NATURAL SPEECH SYNTHESIS BASED ON HYBRID APPROACH WITH CANDIDATE EXPANSION AND VERIFICATION

*Chung-Hsien Wu, Yi-Chin Huang, Shih-Lun Lin and Chia-Ping Chen**

Department of Computer Science and Information Engineering,
National Cheng Kung University, Taiwan, R.O.C.

*Department of Computer Science and Information Engineering,
National Sun Yat-Sen University, Taiwan, R.O.C.

ABSTRACT

A hybrid Mandarin speech synthesis system combining concatenation-based and model-based methodology is investigated in this research. To effectively exploit a small-size corpus, the candidate sets for unit selection are expanded via clusters based on articulatory features (AF), which are estimated as the outputs of an artificial neural network. This is followed by a filtering operation incorporating residual compensation, to remove unsuitable units. Given an input text, an optimal unit sequence is decided by the minimization of a total cost, which depends on the spectral features, contextual articulatory features, formants, and pitch values. Furthermore, prosodic word verification is integrated to check the smoothness of the output speech. The units failing to pass the prosodic word verification are replaced by model-based synthesized units for better speech quality. Objective and subjective evaluations have been conducted. Comparisons among the proposed method, the HMM-based method, and the conventional hybrid method clearly show that candidate set expansion based on articulatory features lead to more units suitable for selection, and the verification process is effective in improving the naturalness of the output speech.

Index Terms—Hybrid speech synthesis, Candidate expansion, Residual compensation

1. INTRODUCTION

Unit selection-based [1-3] and statistical methods, especially the Hidden Markov Model (HMM)-based method [4-6], have proven to achieve satisfactory intelligibility and flexibility in recent years. Voice conversion methods [7-8] to convert the speech signals uttered by a speaker to the other speaker with limited speech data have been one of alternatives recently. Among these approaches, the HMM-based method generally requires less effort on recording speech utterances compared to the unit selection-based method. On the other hand, even though unit-selection-based synthesis method requires a very large speech database to obtain various instances of phonetic or prosodic information, the resulting synthesized speech can have better voice quality than the HMM-based systems. Recently, numerous researchers have focused on developing a hybrid approach by combining both methods to obtain high-quality synthesized speech [9-12]. The main research interest in the hybrid method is to calculate the target cost based on the synthesized speech from the statistical synthesis method. Spectral parameters, *F*0 values, and duration information are generated from the HMM-based TTS system as the "targets" for unit selection. Generally, these studies used maximum-likelihood (ML)-estimated HMMs to predict the targets or calculate the costs between natural phone units and the corresponding synthesized phone units. In hybrid approaches, the requirement of large corpora remains a problem as they depend partially on unit selection. This problem can be solved in two aspects, collecting a large corpus or taking full advantage of a small corpus. In fact, collecting and labeling a large corpus is highly time-consuming and labor-intensive. Therefore many researchers made efforts on how to use a small corpus to achieve the optimal results in their research.

In the unit selection methods, linguistic information is generally employed to select the units with minimum target and concatenation costs. In a small corpus, the number of units with linguistic features matching the target unit and concatenation requirement is quite few. Suitable expansion of the potential units for selection can take full advantage of the small database to achieve a better synthesis performance. Based on a preliminary analysis on the corpus, some units with unmatched linguistic features are sometimes more suitable for concatenation to match the target units than the units with matched linguistic features. Articulatory features (AF), having been proven useful to detect the pronunciation variation of speech signals, are thus considered and used to expand the potential candidates for unit selection [13].

With the limited number of suitable units in the unit selection method, some outlying units in the optimally selected unit sequence are unmatched with the other units and result in spectral and prosodic discontinuities. The results imply that the performance of the unit selection method using a small corpus sometimes is inferior to that of the HMM-based synthesizer due to some outliers in the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

optimally selected units. To solve this problem, a prosodic word verification criterion is proposed to verify the smoothness of the synthesized speech. Prosodic word information from the training corpus is extracted and used to verify the existence of the outliers in the synthesized speech. If any part in the synthesized speech from the unit sequence with minimum cost is still not smooth enough, the synthesized units from the HMM-based synthesizer is used for replacement to obtain better speech quality and naturalness of the synthesized speech.

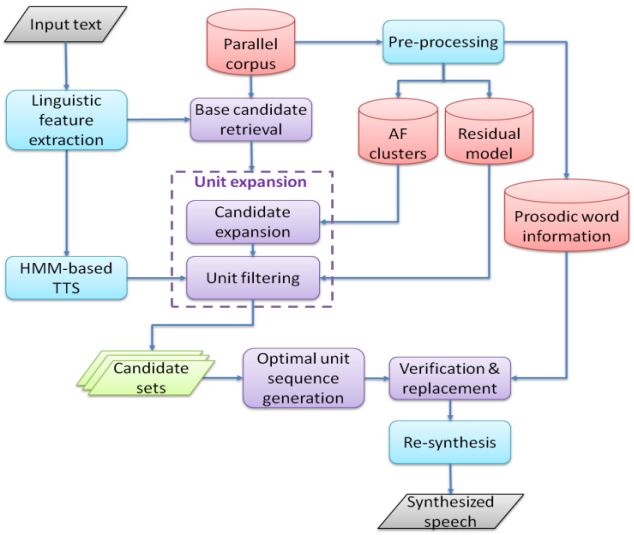

Fig. 1 Block diagram of the proposed system

2. PROPOSED METHOD

In this study, the block diagram of the proposed hybrid TTS system is shown in Fig. 1. First, linguistic features of the input text are extracted to select the base candidate units from the natural speech corpus and to generate the corresponding speech from an HMM-based TTS system. Here, the adopted units are Mandarin sub-syllables. We then expand the candidate units of the base candidate unit in the AF clusters, in which the expanded units have the same AF features as those of the base candidate units. Then, an optimal unit sequence of the input text can be obtained using the dynamic programming algorithm with the dedicated 2-level cost function. Finally, the prosodic word verification process is further adopted to verify the reliability of each concatenated unit. The unreliable units will be replaced by the HMM-based synthesized units to obtain more natural speech output.

2.1. Candidate Expansion Based on Phone Clustering

The goal of candidate expansion is to obtain the units unreachable using the linguistic information but potentially suitable for unit selection. Since the AFs of the same phonemes are intrinsically similar, they could be adopted to cluster the base units. In AF extraction, an artificial neural network (ANN) for different articulatory attributes is constructed [14]. The static and dynamic mel-general cepstral (MGC) coefficients are fed to the input layer. The two-level hidden layers, each consisting of 30 nodes, are used. The nodes at the output layer represent the posterior probabilities of 35 articulatory attributes, based on those used in [14], with additional Mandarin specific attributes. The output likelihoods of the articulatory attributes are defined as AFs.

For phone clustering, first, the phone identity and phone duration are considered. The K-means clustering method based on the Euclidean distance of articulatory features between two phones is adopted. In order to consider the temporal transition between consecutive frames, the AF vector contains the AF values, the delta and the means of AFs. The units in the same cluster are regarded as the expanded units.

2.2. Unit filtering Based on Residual Compensation

Practically, directly using the synthesized speech from the trained HMMs for filtering unsuitable units could be error-prone because of the residuals between natural units and the synthesized speech of the corresponding models. In order to compensate the difference between model-based synthesized units and natural units, the residual compensation concept proposed by Drugman *et al.* [12] is adopted. First, the trained HMMs are used to generate a pseudo-parallel corpus of the original natural corpus. Then, with forced alignment between synthesized speech and natural speech at the state level, the residuals can be calculated for each unit. In order to reduce the footprints of residuals, these residuals are clustered using the clustering method similar to that used for AF-based phone clustering described in Section 2.1. The resultant p-values of residual clusters are larger than 0.88, which means the residuals in the same cluster are consistent.

2.3. Phone Level Cost Estimation

At the first level for optimal unit sequence selection, the costs between target and natural units are estimated. Besides traditional concatenation cost estimation based on spectrum and pitch, the costs based on the following features are considered.

- **Formant:** the formant continuity between phoneme units in the same word is useful to check the smoothness of each word. In this study, the Euclidian distance for the first four formants is used.
- **Articulatory features** – 35 AFs are used to estimate the Euclidian distances.

The concatenation cost is estimated as follows.

$$\boldsymbol{Con}(\boldsymbol{x_n}, \boldsymbol{x_{n+1}}) = \boldsymbol{\alpha} \cdot \boldsymbol{D}(\boldsymbol{x_n}, \boldsymbol{x_{n+1}}) + \boldsymbol{\beta} \cdot \boldsymbol{A}(\boldsymbol{x_n}, \boldsymbol{x_{n+1}}) + \boldsymbol{\gamma} \cdot \boldsymbol{F}^*(\boldsymbol{x_n}, \boldsymbol{x_{n+1}}), \quad (1)$$

$$\boldsymbol{F}^*(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) = \begin{cases} \boldsymbol{0}, & \boldsymbol{p_{t+1}^{phone} = 1} \\ \boldsymbol{F}(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}), & \boldsymbol{otherwise} \end{cases}, \quad (2)$$

Where x_n represents the n-th candidate unit, $D(\cdot)$ is the distance for spectrum and pitch, $A(\cdot)$ is the distance for AF,

$F(\cdot)$ is the distance for formant, and p_t^{phone} is the phone position of the *t*-th word. The resultant phone candidates are then used for prosodic word level selection for higher level naturalness.

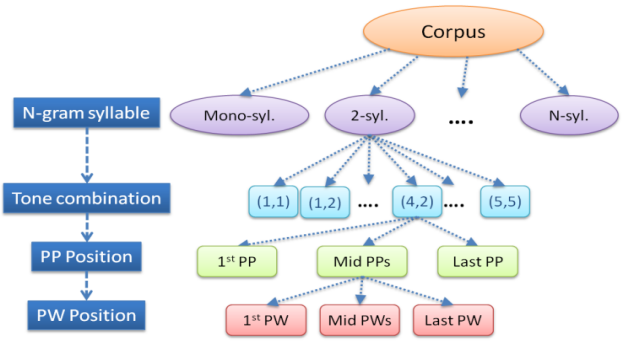

Fig. 2 Illustration of the constructed pitch decision tree

2.4. Prosodic Word Level Cost Estimation

The prosodic word level cost function, used after phone level cost function, includes concatenation (*Con*) cost between words and substitution (*Sub*) cost of the word. The optimal word sequence $\hat{w}^N$ can be obtained as

$$\hat{\boldsymbol{w}}^{\boldsymbol{N}} = \boldsymbol{argmin}_{w^N} \sum_{n=2}^{N} \boldsymbol{Cost}(\boldsymbol{w_n}, \boldsymbol{w_{n-1}}), \quad (3)$$

$$\boldsymbol{Cost}(\boldsymbol{w_n}, \boldsymbol{w_{n-1}}) = \boldsymbol{\alpha} \cdot \boldsymbol{Con}(\boldsymbol{w_n}, \boldsymbol{w_{n-1}}) + \boldsymbol{\beta} \cdot \boldsymbol{Sub}(\boldsymbol{w_n}), \quad (4)$$

$$\boldsymbol{Con}(\boldsymbol{w_n}, \boldsymbol{w_{n-1}}) = \boldsymbol{D^C}\left(\boldsymbol{w_n^{MGC}}, \boldsymbol{w_{n-1}^{MGC}}\right) + \boldsymbol{D^C}\left(\boldsymbol{w_n^{Pitch}}, \boldsymbol{w_{n-1}^{Pitch}}\right) \quad (5)$$

$$\boldsymbol{Sub}(\boldsymbol{w_n}) = \boldsymbol{D^S}(\boldsymbol{w_n}, \boldsymbol{w_n^{Mean}}) \quad (6)$$

$$\boldsymbol{D^C}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=-2}^{1} \boldsymbol{Dist}(\boldsymbol{a_i}, \boldsymbol{b_{N+1+i}}) \quad (7)$$

$$\boldsymbol{D^S}(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{N} \boldsymbol{Dist}(\boldsymbol{a_i}, \boldsymbol{b_i}) \quad (8)$$

where w_n is the *n*-th candidate word.

2.5. Prosodic Word Verification

The verification operation will verify the prosody smoothness and energy consistency in the synthesized speech. The prosodic words are selected as the verification units and can be identified using the approach proposed in [13]. The mean values of the pitch and energy features of the prosodic words in the natural corpus are extracted and used to construct the corresponding decision tree. The decision trees are constructed by the tone combination of the prosodic words. As the research proposed by Hoole *et al*. [15], pitch contours of tone combination are correlated with the manner and place of phone articulation. The decision tree for pitch is constructed based on the articulatory features of the phones. The pitch decision tree is shown in Fig. 2. The number in this figure represents the tone number, e.g., "1" is tone 1 and "(1,2)" represents the tone combination of tone 1 and tone 2 for a two-syllable word.

Energy of a word/phrase is generally related to the position of the word/phrase in an utterance. The position of a prosodic phrase and the position of a prosodic word are adopted to construct the energy decision tree. Because the energy of the first word in a phrase is usually larger and the last word is usually smaller compared to other words, three types of positions for prosodic phrases and prosodic words are defined, including First, Middle and Last, to construct the energy decision tree. The structure of the energy decision tree is constructed similarly to the pitch decision tree.

For verification, the prosodic word *S* is assumed to have p syllables $\{s_1, \ldots, s_p\}$, and each syllable s_i has the corresponding pitch and energy, p_i and e_i. If there are more than 50% of the frames with their pitch values exceeding one standard deviation from the mean pitch value or at least one frame with the pitch value exceeding two standard deviations from the mean pitch value in a syllable of the corresponding prosodic word with the same tone combination, this syllable is regarded to fail the verification. If the mean energy of a syllable in the synthesized speech exceeds the mean energy of the standard prosodic word with one standard deviation, then this syllable cannot pass the verification. Only the syllable passing the pitch and energy verification simultaneously can be kept in the output speech.

$$\boldsymbol{v^p}(\boldsymbol{p_i}) = \begin{cases} \boldsymbol{reject},\ \boldsymbol{if}\ \sum_{i=1}^{n} \boldsymbol{v_1^p}(\boldsymbol{p_i}) > \boldsymbol{n} \times \boldsymbol{50\%} \\ \boldsymbol{reject},\ \boldsymbol{if}\ \sum_{i=1}^{n} \boldsymbol{v_2^p}(\boldsymbol{p_i}) > 0 \\ \boldsymbol{accept},\ \boldsymbol{otherwise} \end{cases} \quad (9)$$

$$\boldsymbol{v_1^p}(\boldsymbol{p_i}) = \boldsymbol{Dist}(\boldsymbol{p_i}, \boldsymbol{p_i^{mean}}) > \boldsymbol{p_i^{std}} \quad (10)$$

$$\boldsymbol{v_2^p}(\boldsymbol{p_i}) = \boldsymbol{Dist}(\boldsymbol{p_i}, \boldsymbol{p_i^{mean}}) > 2 \times \boldsymbol{p_i^{std}} \quad (11)$$

$$\boldsymbol{v^e}(\boldsymbol{e_i}) = \begin{cases} \boldsymbol{reject},\ \boldsymbol{if}\ \boldsymbol{Dist}(\boldsymbol{e_i}, \boldsymbol{e_i^{mean}}) > \boldsymbol{e^{std}} \\ \boldsymbol{accept},\ \boldsymbol{otherwise} \end{cases} \quad (12)$$

$$\boldsymbol{v}(\boldsymbol{s_i}) = \begin{cases} \boldsymbol{accept},\ \boldsymbol{if}\ \boldsymbol{v^p}(\boldsymbol{p_i}) = \boldsymbol{accept}\ \boldsymbol{and}\ \boldsymbol{v^e}(\boldsymbol{e_i}) = \boldsymbol{accept} \\ \boldsymbol{reject},\ \boldsymbol{otherwise} \end{cases} \quad (13)$$

The units which fail the verification will be replaced by the HMM-based synthesized unit outputs.

3. EXPERIMENTAL RESULTS

3.1 Speech Data Collection and Experimental Setup

The phonetically balanced sentences from the TsingHua-Corpus of Speech Synthesis (TH-CoSS) [16] produced by a female speaker, including 5,406 utterances with 98,749 syllables, were used. The context-dependent phone labels were constructed automatically based on the linguistic features extracted from the database. For Mandarin speech synthesis, 107 phone units (including one pause unit) were selected to construct a Mandarin HMM-based TTS system [17-18]. Each Mandarin syllable consists of two vowel phone units and one optional consonant phone unit. The sampling rate of the speech signals was 16kHz and the smoothed spectral coefficients were extracted using the STRAIGHT algorithm. A 5-state left-to-right HMM was adopted to model the acoustic features. The HMM-based Mandarin TTS system was trained using the speech data with numbers 1 to 3,000 in TH-CoSS corpus. The HMMs

were used to generate the same 3,000 sentences for calculating the residuals. The remaining utterances with numbers 3,001 to 5,406 were used as the test corpus.

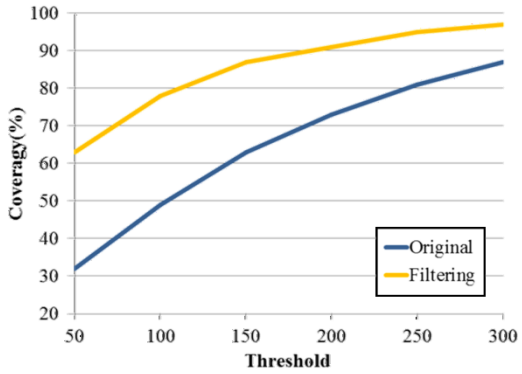

Fig. 3 Relationship between the filtering thresholds and the percentages of correctly extracted units

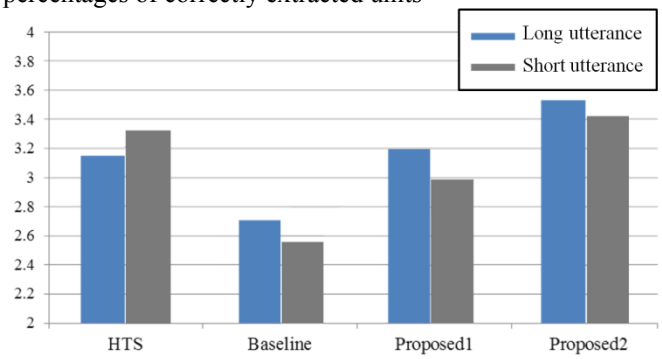

Fig. 4 MOS results for the comparison of different approaches using long/short utterances

3.2. Objective Test

To evaluate the performance of unit filtering and set suitable filtering threshold, we evaluated the relation between the number of the remaining candidates units after filtering and the threshold values. The recall rate of the correct units was adopted to decide the threshold. In Fig. 3, different thresholds for unit filtering were evaluated. The value of the threshold was set to 250 which retrieved 95% correct units in the training corpus.

3.3. Subjective Test

In subjective test, long utterances (≥18 syllables) and short utterances (<18 syllables) were used to evaluate the quality of the proposed approaches for comparison. Long utterances were adopted to evaluate the smoothness of the synthesized speech, while short utterances, which consist of some specific unseen words such as proper name, were adopted to evaluate the verification performance. 12 native subjects were invited to evaluate the Mean Opinion Score (MOS) and preference test. Comparisons among the following approaches were conducted to evaluate the performance of each factor.

- **HTS**: HMM-based synthesizer
- **Baseline**: Hybrid baseline using only linguistic features
- **Proposed1**: Hybrid approach with unit expansion
- **Proposed2**: Hybrid approach with unit expansion and prosodic word verification

The MOS of naturalness for these approaches evaluated by the listeners are shown in Fig. 4. The results show that listeners perceived stable synthesized speech generated by the **HTS** system, while some inevitable noises were generated by the **Baseline** system, which retrieved the units using linguistic features only. Thus the MOS value is higher for the **HTS**. However, through candidate expansion, the synthesized speech generated from **Proposed1** is more acceptable than the **Baseline**. The naturalness of the synthesized speech generated by **Proposed2** was higher than that by the **HTS**.

Preference test compared the following combinations of the approaches in regard to naturalness of the synthesized speech: 1) **HTS vs. Baseline**; 2) **HTS vs. Pro1**; 3) **HTS vs. Pro2**; 4) **Pro1 vs. Pro2**.

Pro1 and **Pro2** stand for systems **Proposed1** and **Proposed2**, respectively. The results for preference test are shown in Fig. 5. The **HTS** system outperformed the **Baseline** system in both long utterance and short utterance cases. With prosodic word verification, **Proposed2** outperformed the **HTS**. Comparing the approaches with and without verification, one can see that verification can enhance the naturalness of synthesized speech significantly.

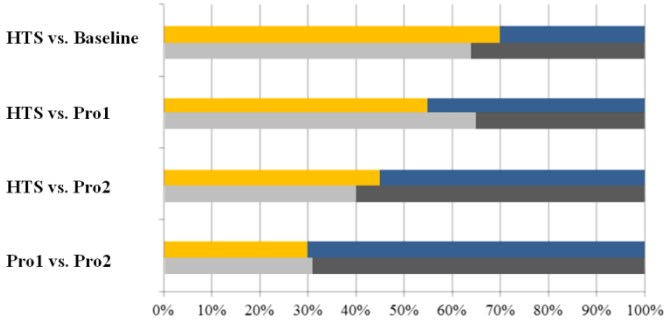

Fig. 5 Comparisons of the preference test on the approaches for long utterances (yellow/blue) and short utterances (grey/black).

4. CONCLUSION

In this paper, an AF-based unit expansion method to enhance the performance of unit selection and achieve the goal to fully use a small corpus is proposed. Residual-compensated synthesized target unit were employed for effective unit expansion. Prosdic word verification is further employed to verify each unit in the concatenated speech to obtain the speech outputs with better naturalness.

Objective and subjective experimental results show that candidate expansion can retrieve more suitable units from the corpus unreachable by the linguistic features. The verification process is also useful to improve the naturalness and quality of the synthesized speech.

5. REFERENCES

[1] S. Sakai and H. Shu, "A Probabilistic Approach to Unit Selection for Corpus-based Speech Synthesis," *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, pp. 81-84, 2005.

[2] F. Campillo D. and Eduardo R. Banga, "A method for combining intonation modeling and speech unit selection in corpus-based speech synthesis systems," *Speech Communication*, vol. 48, pp. 941-956, 2006.

[3] C.-H. Wu, C.-C. Hsia, J.-F. Chen and J.-F. Wang, "Variable-length unit selection in TTS using structural syntactic cost," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1227–1235, May 2007.

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

[5] 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.

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

[7] 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 Trans. Audio, Speech, and Language Processing*, Vol. 14, No. 4, July, 2006, pp.1109~1116.

[8] 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 Trans. Audio, Speech, and Language Processing*, Vol. 18, No. 6, August 2010, pp. 1394~1405.

[9] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A Hybrid Text-to-Speech System That Combines Concatenative and Statistical Synthesis Units," *IEEE Transactions on Audio, Speech, And Language Processing*, vol. 19, no. 5, pp. 1278-1288, 2011.

[10] E. Guner, A. Mohammadi, and C. Demiroglu, "Analysis of Speaker Similarity in the Statistical Speech Synthesis Systems Using A Hybrid Approach," *20th European Signal Processing Conference (EUSIPCO 2012)*, pp. 2055-2059, 2012.

[11] I. Sainz, D. Erro, E. Navas, I. Hernáez, "A Hybrid TTS Approach for Prosody and Acoustic Modules," *12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pp. 333-336, 2011.

[12] T. Drugman, A. Moinet, T. Dutoit, and G. Wilfart, "Using A Pitch-Synchronous Residual Codebook For Hybrid HMM/Frame Selection Speech Synthesis," *2009 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pp. 3793-3796, 2009.

[13] Y.-C. Huang, C.-H. Wu and Y.-T. Chao, "Personalized Spectral and Prosody Conversion using Frame-Based Codeword Distribution and Adaptive CRF," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 21, No. 1, January 2013, pp. 51~62.

[14] Y.-C. Huang, C.-H. Wu, S.-L. Lin, "Residual Compensation based on Articulatory Feature-based Phone Clustering for Hybrid Mandarin Speech Synthesis," in *Proceeding of the 8th ISCA Speech Synthesis Workshop*, August 31st - September 2nd, 2013, Barcelona, Spain.

[15] P. Hoole and F. Hu, "Tone-Vowel Interaction in Standard Chinese," *International Symposium on Tonal Aspects of Languages with Emphasis on Tone Languages (TAL 2004)*, pp. 89-92, 2004.

[16] 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.

[17] 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.

[18] 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, Nov. 2010.