A NOVEL HYBRID MANDARIN SPEECH SYNTHESIS SYSTEM USING DIFFERENT BASE UNITS FOR MODEL TRAINING AND CONCATENATION

Ran Zhang, Jianhua Tao, Ya Li, Zhengqi Wen

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

ABSTRACT

The hybrid speech synthesis system, which uses the acoustic model trained according to the criterion of Maximum Likelihood to select the proper candidates from the corpus, has become a hot topic in recent days. For this hybrid system, the performance is affected by the size of the base training unit and the base candidate unit. Most of existed hybrid systems use the same kind of base unit such as syllable or phone for both model training and concatenation. In Mandarin, initials and finals form the fundamental elements of pronunciation, and are always chosen as the base training unit for statistical parametric TTS system. In this paper a new hybrid Mandarin TTS system is proposed, which uses initial/final for model training and syllable for concatenation. Objective and subjective evaluations are conducted and the comparison results show that the hybrid system we proposed outperforms the traditional systems which use the same base unit for both processes with 4000 and 6000 sentences' corpus.

Index Terms—hybrid speech synthesis system, HMM, syllable, Mandarin speech synthesis

1. INTRODUCTION

While HMM-based speech synthesis system has been very successful in generating intelligible and stable speech[1, 2], the synthesized results are still lack of naturalness and quality. The hybrid speech synthesis system[3], which combines the unit-selection and HMM-based methods have eased these disadvantages and become a new spot in recent research.

In these hybrid systems, model training and unit selection are two essential and separable processes. The former process is exactly the same with the model training process in HMM-based Mandarin TTS system, and the posterior process also consist of target cost and concatenation cost like the traditional concatenation TTS system but with various definitions. So it's reasonable to suggest that base unit that works well in the above two mainstream TTS systems will also work well in this new hybrid system.

Mandarin is a tonal language, and each character in Mandarin is pronounced as a syllable. Most of syllables consist of an initial and a final and a tone, the others contain only a final and a tone. The initial part is treated as a consonant, while the final part is treated as a vowel. Since 424 base syllables and 5 tones are concerned, there are at most 2120 tonal syllables theoretically. However, many combinations are not existed at all, and only about 1300 tonal syllables are in use. So traditionally syllable is used as the base candidate unit for concatenation for its limited number and intra-syllable co-intelligibility[4, 5].

However, in most reported HMM based Mandarin TTS systems, initial/final or demi-syllable is used as the base unit for training and synthesis [6, 7] for two reasons:

1) Syllable-based HMM needs a much larger corpus to train than initial/final-based HMM. The total num of initial/final in Mandarin is 67, which is much smaller than syllable's num.

2) For some uncommon syllables, there might be few or none instances in the corpus, so the models of these syllables can't be trained very well, and the application field of the system will also be limited.

Although inconsistency does exist in the preferences of base unit in model training and unit selection for Mandarin hybrid speech synthesis system, to date, there are few published research papers that attempt to use different base units in those two processes.

The work by Ling and his fellows [8, 9] consider hierarchical unit selection and their system can choose base units of different size (initial/final or frame) by introducing several preset thresholds. And according to their conclusion, by increasing the size of base unit from frame to phone, the synthetic speech quality can be improved. However, limited by the corpus size (1000 sentences), the system do not choose syllable as the base unit for either training or concatenation and only consider smaller units for concatenation.

In this paper, a hybrid speech synthesis system which uses initial/final for training and syllable for concatenation is proposed. At the training stage, small base unit are used to train the HMMs, at the synthesis stage, syllable are used as the base unit for unit selection to reduce the search space and provide the intra-syllable smoothness. A set of experiments are conducted to further examine the differences of speech quality and naturalness caused by the choice of different base units in two processes.

Initial/final instead of phone is chosen as base unit for

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

training, because in our previous research[10], initial/final-based HMM outperforms phone-based HMM for Mandarin with 4000 and 6000 sentences' training corpus.

The rest of the paper is organized as follows. The next section further describes the system we proposed. Section 3 specifies the training process based on initial/final. Section 4 introduces the selection process based on syllable. The objective and subjective evaluation of proposed system and discussions are given in Section 5. Conclusions and future work are presented in Section 6.

2. SYSTEM OVERVIEW

The HMM-based hybrid speech synthesis system can be divided into two stages which are training stage and synthesis stage. Fig.1 gives a brief outline of the proposed system.

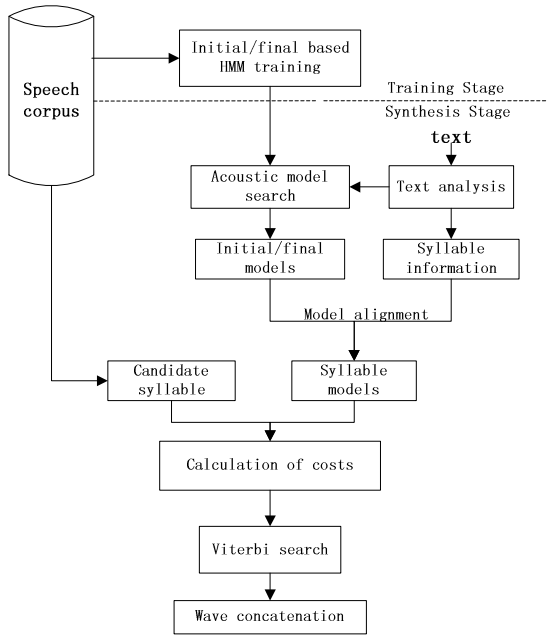

Figure 1: System overview

In the training stage, initial/final based context-dependent models is trained using acoustic features and label information. The feature vector is composed of F0 part and spectrum part. The F0 part includes a log-scale F0, its delta and delta-delta coefficients and it is modeled by multi-space probability distribution (MSD)[11]. The spectrum part consists of line-spectrum-pairs (LSP)[12], their delta and delta-delta coefficients. We then segment all the sentences in the speech database into states using the trained HMMs and get the initial/final boundary information. Finally, the state duration models are also trained.

In the synthesis stage, input text will be analyzed into labels, which are used to search relevant context-dependent models. After all the models are selected for the given text, they will be aligned according to the syllables they belonged to Finally, the target cost and concatenation cost of each candidate are calculated, and a Viterbi search [13] is conducted to get the best candidate for the concatenation. The definitions of cost functions will be specified in Section IV.

3. INITIAL/FINAL BASED MODEL TRAINING

A 7-state left-to-right with no skip HMM is applied to model each initial/final for the proposed hybrid TTS system.

The prosodic structure of this system includes 6 layers: initial/final (I/F), syllable (SYL), prosodic word (PW), prosodic phrase (PP), intonation phrase (IP) and sentence (SEN), 66 prosodic structure features and 24 pronunciation features are considered, all of them are shown in Table 1. The PSF stands for prosodic structure features, and PF indicates the pronunciation features.

Table 1: Context for HMM training

	Contexts
PSF	{position, reverse position} of I/F 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 I/F 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	{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}Articulator method of final

It is time-consuming to label the initial/final boundary manually, because the C-V boundary of the Mandarin syllable is hard to tell by human, and the labeling results always lack of consistency. Since the model quality will deteriorate a lot if the standard of the speech unit boundary definition is not consistent, auto segmentation results are used for model training.

4. SYLLABLE BASED UNIT SELECTION

In the late stage of training, the start/end timings of initial/final in each sentence of the corpus are provided by state segmentation. Since the mapping between syllables

and their initial/finals is already known, it is easy for us to get the syllable boundary; a syllable based candidate corpus can be built for selection.

Assuming target syllable n contains both initial and final, then the relative syllable model λ_i is composed of initial model λ_n^i and final model λ_n^f, and the candidate $u_n = \{u_1, ..., u_T\}$ can also be divided into the initial part u_n^i and final part u_n^f. So the likelihood of the candidate can be calculated as follows:

$$LL(u_n, \lambda_n) = LL(u_n^i, \lambda_n^i) + LL(u_n^f, \lambda_n^f) \quad (1)$$

If the target syllable only has the final part, there are neither λ_n^i nor u_n^i. However, we can still use Eq. (1) to calculate the likelihood of n by letting $LL(u_n^i, \lambda_n^i)$ to be 0. For a whole sentence that contains N syllables, the syllable candidate sequence can be written as $u = \{u_1, ..., u_N\}$, the syllable candidate sequence can be written as u^*, and the optimal one is determined using Eq. (2):

$$\begin{aligned} u^* &= \arg\max_u \sum_{n=1}^{N} LL(u_n, \lambda_n) \\ &= \arg\max_u \sum_{n=1}^{N} [LL(u_n^i, \lambda_n^i) + LL(u_n^f, \lambda_n^f)] \end{aligned} \quad (2)$$

$$LL(u_n^i, \lambda_n^i) = \log P(o_n^i \mid \lambda_n^i, Q_n^i) + \log P(T_n^i \mid \lambda_n^{i,dur}) \quad (3)$$

$$LL(u_n^f, \lambda_n^f) = \log P(o_n^f \mid \lambda_n^f, Q_n^f) + \log P(T_n^f \mid \lambda_n^{f,dur}) \quad (4)$$

, where o_n^i and o_n^f present the acoustic feature vectors of u_n^i and u_n^f; Q_n^i and Q_n^f stand for their state allocations. The likelihood of initial part and final part are calculated separately and treated equally. In order to facilitate unit search progress, Eq. (2) can be converted to the traditional form of a sum of "target cost" and "concatenation cost" [9]as follows:

$$u^* = \arg\min_u \{\sum_{n=1}^{N} TC(u_n) + \sum_{n=2}^{N} CC(u_{n-1}, u_n)\} \quad (5)$$

, where $TC(u_n)$ indicates the weighted sum of likelihood in u_n from 2nd frame to last but one frame, and $CC(u_{n-1}, u_n)$ calculates the sum of likelihood of last frame in u_{n-1} and first frame in u_n. Both costs can be calculated directly once the acoustic feature vectors are aligned towards their acoustic models frame by frame.

Dynamic programming search can be realized after the calculation of target cost and concatenation cost mentioned above. Then the same cross-fade technique for frame sized unit concatenation [14] is used to generate the speech waveform.

5. EXPERIMENT EVALUATION

5.1. Experiment Conditions

The database used for HMM training and unit selection consists of 10000 phonetically balanced Mandarin sentences. The speech signal is analysis at 5 ms frame shift and LSP order is 24 plus one extra order for energy. 7-state left-to-right with no skip HMM structure is adopted for each initial/final. To evaluate the effectiveness using different size of training database and corpus, 2000/4000/6000 sentences were selected separately from these corpora.

For comparison purpose, two traditional hybrid speech synthesis systems are also built. IF_1 uses initial/final as the base unit for both training and concatenation, while SYL_1 uses syllable as the base unit for both training and concatenation.

The training process of IF_1 is exactly the same with the proposed system, while the SYL_1's model Training is a little different. A 10-state left-to-right with no skip HMM structure is used for each syllable, and features of I/F layer are removed from the context.

5.2. Preference Evaluation

In the preference evaluation, listeners are asked to indicate their preference of the naturalness and the intelligibility of the synthetic speech towards proposed systems, SYL_1 and I/F_1, all the systems are trained by 2000 sentences, which are also used as candidate corpus.

Ten listeners participated in the test. The listeners are all graduate students whose major is speech, audio and language processing, and they are also native Mandarin speakers. Each listener evaluates 10 synthesized sentences for each system.

Figure 2 gives the evaluation result. In naturalness evaluation, proposed system and SYL_1 both have a good performance, and they both outperform the naturalness of I/F_1. In intelligibility evaluation, I/F_1 ranks the first, while SYL_1 ranks the last.

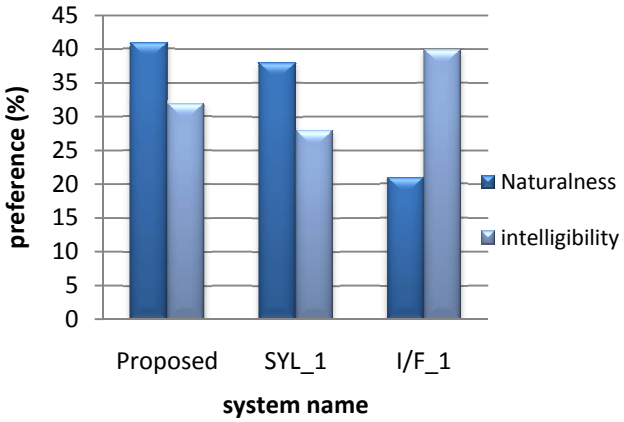

Figure 2: Preference evaluation results for the three systems

5.3. MOS Evaluation

In the MOS evaluation, 9 voices were evaluated together: (1-3) synthetic speech of the proposed system with 2000/4000/6000 sentences' training data; (4-6) synthetic speech of SYL_1 with 2000/4000/6000 sentences' training data;(7-9) synthetic speech of I/F_1 with 2000/4000/6000 sentences' training data.

The same ten listeners are asked to assess the overall quality of each sentence and required to give an opinion score from 1(bad) to (good). Each listener has to evaluate ten held out sentences for each system. The final average scores for each system are shown in Figure 3.

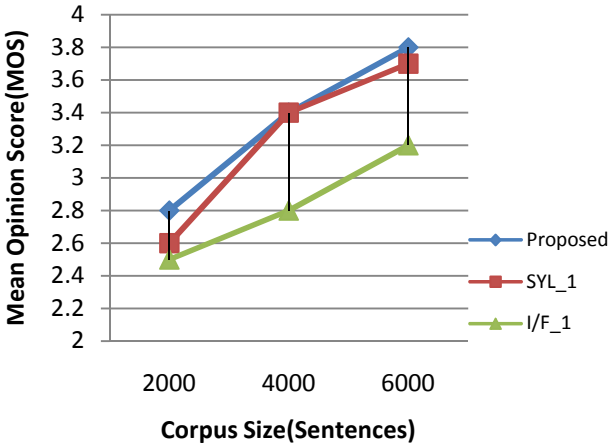

Figure 3: MOS test on different size of corpus

These results show that all of three systems perform better with the increase of corpus size. With 2000 sentences' training data, proposed system and SYL_1 get a similar score, and they perform slightly better than I/F_1.With 4000 sentences and 6000 sentences' training data, the MOS scores of proposed system and SYL_1 are almost the same, and they both much higher than I/F_1's MOS score.

From the Preference evaluation results, we can see that:

1) By using syllable as the base unit for concatenation, the naturalness of synthesis speech can be improved due to the intra-syllable co-intelligibility and the decrease of number of concatenation points.

2) Because the number of syllable is much larger than the number of the phones and our corpus are only 2000 sentences, the lack of candidates and training data can lead to degradation of intelligibility of SYL_1. That's why I/F_1 outperform the other two systems. Proposed system's models are the same with I/F_1, which can be well trained using small size of training data, so it works slightly better than SYL_1 on intelligibility evaluation.

In MOS evaluation, when the size of training data grows from 2000 sentences to 6000 sentences, the performance of all three systems has been improved, but the SYL_1 and proposed system increase much more rapidly than I/F_1, the reasons are as follows:

1) The increase of corpus has eased the shortage problem of candidate and training data.

2) I/F_1 use initial/final as base candidate unit, so most of concatenation points are located at C-V boundary. However, the listeners show small intolerance towards the discontinuity of concatenation points in C-V boundary. Although the acoustic model has been refined a lot with the growth of training data, the growth of MOS score is much slower than its contrasts.

According to the listeners' feedbacks, the proposed system can always choose more reasonable candidates than SYL_1, and the overall intonation is better, that's why the proposed system outperforms the SYL_1 with all three sizes of corpus.

6. CONCLUSION

In this paper, a novel hybrid Mandarin speech synthesis system using initial/final for training and syllable for concatenation is proposed. We specify the training process and unit selection process of the proposed system and their implementations. The evaluation results show that by introducing different base units in the training process and concatenation process, both the naturalness and intelligibility of hybrid system can be improved when the size of corpus increases.

However, the definition of target cost and concatenation cost are quite simple (weighted-sum-of-factors) and still need further tuning if better performance is expected. Since final attaches more significance to the pronunciation of syllable than initial, they should be treated differently, while in this paper, initial and final are treated as two equal parts. Last but not the least, I/F_1's poor performance on MOS test is largely caused by discontinuity of concatenation points at C-V boundary, a smooth algorithm may ease this problem and improve the performance of this system a lot. All of the problems mentioned above will be the tasks of our future work.

7. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC)(No.61273288, No.61233009, No.61203258,No. 61011140075, No. 90820303), and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

8. REFERENCES

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1229-IV-1232.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, pp. 1039-1064, 2009.

[3] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, *et al.*, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007.
[4] J. Tao, J. Yu, L. Huang, F. Liu, H. Jia, and M. Zhang, "The WISTON text to speech system for Blizzard 2008," in *Blizzard Challenge Workshop*, 2008.
[5] F.-c. Chou and C.-y. Tseng, "Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, pp. 893-896.
[6] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *Chinese Spoken Language Processing*, ed: Springer, 2006, pp. 223-232.
[7] J. Yu, M. Zhang, J. Tao, and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-709-IV-712.
[8] Z.-H. Ling and R.-H. Wang, "HMM-based unit selection using frame sized speech segments," in *INTERSPEECH*, 2006.
[9] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1245-IV-1248.
[10] S.-F. Pan, "Research on High Naturalness Statistical Parametric Speech Synthesis," Doctor, Institute of Automation, Chinese Academy of Sciences, 2011.
[11] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 229-232.
[12] I. V. McLoughlin, "Line spectral pairs," *Signal processing,* vol. 88, pp. 448-467, 2008.
[13] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 373-376.
[14] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.