Evaluation of Flat Start Labeling for Phoneme based Mandarin HTS System

Yong Guan, Jilei Tian
Nokia Research Center, Beijing
{ext-yong.guan, jilei.tian}@nokia.com

Abstract

In this paper, we proposed a phoneme based Mandarin HTS speech synthesis system trained with flat start scheme. Conventionally the *full context labels with phonetic time segmentation are required for HTS training. The segmentation is generated by ASR force alignment using the pre-trained ASR models. Thus it brings the dependency on ASR while developing HTS system and causes different label in HTS between training and testing. Flat start labeling, which uses uniformed segmentation in label, was proposed and evaluated by comparing with segmentation using ASR mode as a reference. The subject listening test results showed that flat start scheme performs equally well as the reference system using ASR force alignment when realignment labeling using trained HTS model is iteratively applied. This result is very promising for efficiently developing and porting HTS system to a new language.*

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has become a mainstream method of speech synthesis because of its natural-sounding synthetic speech and its flexibility. Many efforts have been done to improve the performance of HMM based speech synthesis systems (HTS).

To reduce buzziness, mixed excitation was integrated into the basic system to replace the simple pulse or noise excitation [2]. A high quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [3] was also used. This enables the generation of better spectral parameters and consequently more natural synthetic speech. To alleviate over-smooth problem due to statistical processing, a parameter generation algorithm that considers the global variance (GV) of the trajectory being generated was proposed [4]. In order to reflect within-frame correlations and optimize all the acoustic feature dimensions together, semi-tied covariance (STC) modeling [5] was employed to enable the use of full-covariance Gaussians in the HSMMs [6]. To adapt the HTS system to a specific speaker model with a small amount of data, a speaker-adaptive approach and a series of related adaptive and adaptation algorithm have also been developed [7][8]. Compared with early buzzy and muffled HMM-based speech synthesis, the latest systems have dramatically improved quality of synthetic speech. More importantly, the development cost was significantly reduced when it was ported to another language, assuming the availability of a text processing front end and generated labels for the training and test data.

At the same time, it is painful to have dependency on ASR while developing HTS, since the full context labels for training need segmentation of the acoustic unit. The segmentation requires the ASR system that has to be developed including acoustic model, lexicon and the training corpus. It is definitely big task and high cost especially for new languages. In this study, we want to show research community that flat start scheme, which uses uniformed segmentation, can work equally well compared to reference system. Then we can use identical label in HTS for both training and testing meanwhile getting rid of ASR dependency. It is interesting to see how good HTS model can realign the segmentation since it is far different than ASR model because HTS models usually have only one or two mixtures and are trained on a limited amount of training data. In flat start scheme, the HTS models are initially trained with uniformed time labeling. On the other hand, the timing information of the segmentation is arbitrarily assigned.

In this paper, we used flat start uniformed labels to initially train a phoneme based Mandarin HTS system. The force alignment using the initially trained HTS models is applied for segmentation of the training utterances. Repeatedly the HTS models can be iteratively re-trained on the realigned training data. We

evaluated the flat start scheme by comparing them with the systems trained with commonly used ASR labeled system as the reference system. The subject listening test results showed that the flat start labeling scheme, particularly with refined HTS modeling process, can perform equally well to the reference approach.

The rest of the paper is organized as follows. First, general system is briefly overviewed in Section 2. In Section 3, several labeling methods, such as flat start, realignment with HTS models and segmentation with ASR models are experimented and evaluated. Finally, the conclusions are drawn in Section 4.

2. System description

In this paper, we evaluate phoneme based Mandarin HTS with flat start labeling and compare it with re-segmentation using HTS and ASR models.

The whole HTS speech synthesis system includes text processing front-end and HTS back end system. The HTS back end system includes training part and synthesis part. The training part is to train acoustic models given full context labels and acoustic features. The synthesis part is to generate speech waveforms given full context labels. The text processing front-end generates full context labels for HTS training and synthesis. Figure 1 illustrates an overview of the basic HMM-based speech synthesis system [1].

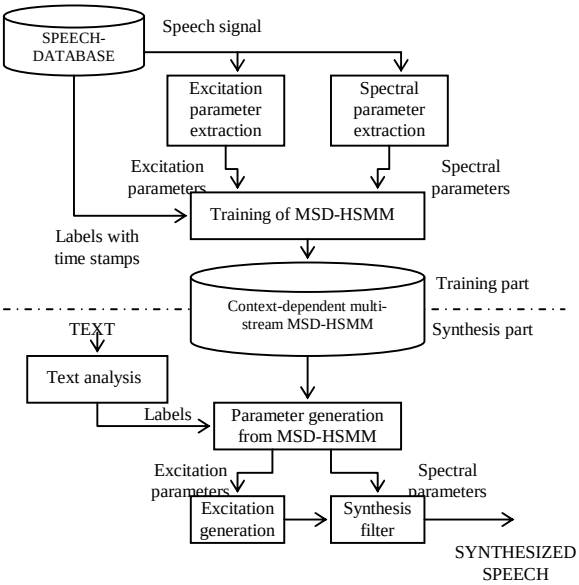

Figure 1: An overview of the basic HMM-based speech synthesis system.

2.1. Speaker dependent HTS back-end

For phoneme-based system we conducted a preliminary experiment with 5-state models. Following algorithms and technologies were applied in the speaker dependent framework:

- Speech analysis: A high quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [3] was used, in conjunction with the mixed excitation [2].
- Training: To simultaneously model the duration for the spectral and excitation components of the model, the MSD hidden semi-Markov model (MSD-HSMM) [9] [10] was applied. In order to reflect within-frame correlations and optimize all the acoustic feature dimensions together, semi-tied covariance (STC) modeling was applied to enable the use of full-covariance Gaussians in the HSMMs [6].
- Speech generation: Generating smooth and natural parameter trajectories from HMMs considering the global variance (GV) [4].

2.2. Phoneme based front-end

2.2.1 Phoneme based full context features

There are always at least a few language-dependent factors in the HTS system. For Mandarin HTS, the sub-word acoustic unit has to be defined. Many of earlier study used initial/final as the basic acoustic units, i.e., there is one HMM for each such unit, in each context. However, the acoustic units of HTS for the most other languages and even Mandarin ASR are usually phoneme-based. Since some of cross language techniques require a common acoustic unit it is beneficial to develop a phoneme-based Mandarin HTS system.

We adopted SAMPA-C phoneme [11] set which includes 23 consonants and 18 vowels/semivowels. An extended LC-STAR lexicon is used. The phonetic and linguistic contexts contain phonetic, segment-level, syllable-level, word-level and utterance-level features including tonal features, which is specific for Mandarin.

2.2.2 Flat start labeling

First, uniformed time labels were developed and used as flat start scheme to initially train speaker dependent HTS models. Then, the initially trained HTS model was used to realign the training data for generating

refined time labeling in the full context labels. The realigned full context labels were then used to repeatedly train the HTS models for a second round.

As a reference system, full context labels with real time labels generated by force-alignment with ASR model, were used to train HTS models.

3. Experiments and Evaluations

3.1. Database and measurement description

The speech database used for building the current phoneme based Mandarin HTS systems is licensed from iFlyTek. This database is specifically designed for speech synthesis and contains data from 6 speakers having 3 male and 3 female speakers with 1000 phonetically balanced utterances per speaker. In total, there are 6000 utterances amounting to 12 hours of speech. The recordings are high quality, made in sound-proof rooms using high-quality microphones, and the waveforms are 16kHz/16bit. For current speaker-dependent training, the speech data from female speaker 'f2' and male speaker "m2", comprising 1,000 utterances respectively, are used.

To evaluate the performance of the time labeling using both flat start scheme and realignment with HTS models compared to the reference system, formal subjective listening test was conducted. We compared these two kinds of synthesized voices with the reference ones, which were trained with segmentation using ASR models. The criterions are designed as shown in Table 1, with the voices segmented using ASR system as reference voices:

Table 1: Listening test measurement criterion (comparing with ASR model based time labeling)

Score	*Description*
1	obviously worse
2	slightly worse
3	no difference
4	slightly better
5	obviously better

For each HTS system, 10 test sentences, which were not contained in the training data, were synthesized, respectively. Subjects including 5 testees were presented a pair of synthesized speech from different methods in the random order, and then asked if the target voices are obviously worse, slightly worse, no difference, slightly better or obviously better than the reference voices.

3.2. Experiments and analysis

3.2.1. Flat start labelin*g*

First, we evaluated the performance of the flat start labeling, i.e. with uniformed timed label. Figure 2 shows the listening test results of the 10 sentences for f2 and m2 respectively. It can be easily seen that the performance is different between case f2 and m2. For speaker f2, the most synthesized voices with flat start labeling are obvious worse than the reference voices. However, for speaker m2, the average scores show that the flat start voices are slightly worse or comparable with reference ones.

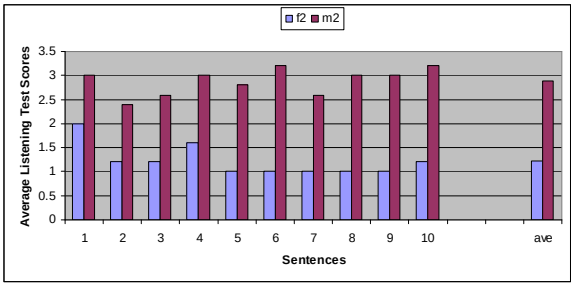

Figure 2 Average listening test scores of flat start, comparing with alignment with ASR model, for speaker f2 and m2

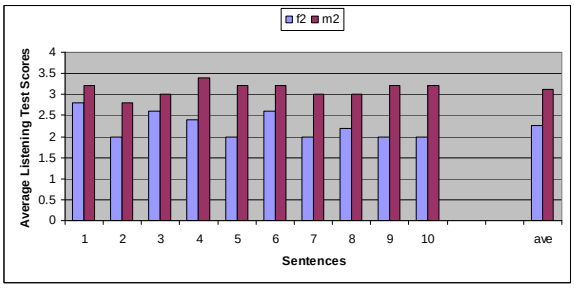

Figure 3 Average listening test scores of realignment with HTS model, comparing with segmentation with ASR model, for speaker f2 and m2

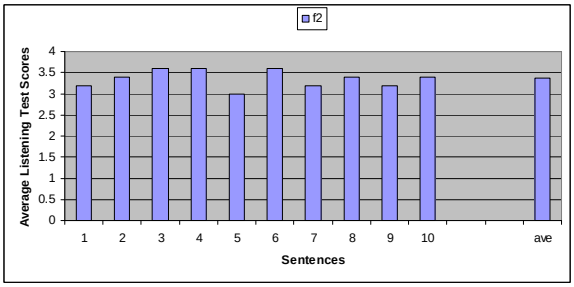

Figure 4 Average listening test scores of the second realignment with HTS model, comparing with alignment with ASR model, for speaker f2

3.2.2. Realigned labeling with HTS model

To investigate the potential performance of flat start labeling, we evaluate the performance of realignment time labeling system, in which the time labels are generated with force-alignment using HTS model initially trained with data having flat start labeling. The listening test results are listed in Figure3. For speaker f2, the average score shows that the generated voices are still worse than the reference ones though the improvement with previous results using initially trained model. For speaker m2, the average score shows that the most generated voices are comparable with reference ones.

3.2.3 Second realigned labeling with HTS model

Since the voice quality using the realigned labeling for speaker f2 are still below the reference, we continue to evaluate the performance of using the second realigned time labeling which the time labels are generated with force-alignment using the HTS models trained on the first realigned data. The listening test results are listed in Figure 4. It can be seen that it slightly outperforms the refernce voices.

3.3. Discussion

We redraw the distribution of listening test scores in Figure 5 which horizontal line indicates the position of average scores. From the listening test score distribution, we could see that system with realigned time labeling performs clearly the improvement step by step for both female speaker f2 and male speaker m2. As consequence, it shows equally well and even slightly better performance compared to the reference system. It can be concluded that the flat start labeling scheme can be used to develop HTS system having good performance without using ASR models, particularly after a few round of realignment to refine the trained HTS models.

4. Conclusions

We have proposed flat start labeling scheme to train a phoneme based Mandarin HTS system. The subject listening test results showed that the flat start labeling, especially after realignment with trained HTS models, can perform 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.

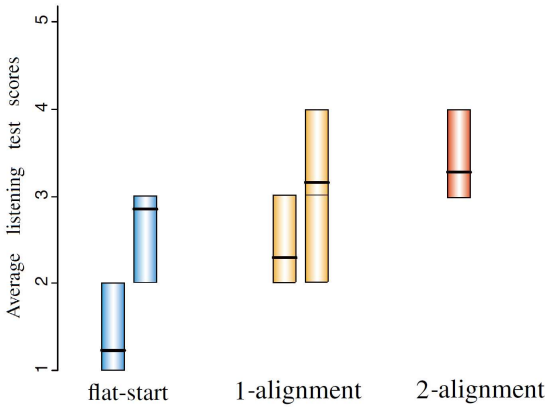

Figure 5 Listening test score distribution of flat start, 1-alignment with HTS model and 2-alignment with HTS model comparing with alignment with ASR model, for speaker f2 and m2

5. Acknowledgement

The work was partly funded from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement 213845 (the EMIME project (http://www.emime.org).

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] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," *in Proc. EUROSPEECH 2001, Sep. 2001*

[3] 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.*

[4] 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.*

[5] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process., vol. 7, pp. 272–281, Mar. 1999.*

[6] 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.*

[7] 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.*

[8] 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.*

[9] 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.*

[10] 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.*

[11] 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.*