AN UNIFIED AND AUTOMATIC APPROACH OF MANDARIN HTS SYSTEM

Yong Guan[1, 2], *Jilei Tian*[2], *Yi-Jian Wu*[3], *Junichi Yamagishi*[4] *and Jani Nurminen*[5]

[1]Beijing University of Posts and Telecommunications, China [2] Nokia Research Center, Beijing, China

[3] Microsoft, China [4] University of Edinburgh, UK [5]Nokia Devices R&D, Finland

ABSTRACT

Most studies on Mandarin HTS (HMM-based text-to-speech system) have taken the initial/final as the basic acoustic units. It is, however, challenging to develop a multilingual HTS in a uniformed and consistent way since most of other languages use the phoneme as the basic phonetic unit. It becomes hard to apply cross-lingual adaptation which need map phonemes from each other, particularly in the case of unified ASR and HTS system due to the phoneme nature of most of the ASR systems. In this paper, we propose a phoneme based Mandarin HTS system, which has been systematically evaluated by comparing it with the initial/final system. The experimental results show that the use of phoneme as the acoustic unit for Mandarin HTS is a promising unified approach, thus enabling better and more uniform development with other languages while significantly reducing the number of acoustic units. The flat-start training scheme is also evaluated to show that the phoneme segmentation problem is solved without any performance degradation for phoneme based Mandarin HTS system. This performs an automatic approach without dependency with particular ASR system.

Index Terms—speech synthesis, Mandarin HTS, flat-start training, speaker adaptation

1. INTRODUCTION

HTS [1] has recently been found to be of comparable quality with the state-of-the-art concatenative text-to-speech (TTS) systems [2][3]. Furthermore, HTS can easily change the voice characteristics of synthesized speech by using a speaker adaptation technique originally developed for speech recognition. It has been shown that supervised speaker adaptation can yield high quality synthetic voices with data of a lower order of magnitude than required to train a speaker-dependent model or to build a basic unit-selection system [4][5][6]. Besides, with uniform framework, HTS system makes multi-lingual research more easily accessible for universal researchers.

Mandarin is monosyllabic language where each syllable can be conventionally decomposed into an initial/final format. The initial is the initial consonant and the final is the vowel (or diphthong) part with an optional medial or a nasal ending in the syllable. There are a smaller number of phonemes that create more difficulty in precise segmentation and imply a larger search space for unit selection. Thus the initial/final is a natural choice of the acoustic unit in concatenative Mandarin TTS since it brings more integrity of acoustic unit, and reasonable search space. Furthermore, the initial/final format is also commonly taken as the acoustic modeling unit in Mandarin HTS systems partially because most of the Mandarin speech synthesis databases were already labeled with the initial/final [3][4][7]. However, a problem is encountered when sharing acoustic units with other language, such as in a multilingual HTS framework and cross-lingual adaptation research work [8][9]. This is simply because most of theother languages take the phoneme as the basic acoustic modeling unit. A similar problem arises when unifying ASR and HTS models or performing joint adaptation across ASR and HTS models due to the phoneme nature of most ASR systems [10]. So we expect one phoneme base Mandarin HTS system which performs at least not worse than the initial/final one.

In this study, we carried out the experiments to compare phoneme with initial/final as the sub-word acoustic modeling unit for Mandarin HTS system. The evaluation results showed that the phoneme-based system performed a little better than initial/final-based system particularly in the adaptation case where a small amount of adaptation data is given. This is a promising result since a phoneme-based system is much easier to use in a cross- and multi-lingual system, where the most other languages use the phoneme as the basic modeling unit. Furthermore, the phoneme boundary is usually difficult to label well and truly compared with initial/final case and depends on a particular ASR system to do force-alignment. However, for some languages, maybe you have no ASR system in hand to do the force-alignment for TTS purpose. In this paper, we investigate the flat-start training scheme for phoneme based Mandarin HTS system to show that the unit segmentation can be easily handled without dependency on ASR force-alignment. On the other hand, we can train the model started with uniformed segmentation instead.

The rest of the paper is organized as follows. In Section 2, the initial/final and phoneme-based Mandarin HTS systems are proposed and evaluated with supervised adaptation and a variety of enrollment data. Furthermore, in section 3, we investigate the flat-start training scheme for phoneme-base Mandarin HTS system. Finally, the conclusions are drawn.

2. PHONEME-BASED MANDARIN HTS SYSTEM

In this paper, the HTS system is built using the framework from the HTS-2007 system [4] [6], which was a speaker-adaptive system entered for the Blizzard Challenge 2007 and Blizzard Challenge 2008. The HTS-2007 system consists of four main components:

speech analysis, average voice training, speaker adaptation and speech generation. To evaluate the phoneme-based HTS, we also built a speaker dependent system, in which we train the HTS model using speech data from a single speaker only.

Following algorithms and technologies were applied in the speaker-adaptive framework:

- Speech analysis: A high quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [11] was used, in conjunction with the mixed excitation [12].
- Training: To simultaneously model the duration for the spectral and excitation components of the model, the MSD hidden semi-Markov model (MSD-HSMM) [13][14] was applied. In order to reflect within-frame correlations and optimize all the acoustic feature dimensions together, semi-tied covariance (STC) modeling [15] was applied to enable the use of full-covariance Gaussians in the HSMMs [16].
- Speaker adaptive training (SAT): A speaker-adaptive approach in which average voice models are created using data from several speakers. The average voice models may then be adapted using speech from a target speaker (e.g. [17]). We adopted several techniques for training the average voice model, such as a SAT algorithm [18].
- Speaker Adaptation: adopting advanced adaptation algorithm called constrained structural maximum a posteriori linear regression (CSMAPLR) [17]
- Speech generation: Generating smooth and natural parameter trajectories from HMMs considering the global variance (GV) [19] [20].

To evaluating the phoneme-based HTS, we also built a speaker dependent system, in which, we train HTS model using speech data from single speaker only.

2.1 Phoneme vs. initial/final

Though several acoustic processing components in HTS are language-independent, there are still a few language dependent issues that need to be addressed. The sub-word acoustic unit needs to be defined in Mandarin HTS. Mandarin is a tonal syllabic language, where one syllable usually consists of one initial and one final. In the Mandarin speech synthesis system, initial/final is commonly deployed as the basic acoustic unit. However, the acoustic units in HTS of most of the other languages and even Mandarin ASR are usually phoneme-based. Some of the cross-language or multi-language techniques require a common acoustic unit, such as the unified phoneme set. Also, in order to bridge the gap between Mandarin HTS and Mandarin ASR, it is beneficial to explore the performance of phoneme-based Mandarin HTS where the same phoneme set can be used as for Mandarin ASR.

We adopted the SAMPA-C phoneme [21] set which includes 23 consonants and 18 vowels/semivowels. An extended LC-STAR lexicon was used. At the same time, for the initial/final set, the inventory consisted of 21 initials and 38 finals.

For the phoneme-based system we conducted a preliminary experiment with both 3-state and 5-state models. The phonetic and linguistic contexts contain phonetic, segment-level, syllable-level, word-level and utterance-level features including tonal features. It was found that the quality of the synthetic speech from these two phoneme-based systems was similar. However, the stability of speech using 5-state models appeared to be better. Therefore, we used 5-state HMM phoneme model in the subsequent experiments. In the initial/final-based system, 5-state left-to-right HMMs are used for each unit.

2.2 Experimental evaluations

The speech database used for building the current Mandarin TTS systems is licensed from iFlyTek. This database is specifically designed for speech synthesis containing data from 6 speakers having 3 male and 3 female speakers with 1000 phonetically balanced utterances per speaker. There are 6000 utterances corresponding to 12 hours speech in total. The labels contain initial/final, tone, part of speech and prosodic boundary information. The labels have been automatically parsed from the text transcription. The speech database is high quality recorded in sound-proof rooms using high-quality microphones, and the waveforms are sampled at 16kHz and coded as 16 bit.

The experiments are conducted to assess the performance of speaker-dependent and speaker-independent with supervised adaptation. It compares the performance of phoneme-based and initial/final Mandarin HTS. For speaker-dependent training, the speech data from female speaker ‘f2’, comprising 1,000 utterances, is used. For average voice model training, the remaining speech data from the other five speakers is used. For speaker adaptation, we compared the performances obtained using varying amounts of adaptation data ranging from 10 to 1000 utterances from the target speaker.

The subjective listening test is designed as follows. 697 test sentences, including 647 sentences from the news genre used to evaluate naturalness and similarity and 50 semantically unpredictable sentences used to evaluate intelligibility, are synthesized. The listening test set is randomly selected as a subset.

To evaluate naturalness and similarity, a 5-point mean opinion score (MOS) and comparison category rating (CCR) tests were conducted. The scale for the MOS test ranged in scale from 1 to 5 where 5 denotes completely natural and 1 stands for completely unnatural. The scale for the CCR tests were defined so that 5 indicates it sounds like exactly the same person and 1 indicates it sounds like a totally different person, compared to a few natural example sentences from the reference speaker. To evaluate intelligibility, the subjects were asked to transcribe semantically unpredictable sentences and average pinyin and tone error rate (PTER) was calculated.

The evaluations were conducted using a standard web service through the browser and the number of listeners was 91, of which almost all took the test in a quiet laboratory environment using headphones.

We built and evaluated 10 voices denoted as A~J. Among them, A, B, C and D are speaker adaptive initial/final systems adapted using 10, 20, 100 and 1000 sentences separately, E, F, G and H are corresponding phoneme based systems, and I and J are initial/final and phoneme based speaker dependent systems. The experimental results are summarized in Table 1.

By comparing phoneme-based system with corresponding initial/final system in Table 1, we found that phoneme-based speaker adaptive system outperforms initial/final system, especially with 10, 20 and 100 adaptation sentences, both in MOS and PTER. However, not much difference was found for speaker dependent

HTS voices between initial/final and phoneme based systems in terms of MOS and PTER.

Table 1. *Subjective listening evaluation results for Mandarin HTS. SD refers to speaker-dependent and SA refers to supervised speaker-adapted system.*

	Adapt	Sub-word	#Sent	MOS	PTER
A	SA	Initial/Final	10	1.7	39
B	SA	Initial/Final	20	1.5	56
C	SA	Initial/Final	100	2.6	37
D	SA	Initial/Final	1000	3.0	34
E	SA	Phoneme	10	2.3	39
F	SA	Phoneme	20	2.1	33
G	SA	Phoneme	100	2.9	32
H	SA	Phoneme	1000	3.1	33
I	SD	Initial/Final	1000	3.9	19
J	SD	Phoneme	1000	3.8	19

MOS scores for the naturalness in the subjective listening test are also summarized in Figure 1. It can be clearly seen that phoneme based E and F outperform initial/final counterpart, A and B in naturalness having a small number of adaptation data. For other comparisons between phoneme and initial/final systems, similar performances are found.

For the similarity listening test, CCR scores are summarized in Figure 2. We can see that generated waves from phoneme based systems G and H have higher similarities with reference speeches than those from initial/final system, denoted as C and D. Consistent observations are found for other comparable settings between phoneme and initial/final systems.

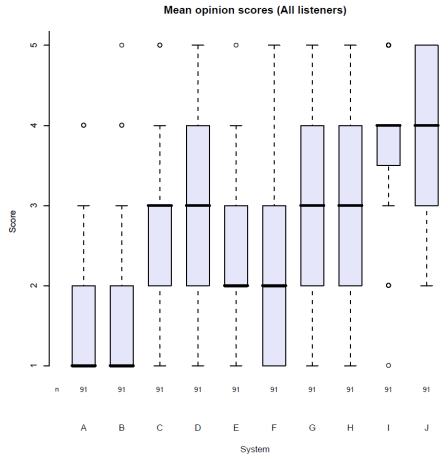

Figure 1: *Listening test results for Mandarin: naturalness.*

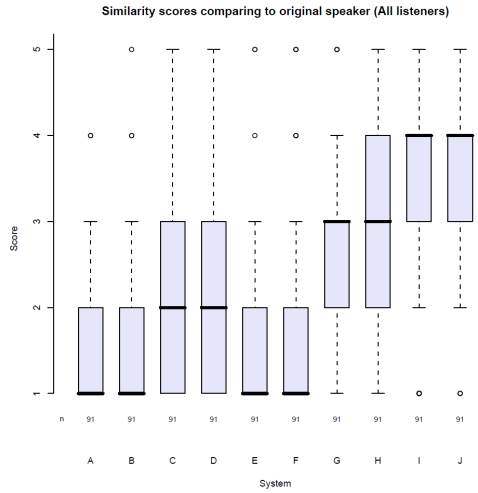

Figure 2: *Listening test results for Mandarin: similarity.*

Clearly the phoneme-based HTS has performed better than the initial/final system, given the same amount of adaptation data, and the difference is significant particularly for the case where a small amount of enrollment data is applied. This is a promising result since a phoneme-based HTS would be much easier to use in a cross- and multilingual HTS framework and offer several benefits as discussed above.

For small amounts of adaptation data, the initial/final-based HTS performs worse than the phoneme-based system. In general, initial/final-based HTS appears to require more adaptation data than phoneme-based HTS for supervised adaptation since initial/final systems have many more units considering tone. In comparison with Figure 1 and Figure 2, the phoneme-based Mandarin HTS can reach more reasonable performances of naturalness than of similarity especially when small amounts of adaptation are used.

By comparing speaker dependent with speaker adaptive systems, we found that the two speaker-dependent systems are significantly better than all speaker-adapted systems in terms of naturalness and similarity, and better than almost all in terms of intelligibility, even with all 1000 sentences as adaptation data. However, comparing the results to those from experiment in English [6], we may conjecture that the poor result in Mandarin speaker adaptive systems can be attributed to the use of a small number of training speakers for the average voice model. Future experiments will use large amounts of ASR data to train the average voice model, and we expect this to improve results as was the case for English.

3. FLAT-START TRAINING

Considering phoneme segmentation for Mandarin speech synthesis, we conducted a preliminary flat-start training experiment in speaker dependent phoneme-based Mandarin HTS system. We trained speaker dependent phoneme-based Mandarin HTS models for 6 speakers in iFlyTek.

To evaluate the flat-start training, uniformed time labels were used as a flat-start scheme to initially train speaker dependent HTS models. Then, the trained HTS model was used to realign the training data for generating refined time labels in the full context

labels. The realigned full context labels were used to train the HTS models for a second pass. We can iteratively obtain high quality results for the refined context labels and refined HTS models.

A formal subjective listening test was conducted to evaluate the synthesized voices with flat-start training. As a reference system, full context labels with time labels generated by force-alignment using the ASR model were used to train HTS models, referred as ASR_aligned_labeling.

For each HTS system, 10 test sentences which were excluded in the training data were synthesized, respectively. 5 Subjects were presented with a pair of synthesized speech from different training methods in random order, and then asked if the target voices are obviously worse 1, until obviously better 5, compared with the reference voices. Figure 3 shows the evaluation results.

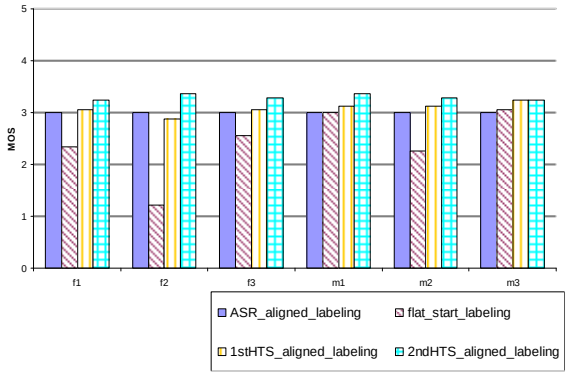

Figure 3: MOS scores of synthesized speech from flat-start training models.

Firstly, we evaluated the performance of the flat-start training system referred as flat_start_labeling system with uniformed timed label. It can be easily seen that the performance is different for case f2 than for the others. For speaker f2, most of the synthesized voices with flat-start labeling are obviously worse than the reference voices. However, for other speakers, the average scores show that the flat-start voices are slightly worse or comparable with reference ones.

To investigate the potential performance of flat-start training, we evaluate the performance of realigned time labeling system denoted as 1stHTS_realigned_labeling system, in which the time labels are generated with force-alignment using the HTS model using flat-start labeling. For speaker f2, the average score shows that the generated voices are still worse than the reference ones though the improvement to previous results is visible. For other speakers, the average score shows that the generated voices are comparable with reference ones. Since the voice quality of the 1stHTS_realigned_labeling system for speaker f2 is still something below the reference, we continue to evaluate the performance of having the second iteration of realigned time labeling denoted as 2ndHTS_aligned_labeling system in which the time labels are generated with force-alignment using the 1stHTS_realigned_labeling HTS models. It slightly outperforms the reference system.

From the score distribution of listening test, we can see that the system with realigned time labeling significantly improved with more iterations across all speakers. As a consequence, it performs equally well and or even slightly better than the reference system. It can be concluded that the flat-start training scheme can be used to develop the HTS system having good performance without dependency on ASR models, particularly after two rounds of realignment. The experimental result is promising not only for porting HTS systems to a new language without dependency on ASR system but also for unifying the labels between training and testing.

4. CONCLUSIONS

We carried out experiments to evaluate phoneme based Mandarin HTS systems compared with initial/final sub-word unit system and found that phoneme based Mandarin HTS systems performed similarly to the initial/final system. It has also found that the phoneme-based speaker adaptive system performed better when adapted with a small quantity of enrollment data. The initial/final system requires more adaptation data to reach good level of performance. The experimental results indicate that the use of phoneme-based Mandarin HTS is a promising approach, particularly when integrating the system with the other phoneme-based multilingual HTS and ASR systems.

We have also investigated a flat-start training scheme to train a phoneme based Mandarin HTS system. The subject listening test results show that flat-start training can perform at least equally well as the reference baseline system. The experimental result is promising not only for porting HTS systems to a new language without dependency on ASR system, but also unified the labels between training and testing.

5. ACKNOWLEDGEMENT

The authors would like to thank ANHUI USTC iFlyTek Co., Ltd. for permission to use their Mandarin speech database. The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project (http://www.emime.org). Junichi Yamagishi is partially supported by EPSRC.

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,” Proc. of EUROSPEECH, 1999.

[2] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” IEICE Trans. Inf. & Syst., vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[3] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, “USTC system for Blizzard Challenge 2006 an improved hmm based speech synthesis method,” in Proc. Blizzard Challenge 2006, Sep. 2006.

[4] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, “Speaker independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007,” in Proc. BLZ3-2007, Aug. 2007.

[5] Junichi Yamagishi, Takashi Nose, Heiga Zen, Tomoki Toda, Keiichi Tokuda ``Performance Evaluation of The Speaker-Independent HMM-based Speech Synthesis System "HTS-2007" for the Blizzard Challenge 2007," Proc. ICASSP 2008, 2008.

[6] J. Yamagishi, H.Zen, Y. J. Wu, T. Toda, K. Tokuda, “The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge” in Proc. BLZ4-2008, Sep. 2008.
[7] Y.-J. Wu, “Research on HMM-based speech synthesis,” in Ph.D. Thesis, University of Science and Technology of China, 2006.
[8] H. Ye and S. Young, Improving Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning." In Proc Interspeech 2005, Sep. 2005
[9] Y. J. Wu, S. King, K.Tokuda, “Cross-lingual speaker adaptation for HMM-based speech synthesis” in Proc. ISCSLP 2008, Dec. 2008
[10] J. Olsen, Y. Cao, G. Ding, X. Yang “A Decoder for large vocabulary continuous short massage dictation on embedded devices.” in Proc. ICASSP 2008, Sep. 2008.
[11] H. Kawahara, I. Masuda-Katsuse, and A. Cheveign´e, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187–207, 1999.
[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for HMM-based speech synthesis,” in Proc. EUROSPEECH 2001, Sep. 2001
[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multispace probability distribution HMM,” *IEICE Trans. Inf. & Syst., vol. E85-D, no. 3, pp. 455–464, Mar. 2002.*
[14] 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, no. 5, pp. 825–834, May 2007.*
[15] M. Gales, “Semi-tied covariance matrices for hidden Markov models,” *IEEE Trans. Speech Audio Process., vol. 7, pp. 272–281, Mar. 1999.*
[16] H. Zen, T. Toda, and K. Tokuda, “The Nitech-NAIST HMM based speech synthesis system for the Blizzard Challenge 2006,” *IEICE Trans. Inf. & Syst., vol. E91-D, no. 6, pp. 1764–1773, Jun. 2008.*
[17] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Trans. Speech, Audio & Language Processing, Jan. 2009*
[18] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A training method of average voice model for HMM-based speech synthesis,” *IEICE Trans. Fundamentals, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.*
[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *in Proc. ICASSP 2000, Jun. 2000, pp. 1315–1318.*
[20] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” *IEICE Trans. Inf. & Syst., vol. E90-D, no. 5, pp. 816–824, May 2007.*
[21] X. X. Chen, A. J. Li, G. H. Sun, and Z. G. Yu, “An Application of SAMPA-C for Standard Chinese”. in Proc. of International Conference on Spoken Language Processing (ICSLP), Oct. 2000, Beijing.