method, we integrated the proposed method with an HMM-based Mandarin TTS system, as shown in Fig. 2. The generated pitch contour is aligned to the pitch contour of the HMM-based model for each pitch pattern in the synthesized utterance. For example, for each PW, the generated pitch contour is scaled to match the length of the synthesized speech while maintaining the original length ratio of syllables in each PW.

For objective evaluation, we used 406 utterance to calculate the Root Mean Square Error (RMSE) of the pitch contour between two methods for comparison to the real utterances. The mean and variance of RMSE of the proposed method is 0.3 and 0.02, which outperforms the HMM-based method, which is 1.4 and 0.12.

The 5-point Mean Opinion Score (MOS) subjective evaluation was conducted to compare the effectiveness of the proposed method and the HMM-based method. There were 10 speech utterances, selected randomly from the daily newspapers, synthesized by the proposed method and the HMM-based method. Ten native subjects were invited to participate in the evaluation. The participants were asked to score speech quality for each utterance.

The ABX preference test was also conducted to compare the speaker similarity for different methods. Subjects were presented some utterances of the target speaker as the reference, and then a pair of synthesized utterances was presented. Subjects were asked to select which utterance is more similar to the target voice, especially in terms of pitch variation.

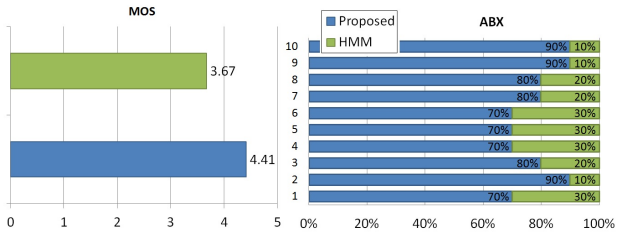

Fig. 3. Left plot: speech quality comparison results of MOS test. Right plot: Speaker similarity results of ABX test (The average preference of the proposed method is 79%).

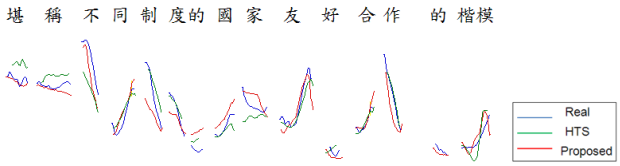

Fig. 4. Pitch contours of an outside utterance "堪稱不同制度的國家友好合作的楷模" generated by different methods.

The results of MOS evaluation are shown in the left plot of Fig. 3. The proposed method achieved better scores to that of the HMM-based method, and from the feedback from subjects, the prosodic perception of the synthesized utterances from the proposed method were more natural than those from the HMM-based method. The right plot of Fig. 3 shows the outside test results of ABX evaluation, which shows that the generated speech of the proposed method is perceived more similar to the speech of the training corpus than that of the HMM-based method. Fig. 4 shows an example of the pitch contour for an outside utterance "堪稱不同制度的國家友好合作的楷模 (kan1 cheng1 bu4 tong2 zhi4 du4 de5 guo2 jia1 you3 hao3 he2 zuo4 de5 kai3 mo2)", which shows the pitch contour generated by the proposed method is similar to the real pitch contour and its dynamic range is larger than that from the HMM-based method.

5. CONCLUSIONS

A hierarchical pitch pattern retrieval is proposed to generate personalized speech with natural pitch contour. With a modified Fujisaki model, the pitch contour can be regenerated by hierarchical prosodic commands and preserves pitch dynamic range. The proposed prosodic codeword retrieval is useful for selecting suitable pitch patterns, while the pitch contour generated by the HMM-based method has the over-smoothing problem of generating natural pitch contour. With the proposed pitch pattern retrieval method, combining with a HMM-based TTS system, the resultant pitch contour is more similar to that of the real speech while maintaining the speech quality. Future work will be focused on adding natural spectral information into the proposed pitch contour for further improvement of speech perception.

6. REFERENCES

[1] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, april 2007, vol. 4, pp. IV–1229 –IV–1232.

[2] C.-H. Wu and J.-H. Chen, "Automatic generation of synthesis units and prosodic information for chinese concatenative synthesis," *Speech Communication*, vol. 35, no. 3 - 4, pp. 219 – 237, 2001.

[3] K. Tokuda, H. Zen, and A.W. Black, "An hmm-based speech synthesis system applied to english," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, September 2002, pp. 227 – 230.

[4] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proceedings of Fourth International Conference on Spoken Language*, October 1996, vol. 2, pp. 1137–1140.

[5] C.-C. Hsia, C.-H. Wu, and J.-Q. Wu, "Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1245 –1254, September 2007.

[6] Y. Qian, J. Xu, and F.K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 5120 –5123.

[7] C.-P. Chen, Y.-C. Huang, C.-H. Wu, and K.-D. Lee, "Cross-lingual frame selection method for polyglot speech synthesis," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, march 2012, pp. 4521–4524.

[8] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-hmms for expressive speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, July 2006.

[9] Y.-C. Huang, C.-H. Wu, and Y.-T. Chao, "Personalized spectral and prosody conversion using frame-based codeword distribution and adaptive crf," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 51–62, January 2013.

[10] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in hmm-based speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1994 –2003, November 2010.

[11] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1394–1405, August 2010.

[12] M. Chu, H. Peng, H.-Y. Yang, and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 785–788.

[13] F.-C. Chou, C.-Y. Tseng, and L.-S. Lee, "Automatic generation of prosodic structure for high quality mandarin speech synthesis," in *Proceedings of Fourth International Conference on Spoken Language*, oct 1996, vol. 3, pp. 1624–1627.

[14] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for mandarin speech," *The Journal of the Acoustical Society of America*, vol. 117, no. 2, pp. 908–925, 2005.

[15] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proceedings of ISCA SSW6*, August 2007.

[16] C. Huang, Y. Shi, J. Zhou, M. Chu, T. Wang, and E. Chang, "Segmental tonal modeling for phone set design in Mandarin LVCSR," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 901–904.

[17] F.-Z. Liu, H.-B. Jia, and J.-H. Tao, "A maximum entropy based hierarchical model for automatic prosodic boundary labeling in mandarin," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, December 2008, pp. 1–4.

[18] H. Fujisaki and H. Kawai, "Realization of linguistic information in the voice fundamental frequency contour of the spoken japanese," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1988, pp. 663–666.

[19] L. Cai, D. Cui, and R. Cai, "TH-CoSS, a mandarin speech corpus for tts," *Journal of Chinese Information Processing*, vol. 21, no. 2, pp. 94–99, 2007.