USTC System for Blizzard Challenge 2006
an Improved HMM-based Speech Synthesis Method

Zhen-Hua Ling, Yi-Jian Wu, Yu-Ping Wang, Long Qin, Ren-Hua Wang

University of Science and Technology of China
zhling@ustc.edu, jasonwu@mail.ustc.edu.cn, ypwang2@ustc.edu,
qinlong@mail.ustc.edu.cn, rhw@ustc.edu.cn

ABSTRACT

This paper introduces the USTC speech synthesis system for Blizzard Challenge 2006. The HMM-based parametric synthesis approach was adopted for its convenience and effectiveness in building a new voice, especially for the non-native developers. Some useful techniques were also integrated into our system, such as minimum generation error (MGE) training, phone duration modeling and linear spectral pair (LSP) based formant enhancement. The evaluation results show that the proposed system is able to synthesize speech with high naturalness and intelligibility by using either full database or only ARCTIC subset.

1. INTRODUCTION

In recent years, HMM-based speech synthesis method has been proposed and applied successfully in the synthesis system of different languages [1-3]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [1] and the parameters are generated from HMMs under maximum likelihood criterion by using dynamic features [4]. Then parametric synthesizer is used to reproduce speech signals. This method is able to synthesize highly intelligible and smooth speech. Besides, the voice character of synthetic speech can be controlled flexibly by employing some model adaptation methods [5].

We adopted this approach to construct the system for Blizzard Challenge 2006 also for the following reasons:

1) This method is able to produce speech with high smoothness and naturalness and the system can be constructed in a short space of time.
2) The system building process is almost fully-automatic. This advantage is especially important and convenient for us non-native developers compared with conventional cost-based unit selection method, where a lot of cost tables or weights need to be tuned manually.

In order to improve the performance of baseline HMM based synthesis system, especially the voice quality of synthesized speech, some new techniques, such as MGE training [6], phone duration modeling and LSP-based formant enhancement [3], were integrated into our system. The flowchart of proposed system is shown in Fig. 1.

This paper is organized as follows. Section 2 describes the new methods that have been applied in our system. Section 3 gives some detailed introductions of system building. The discussions and conclusions are in section 4 and 5.

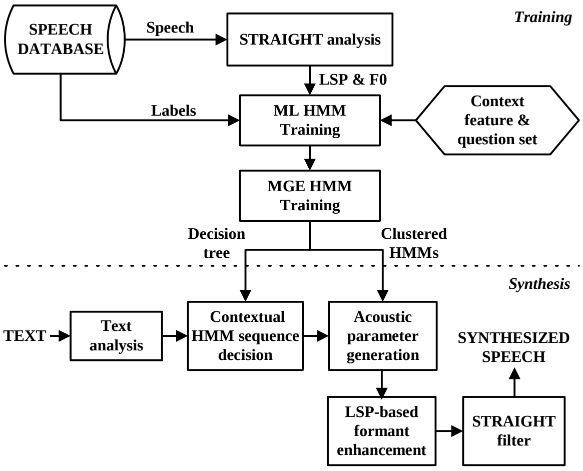

Figure 1: Flowchart of proposed method.

2. METHOD

2.1 Baseline system

The whole system can be divided into training stage and synthesis stage. In training stage, a set of contextual dependent HMMs are estimated according to the acoustic features and label information of the training database under maximum likelihood criterion [1]. The feature vector consists of static and dynamic components of spectrum and F0. The spectrum part is modeled by a continuous probability distribution and the F0 part is modeled by a multi-space probability distribution (MSD) [7]. A decision tree based model clustering method is applied after contextual dependent HMM training to improve the robustness of estimated models. During synthesis, the prosodic and spectral parameters are predicted by ML-based parameter generation method [4] and sent to parametric synthesizer to reproduce waveform.

In order to improve the performance of the baseline system, some new techniques are introduced, which will be discussed in the following paragraphs.

2.2 Minimum generation error training

Under MGE criterion, the model parameters are estimated to minimize the difference between generated parameters and natural ones for the sentences in training set. This method has been proved to improve quality of synthesized speech effectively in our previous work [6]. This improvement can be explained from two aspects. First, this criterion gives better consistency between model training and the purpose of speech synthesis, which is to produce speech signal or parameter sequences as closely as the natural ones. Second, by incorporating parameter generation into the training procedure, the constraints between static and dynamic features are considered in HMM training. Here, we use the ML trained clustered contextual dependent model as initial model as shown in Fig.1 and then apply MGE training, in which Generalized Probabilistic Descent (GPD) algorithms [8] is used to update the model parameters of spectral part.

2.3 Phone duration modeling

In the baseline system, a state duration model is trained to predict the duration of every state in the utterance for synthesis. Considering state is not an explicit and stable phonetic unit, a phone duration model is also constructed in our system and is combined with the state duration model to predict the duration of each state [3].

Assuming *N* is the number of phones in the sentence, *S* is the number of states defined in a phone, $d_{n,s}$ is the duration of state *s* in phone *n* and can be predicted as

$$d^*_{n,s} = \max_{d_{n,s}}[\log p_{n,s}(d_{n,s}) + w\bullet\log p_n(\sum_{i=1}^{S} d_{n,i})] \qquad (1)$$

where $p_{n,s} = N(d_{n,s} \mid m_{n,s}, \sigma^2_{n,s})$ presents the state duration model for state *s* in phone *n* and $p_n = N(d_n \mid m_n, \sigma^2_n)$ presents the phone duration model for phone *n*. *w* is set as the weight between these two models. So the final state duration can be derived from Eq.1 as

$$d^*_{n,s} = m_{n,s} + \rho_n\bullet\sigma^2_{n,s}$$

$$\rho_n = \frac{w\bullet(m_n - \sum_{i=1}^{S} m_{n,i})}{\sigma^2_n + w\bullet\sum_{i=1}^{S}\sigma^2_{n,i}} \qquad (2)$$

2.4 LSP based formant enhancement

STRAIGHT [9] as a high quality speech vocoder is adopted here to extract acoustic features from speech waveforms and to synthesize speech using generated parameters. We select linear spectral pair (LSP) to present each frame's spectral envelop estimated by STRAIGHT because LSPs relate more closely to formant positions and have better smoothness among adjacent frames. At first the linear prediction coefficients are estimated from spectral envelop by all-pole modeling [10] and LSPs are then derived. In order to improving the accuracy of spectral fitting within low frequency band, spectral warping processing is carried out during all-pole modeling.

Because of the averaging effect of statistic modeling, the spectrums reconstructed from ML-based parameter generation are always over-smoothed and the formants are broaden, which make the synthetic speech sounds muffled. Here, the relationship between spectral peaks and LSP, especially the difference between its adjacent orders, is used to enhance the formants of synthesized speech [3]. Define the ML generated LSPs of one frame are $l_i, i = 1,....,D$, D is the order of all-pole modeling. Then the new LSPs can be calculated from order 2 to order $D-1$ recursively as

$$l_i' = l_{i-1} + d_{i-1} + \frac{d^2_{i-1}}{d^2_{i-1} + d^2_i}[(l_{i+1} - l_{i-1}) - (d_i + d_{i-1})]$$

$$d_i = \alpha\bullet(l_{i+1} - l_i), \quad \alpha < 1, \quad i = 2,...D-1 \qquad (3)$$

where α controls the degree of the enhancement. The less α is, the more obvious the enhancement will be. Listening test proves that this formant enhancement method is able to improve the quality and articulation of generated speech effectively.

3. SYSTEM BUILDING

3.1 Speech database

The database for Blizzard Challenge 2006 contains 4273 sentences from 3 domains – novel, newspaper and conversation. Two systems are required, one is built with the full database and one is with the ARCTIC subset which contains novel style sentences. We select only 70% of the sentences from each subset based on the coverage of triphones to train the full set system because our PC for training fails to afford the memory requirement to train with all the 4273 sentences, which contain more than 200,000 phone units.

3.2 Contextual factors and question set designing

In order to get a reliable context dependent HMM, the question set designing is very important and only this part in our system is language dependent. The whole question set can be roughly divided into the following layers according to the available contextual information:

1) Phone layer: the name and type of current and surrounding phonemes; the number and position of phones in syllable.

2) Syllable layer: the stress and accent type of current and surrounding syllables; the number and position of syllables in word.

3) Word layer: the POS of current and surrounding words; the number and position of words in phrase.

4) Phrase layer: the number and position of phrases in utterance; the boundary type of current phrase.

5) Utterance layer: the number of syllables, words and phrases in utterance.

6) Subset layer (*SS_NOV*, *SS_NEWS*, *SS_CONV*): only for the system with full database, to determine the category of subset where the sentence belongs to.

3.3 Parameter Extraction and Generation

In our system, acoustic features are extracted at 5 ms frame shift. The order of LSPs derived from STRAIGHT spectrum is set to 40 and there is another dimension for gain coefficient. So the total size of each feature vector is 126 considering the static, delta and acceleration components of both spectrum and logarithmized F0. In synthesis stage, the w in Eq. 2 is set to 10 for state sequence generation and α in Eq. 3 is set to 0.7 for spectral enhancement. An example of the effect of proposed formant enhancement method for the system with full database is shown in Fig. 2.

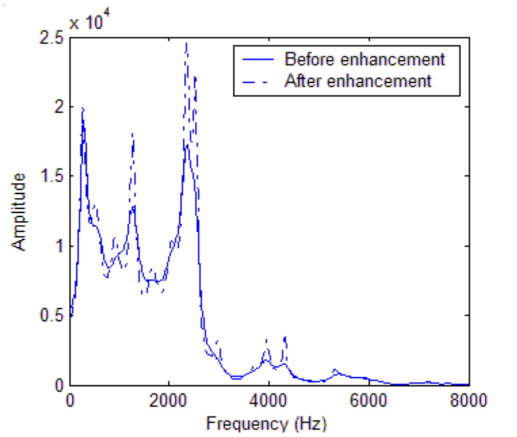

Figure 2: An example of LSP based formant enhancement

3.4 Model training

5-state left-to-right without skip HMM structure is used in our system. Because the scripts and pronunciation of each sub-corpus has its own characters, we add three questions in our question set for full database system to determine whether the sentence is from a novel, a newspaper or a conversation sub-corpus as mentioned in section 3.2. The frequency of these three questions asked in our trained decision trees is summarized in Table 1. Because the sentence length and reading style of sentences from newspaper subset are quite different from other two subsets, the question *SS_NEWS* has the highest frequency of being asked.

Table 1: The frequency of the three subset layer question being asked

Question	Frequency
SS_NOV	118
SS_NEWS	611
SS_CONV	93

After ML training for clustered contextual dependent HMMs, MGE training is carried out where the number of iteration for GPD based parameter updating is set to 20. The convergence property of the MGE training is shown in Fig. 3, where only the results of a few dimensions on close test of full database training are presented.

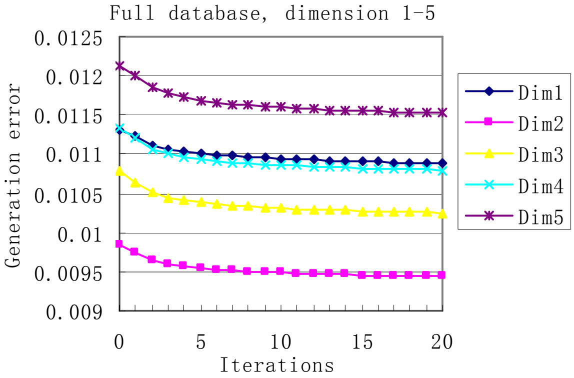

Figure 3: Convergence of MGE training

Because of the introducing of MGE training, the time consuming of whole training procedure increase. The time consumed by baseline training and MGE updating are summarized in Table 2, where a 3.0G CPU PC is used for model training and the time for acoustic feature extraction is not included. From this table, we can see that MGE training will not increase the computation cost of model training significantly, especially when the training set is large, where most of the running time is consumed on decision tree based clustering.

Table 2: Time consuming for model training (hour:minute:second)

System	Size (MB)	Baseline training	MGE training
Full set (70%)	521	77:05:47	16:34:11
ARCTIC subset	141	9:37:27	4:47:18

3.5 System performance

The evaluation results for average MOSs and WERs of our system are shown in Fig.4 and Fig.5. From these figures, we can see that our system achieves the best performance for ARCTIC database in both MOS and WER. For the full database, the difference between our system and the best one of others is not significant.

4. DISCUSSIONS

The evaluation results show the superiority of HMM based synthesis method over conventional unit selection method when a small training set is used. With the size of speech database increasing, the unit selection and concatenative method is able to produce natural speech with high quality, although the stability and robustness are still the problem for these systems. At this time, the superiority of HMM based parametric synthesis method decreases, especially

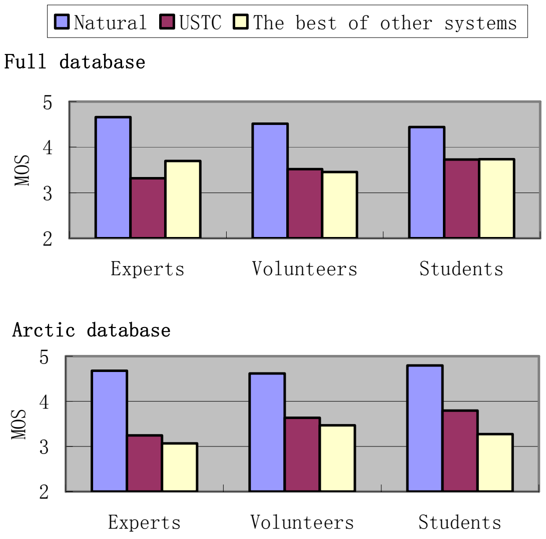

Figure 4: Average MOSs of natural speech, USTC system and the best of other systems

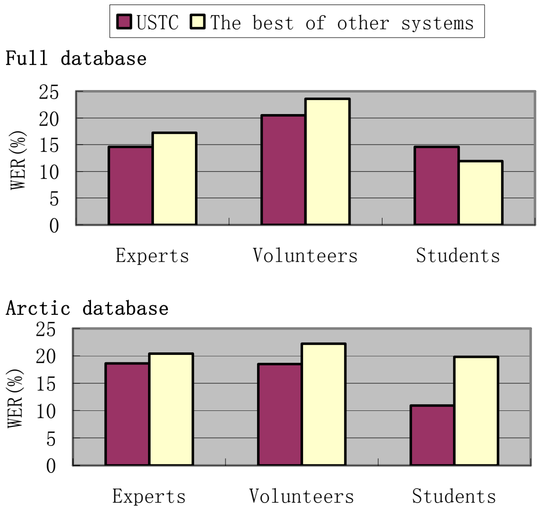

Figure 5: Average WERs of USTC system and the best of other systems

for the expert listeners who are more sensitive to the difference between natural sounds and speech generated by vocoders than common listeners as shown in Fig. 4 and 5. Besides, we met some problems when we were training the system by full database. On one hand, the time and memory consuming increase greatly, which decrease the flexibility of system building. On the other hand, from Fig. 4 and 5, we can see that our system built with full database does not show enough advantage over the system built with only ARCTIC database, especially for non-experts listeners. How to make full use of a large speech database for HMM based synthesis method still needs further work.

5. CONCLUSIONS

This paper introduces the system developed by USTC for Blizzard Challenge 2006. Several new techniques are employed to improve the performance of baseline HMM based synthesis system. The evaluation result confirms the effectiveness of proposed methods. However, there still exist some problems for such HMM based parametric synthesis method, especially when the size of training database is large. Besides, the generated prosodic features still sounds too flat and lack of expressive variation and the performance gap between our system and natural speech is still large. In order to improve the performance of current system further, designing higher performance speech vocoder, improving modeling and parameter generation strategy, combining statistic modeling with some unit selection methods [11] may be the approaches we can attempt in our future work.

6. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in Proc. of Eurospeech, pp. 2347-2350, 1999.

[2] H. Zen, and T. Toda, “An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005”, in Proc. of Eurospeech, pp. 93-96, 2005.

[3] Yi-Jian Wu, “Research on HMM-based Speech Synthesis”, Ph.D Thesis, University of Science and Technology of China, 2006. [in Chinese]

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis,” in Proc. of ICASSP, pp. 1315-1318, 2000.

[5] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis Using MLLR”, in Proc. of ICASSP, pp. 805-808, 2001.

[6] Yi-Jian Wu, and Ren-Hua Wang, “Minimum generation error training for HMM-based speech synthesis”, in Proc. of ICASSP, pp. 89-92, 2006

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, in Proc. of ICASSP, pp. 229-232, 1999.

[8] J. R. Blum, “Multidimensional stochastic approximation methods,” Ann. Math. Stat, vol. 25, pp.737-744, 1954.

[9] H.Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Restructuring speech representations using pitch-adaptive timefrequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187-207, 1999.

[10] R. J. MacAulay, and T. F. Quatieri, Sinusoidal coding, in Speech Coding and Synthesis, Elsevier, Amsterdam, 1995.

[11] Zhen-Hua Ling, and Ren-Hua Wang, “HMM-based unit selection using frame sized speech segments”, to be appeared in ICSLP, 2006.