Comparison of Syllable/Phone HMM based Mandarin TTS

Quansheng Duan, Shiyin Kang, Zhiyong Wu, Lianhong Cai
Department of Computer Science and Technology
Tsinghua University
Beijing, China
{ddandre32, kangshiyin, john.zy.wu}@gmail.com
clh-dcs@mail.tsinghua.edu.cn

Zhiwei Shuang, Yong Qin
IBM China Research Lab
Beijing, China
{shuangzw, qinyong}@cn.ibm.com

***Abstract*—The performance of HMM-based text to speech (TTS) system is affected by the basic modeling units and the size of training data. This paper compares two HMM based Mandarin TTS systems using syllable and phone as basic units respectively with 1000, 3000 and 5000 sentences' training data. Two female speakers' corpora are used as training data for evaluation. For both corpora, the system using syllable as basic unit outperforms the system using phone as basic unit with 3000 and 5000 sentences' training data.**

Keywords-Speech Synthesis; HMM; syllable; Mandarin

I. INTRODUCTION

Mandarin is a syllable based tonal language. Each character is pronounced as a syllable. Syllables consist maximally of an initial consonant, a glide, a vowel, a final and tone. However, as there are rules prohibiting certain phoneme appearing with others, there are only about 1300 tonal syllables. These tonal syllables are the legal combinations of about 410 base syllables and 5 tones. Due to the limited number of syllables and the strong intra-syllable co-intelligence, most Mandarin concatenative TTS systems use syllable as basic units for concatenation.

HMM-based TTS technology has developed rapidly in recent years [1]. Compared to the traditional large-corpus-based concatenative TTS system HMM based TTS needs a lower resource cost, and has less relevance to speakers and languages [2]. The performance of HMM-based TTS system is affected by the basic modeling units and the size of training data. However, in most reported Mandarin HMM based TTS systems, phone or demi-syllable are used as basic units for training and synthesis [3, 4].

To maximally leverage the previous research and development work on syllable based concatenative TTS system, we build a HMM based TTS system using syllable as basic units, which we call Syllable HMM based TTS system [5]. We utilize the unique consonant/vowel structure of Mandarin syllable to improve the voiced/unvoiced decision of HMM states. Meanwhile, we also build a HMM based TTS system using phone as basic unit for comparison, which we call Phone HMM based TTS. To understand the influence of modeling units and size of training data, we make a preference evaluation and MOS evaluation between Syllable HMM based TTS and Phone HMM based TTS with different amount of training data. Two female speakers' corpora are used in evaluation.

The rest of this paper is organized as follows. The Syllable/Phone HMM based TTS systems are described in Section 2. Comparison evaluation of two systems and discussion are given in Section 3. Concluding remarks are presented in Section 4.

II. SYLLABLE/PHONE HMM BASED TTS

TTS system using traditional HMM-based speech synthesis method can be divided into training stage and synthesis stage.

In training stage, parameters are extracted from corpus and used as the training data of HMM models. Acoustic parameters, such as F0 and spectrum [6], as well as their dynamic characteristics, are concerned in HMM model-training [7]. These models are then clustered by decision tree following MDL criterion [8].

In synthesis stage, context information is generated by the text analysis procedure. Using context information, the system predicts HMM sequence by the decision tree [9]. Finally, the speech parameter sequence, which is generated based on the predicted models, is used to synthesis the speech waveform by a vocoder.

Syllable HMM based TTS system and Phone HMM based TTS system are built by referring to the "HMM-based speech synthesis system", except that several modifications are made to improve the performance. In the following introduction, we will focus on special part in our implementation for both systems instead of general system process.

Manually checked prosodic structure information and syllable pronunciation are used for decision-tree based context clustering of HMMs.

A. Phone HMM based TTS system

As shown in Fig. 1, Phone HMM based TTS uses a 5-state left-to-right HMM topology to model each phone.

Prosody structure of Phone HMM based TTS system includes 6 layers: phone (PHO), syllable (SYL), prosodic word (PW), prosodic phrase (PP), intonation phrase (IP) and sentence (SEN). Totally, 65 prosodic structure features and 24 pronunciation features are considered in Phone HMM based TTS system, as shown in Table 1.

TABLE I. CONTEXT FOR PHONE HMM BASED TTS

	Contexts
PSF[a]	{position, reverse position} of PHO in SYL/PW/PP/IP/SEN
	{position, reverse position} of SYL in PW/PP/IP/SEN
	{position, reverse position} of PW in PP/IP/SEN
	{position, reverse position} of PP in IP/SEN
	{position, reverse position} of IP in SEN
	Number of PHO in {previous, current, next} SYL/PW/PP/IP
	Number of SYL in {previous, current, next} PW/PP/IP
	Number of PW in {previous, current, next} PP/IP
	Number of PP in {previous, current, next} IP
PF[b]	{previous, current, next}Pinyin/Tone/Initial/Final
	{previous, current, next}Method of articulatory type of initial
	{previous, current, next}Place of articulatory of initial
	{previous, current, next}Articulatory method of first/last vowel

a. PSF: prosodic structure feature. b. PF: pronunciation features

The method of articulatory type of initial, the place of articulatory of initial, the articulatory method of vowel are based on SAMPA-C Schema, except one silence category is added.

B. Syllable HMM based TTS system

As shown in Fig. 2, Syllable HMM based TTS uses a 10-state left-to-right HMM topology to model each syllable.

Prosody structure of Syllable HMM based TTS system includes 5 layers: syllable (SYL), prosodic word (PW), prosodic phrase (PP), intonation phrase (IP) and sentence(SE N). There are 42 prosodic structure features (PSF) and 24 pronunciation features (PF) in the decision-tree based clustering process, as shown in Table 2.

TABLE II. CONTEXT FOR SYLLABLE HMM BASED TTS

	Contexts
PSF[a]	{position, reverse position, number} of SYL in PW/PP/IP/SEN
	{position, reverse position, number} of PW in PP/IP/SEN
	{position, reverse position, number} of PP in IP/SEN.
	{position, reverse position, number} of IP in SEN.
	Number of SYL in {previous, next} PW/PP/IP
	Number of PW in {previous, next} PP/IP
	Number of PP in {previous, next} IP
PF[b]	{previous, current, next}Pinyin/Tone/Initial/Final
	{previous, current, next}Method of articulatory type of initial
	{previous, current, next}Place of articulatory of initial
	{previous, current, next}Articulatory method of first/last vowel

a. PSF: prosodic structure feature. b. PF: pronunciation features

A voiced/unvoiced decision algorithm is employed in Syllable HMM based TTS [5]. As we know, there is no more than one V/U changing points in the whole syllable. This algorithm helps avoiding inappropriate unvoiced decision inside syllables. The algorithm calculates the unvoiced percentage in the whole syllable and then models this percentage for each syllable by GMM models. Therefore, in synthesis stage, we can check all the voiced/unvoiced changing points, and select the changing point with the highest probability according to the GMM models [10].

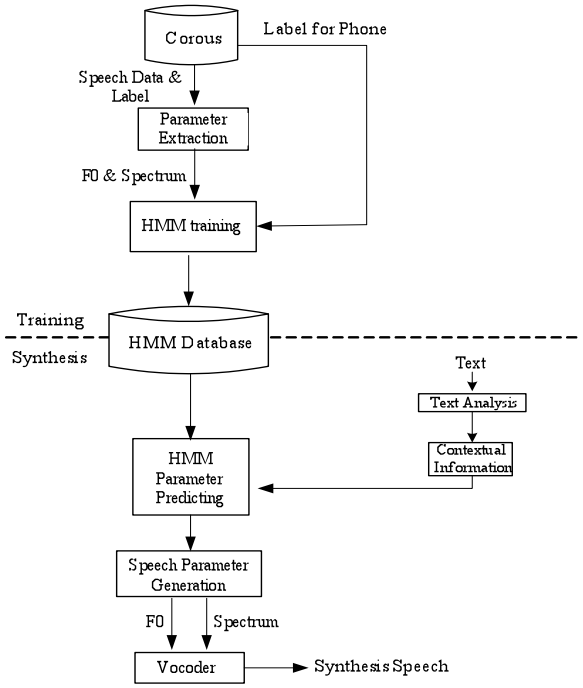

Figure 1. Phone HMM based TTS on Mandarin.

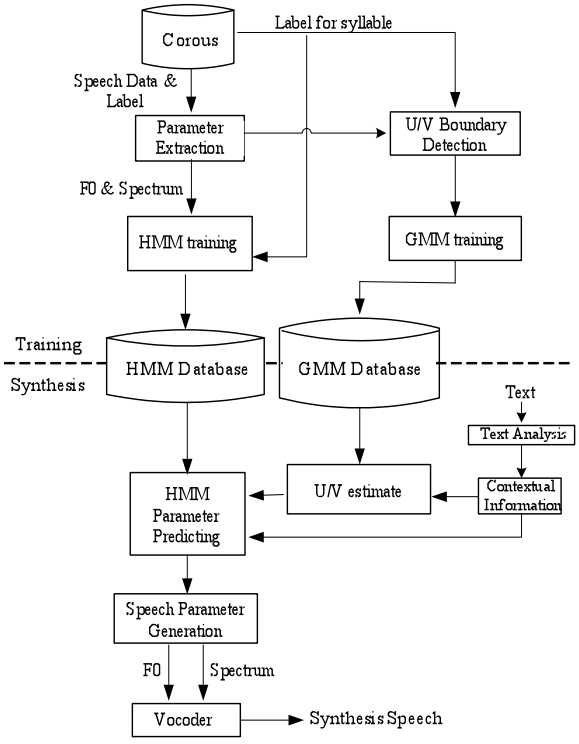

Figure 2. Syllable HMM based TTS on Mandarin

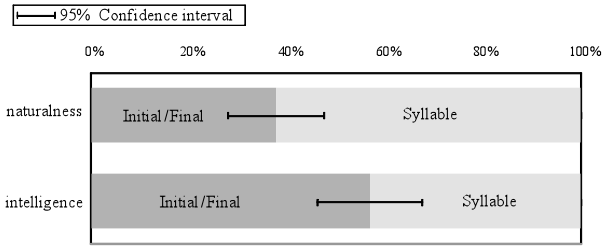

Figure 3. preference test between different units

III. EXPERIMENT EVALUATION

A. Experiment Conditions

The corpora of two female speakers (noted as F1 and F2), are used as training data for evaluation [11]. The speech signals are sampled at a rate of 16 kHz and quantified to 16-bit. MLSA (Mel Log Spectrum Approximation) is used to convert speech signals to mel-cepstral coefficients sequences in 5ms' interval. The speech parameter contained 0th through 24th MGC coefficients, log-scaled F0, including their delta and delta-delta features.

Both corpora contain more than 5000 phonetically balanced sentences. To evaluate the effectiveness using different size of training database, 1000/3000/5000 sentences were selected separately from these corpora.

B. Preference Evaluation

In the preference evaluation, listeners are asked to indicate their preference of the naturalness and the intelligibility between the synthetic speech by two systems:

- Syllable HMM based TTS system trained by 1000 sentences.
- Phone HMM based TTS system trained by 1000 sentences.

Only F1's training data is used in preference test. Ten listeners participated in the test. The listeners are researchers or students aged 20 to 35 years old and are native Mandarin speech experts with no known hearing problems. Each listener evaluated 8 sentences for each method.

Fig. 3 shows the result of the test. With 1000 sentences, Syllable HMM based TTS system performs better in naturalness, while Phone HMM based TTS system performs better in intelligence.

C. MOS Evaluation

In the MOS evaluation, listeners are asked to assess the overall quality of each sentence according to the following scale: (1) bad; (2) poor; (3) fair; (4) good; (5) excellent. The mean opinion score (MOS) is the arithmetic mean of all the scores from each individual. All listeners are native Mandarin speech experts with no known hearing problems.

The MOS evaluations of two corpora were performed separately. For each speaker's corpus, the following 6 voices were evaluated together: (1-3) synthetic speech of Syllable HMM based TTS with 1000/3000/5000 sentences' training data; (4-6) synthetic speech of Phone HMM based TTS with 1000/3000/5000 sentences' training data.

For F1's corpus, 13 listeners participated in the test. Each listener evaluated 5 sentences for each voice. Fig. 4 shows the result of the MOS test. Both Syllable HMM based TTS system and Phone HMM based TTS system performs better with the increase of training data. With 1000 sentences' training data, Phone HMM based TTS system gets a higher score than Syllable HMM based TTS system. With 3000 sentences and 5000 sentences' training data, Syllable HMM based TTS system outperform Phone HMM based TTS system.

For F2's corpus, 7 listeners were invested to evaluate 10 sentences for each voice.

Fig. 5 shows the result of the MOS test with F2's corpus. This time, Syllable HMM based TTS system get a higher score the Phone HMM based TTS system for all training data size. With 1000 sentences' training data, the scores of two systems are close. With 3000 sentences' training data and 5000 sentences, Syllable HMM based TTS system performs significantly better than Phone HMM based TTS.

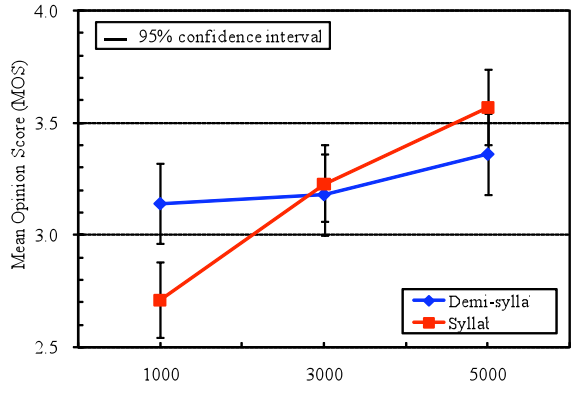

Figure 4. MOS test on F1's corpus

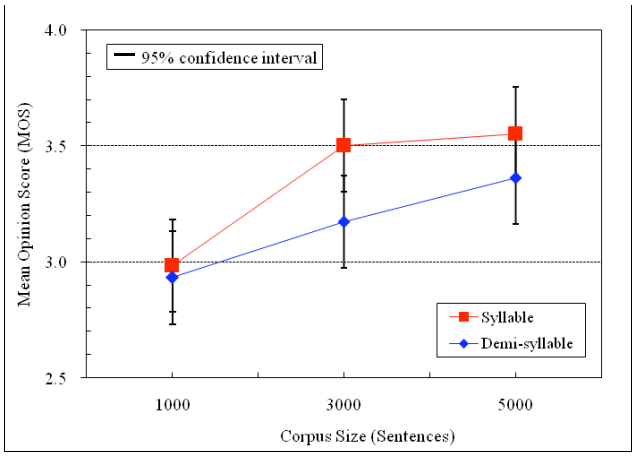

Figure 5. MOS test on F2's corpus

D. Discussion

Preference evaluation result indicates Syllable HMM based TTS gets a better naturalness, which is probably introduced by the more natural prosody inside the syllable. However, because the number of syllables is much larger than the number of the phones, there can be degradation of intelligence caused by lack of data for some syllable with 1000 sentences' training data.

However, as shown in MOS evaluation, when the training data size increases from 1000 sentences to 3000 sentences, the performance of Syllable HMM based TTS system increase much more rapidly than Phone HMM based TTS system. This is probably because that the increase of training data counters the lack of data problem in Syllable HMM based TTS.

We notice that both Syllable HMM based TTS and Phone HMM based TTS have their advantages and disadvantages with 1000 sentences' training data. We are currently researching on HMM based TTS that can automatically choose to use syllable and phone unit according to the training data of specific syllable [12].

IV. CONCLUSIONS

In this paper, we compare the performance of two HMM based Mandarin TTS systems using syllable and phone as basic units respectively with different size of training data. Two female speakers' corpora are used as training data for evaluation. Preference evaluation result shows that Syllable HMM based TTS performs better in naturalness while Phone HMM based TTS performs better in intelligibility with 1000 sentences' training data. However, as shown in MOS evaluation, when the training data size increases from 1000 sentences to 3000 sentences, the MOS score of Syllable HMM based TTS system increases much more rapidly than Phone HMM based TTS system. For two speakers' corpora, Syllable HMM based TTS system outperforms the system using phone as basic unit with 3000 and 5000 sentences' training data.

ACKNOWLEDGMENT

This work was partly supported by the University Joint Research project of IBM China Research Lab. This work was also supported by the National Natural Science Foundation of China (60805008, 60433030, 90820304), the National Basic Research Program of China (973 Program) (No.2006CB303101), the Ph.D. Programs Foundation of Ministry of Education of China (No.200800031015) and the National High Technology Research and Development Program ("863" Program) of China (No. 2007AA01Z198).

REFERENCES

[1] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," Proc. of IEEE Workshop on Speech Synthesis, 2002.

[2] Z.H. Ling, Y.J. Wu, Y.P. Wang, et al. USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method [A]. Proc. of ICSLP Satellite Workshop, Blizzard Challenge[C]. 2006.

[3] Yao. Qian, Soong Frank, Yining Chen, Min Chu, "An HMM-Based Mandarin Chinese Text-To-Speech System," LNCS vol. 4274/2006, pp. 223-232, 2006.

[4] J. Yu, M Zhang, J Tao, X Wang, "A Novel HMM-Based TTS System Using Both Continuous HMMs and Discrete HMMs," Proc. ICASSP, pp.IV-709-IV-712, 2007.

[5] S.Y. Kang, Z.W. Shuang, Q.S. Duan, Y. Qin, L.H. Cai. "Voiced/Unvoiced Decision Algorithm for HMM-based Speech Synthesis," Proc. of INTERSPEECH, pp. 412-415, 2009.

[6] T. Fukada, K. Tokuda, T. Kobayashi, and I. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in Proc. ICASSP, 137-140, 1992.

[7] T. Yoshimura, K. Tokuda, T. Masuko, et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. of Eurospeech. 1999, vol. 5, 2347-2350.

[8] K. Shinoda and T.Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn.(E), vol. 21, pp. 79–86, 2000.

[9] T. Masuko, K. Tokuda, and T. Kobayashi, "Speech synthesis from HMMs using dynamic features," Proc. of ICASSP, 389-392, 1996.

[10] S.Y. Kang, Q.S. Duan, Z.W. Shuang, Y. Qin, L.H.Cai, "A research of F0 extraction and prediction algorithm for HMM-based speech synthesis," in Proc. NCMMSC, 2009.

[11] Lianhong Cai, Dandan Cui, Rui Cai, TH-CoSS,a Mandarin Speech Corpus for TTS[J], Journal of Chinese Information Processing, 2007-02

[12] Q.S. Duan, S.Y. Kang, Z.W. Shuang, Z.Y. Wu, Y. Qin, L.H. Cai, "A modeling-unit selection algorithm approach to HMM-based speech synthesis on Chinese", in Proc. NCMMSC, 2009.