A NOVEL HMM-BASED TTS SYSTEM USING BOTH CONTINUOUS HMMS AND DISCRETE HMMS

Jian Yu[(1)], Meng Zhang[(2)], Jianhua Tao[(3)], Xia Wang[(4)]

[(1)(2)(3)]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[(4)]Nokia Research Centre, China
{jyu[(1)], mzhang[(2)], jhtao[(3)]}@nlpr.ia.ac.cn, [(4)]xia.s.wang@nokia.com

ABSTRACT

The conventional HMM-based speech synthesis system (HTS) has encountered two over-smoothing problems in both time domain and frequency domain. To resolve the two problems, the paper presents a new HTS framework using both continuous HMMs and discrete HMMs. By the replacement of spectral envelope represented by continuous Gaussian distribution with that represented by discrete codevector, the over-smoothing problem in frequency domain can be resolved, and by the replacement of the parameter generation algorithm using dynamic features with a new well-designed codevector selection algorithm, the over-smoothing problem in time domain can also be better resolved. Experimental results show that the using of both continuous HMMs and discrete HMMs significantly improves the voice quality of synthesized speech.

Index Terms **—** Speech Synthesis, Hidden Markov Model, Speech Processing.

1. INTRODUCTION

In recent years, a kind of corpus-based speech synthesis system based on hidden Markov models (HMM) has been developed. In the system, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs, and the parameters are generated from HMMs by using dynamic features [1][2][3]. Although the current performance of HMM-based speech synthesis system is quite good, the synthesized speech is still a little buzzy and muffled compared with the result of concatenation-based system. This fact comes from two over-smoothing problems in both frequency domain and time domain, in which the over-smoothing in frequency domain makes the formant position unclear and the over-smoothing in time domain makes the synthesized speech lose too much detailed information. As a consequence, the formant trajectory of synthesized speech becomes unclear and unstable, which makes the speech sound muffled. Some research efforts have been carried out for handling these problems: (i) designing a new speech parameter generation algorithm considering global variance to improve the variance of the generated trajectory [4]; (ii) applying a minimum generation error based HMM training method to solve the inconsistency between training and synthesis [5].

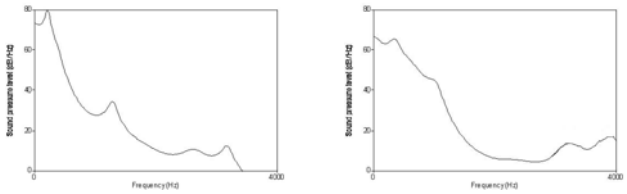
The paper presents a new HMM-based TTS system using both continuous HMMs and discrete HMMs. In the method, the real continuous spectrum parameters are represented as discrete codevector indexes based on vector quantization (VQ). The codevector comes from real spectrum parameter, so it keeps entire details in spectral envelope, while in conventional HTS system, the spectral envelope is represented by Gaussian distribution whose parameters are excessively smoothed by statistic processing. By the method, the over-smoothing problem in frequency domain is resolved. On the other hand, in parameter generation part, a well-designed codevector selection algorithm is proposed to replace the generation algorithm using dynamic features. Discrete HMMs are constructed to obtain the output probability of codevector and multi-space probability distribution HMMs (MSD-HMM) are constructed to generate the formant trajectory, which are both important criterions for the codevector selection. In addition, some statistic results such as the concatenation probability of codevectors are also taken into accounted. Based on these kinds of criterions, the generated spectrum parameter keeps more detailed information and the corresponding formant trajectory is stable and clear. By this method, the over-smoothing problem in time domain is also better resolved.

The paper is organized as follows: In section 2, we briefly review the two over-smoothing problems encountered by conventional HTS system. Section 3 introduces the structure of the new HTS system in detail, including both training part and synthesis part. Section 4 makes an evaluation which shows the performance of the new system. Finally, the summarization of the paper is presented in section 5.

2. THE TWO OVER-SMOOTHING PROBLEMS

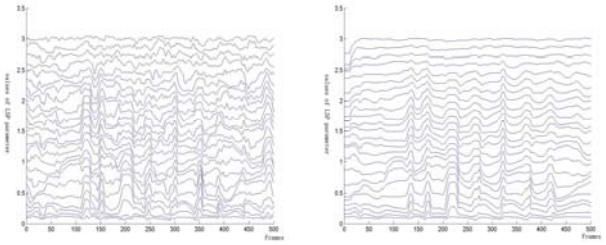
2.1. Over-smoothing problem in frequency domain

It has been studied that the formant position is very critical to the voice quality of synthesized speech. However, in



(a) Real spectrum (b) Synthesized spectrum

Fig.1. the over-smoothing problem in frequency domain



(a) Real LSP trajectory (b) Synthesized LSP trajectory

Fig.2. the over-smoothing problem in time domain

conventional HTS system, the spectral envelope of each frame is excessively smoothed due to the statistical processing in Baum-Welch re-estimation [10], which leads to unclear formants and muffled speech. Fig 1 shows a comparison between real spectrum and synthesized one.

2.2. Over-smoothing problem in time domain

The over-smoothing problem in time domain means that the predicted spectrum parameter trajectory, particularly, 24-order Linear Spectral Pair (LSP) trajectory in this paper, is too smooth to carry enough detailed information. In current HTS system, a phone is always represented by 5-state left-to-right HMM structure. If the state duration is too long, only one or several Gaussian functions can not depict tiny variations of LSP trajectory, which causes the over-smoothing problem in time domain, as shown in Fig 2.

3. THE STRUCTURE OF THE NEW HTS SYSTEM

In the new HTS system, to resolve the over-smoothing problem in frequency domain, codevectors which come from real speech are used to represent spectral envelopes, and to resolve the over-smoothing problem in time domain, a codevector selection algorithm is proposed to make sure that the selected codevector sequence has clear and stable formant trajectory. Figure 3 shows the framework of the new HTS system, which consists of two procedures: the training and synthesis procedure.

3.1. The training procedure

In the training procedure, the continuous-valued LSP vectors are represented by indexes of the codevector with the smallest distortion. Then, discrete HMMs, continuous

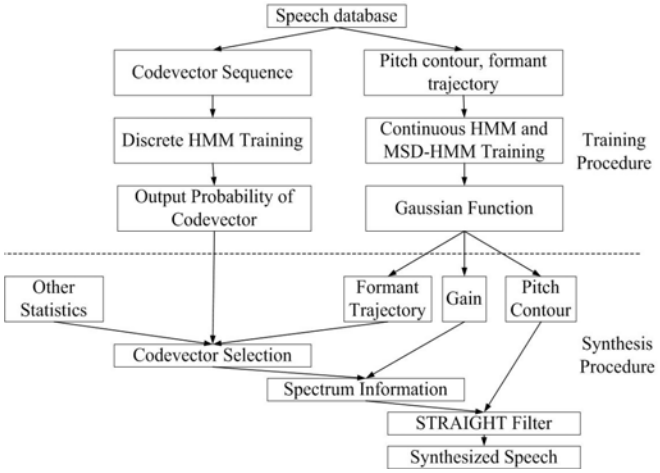


Fig.3. An overview of the new HTS system

Original	First VQ stage	Second VQ stage
0.09	0.03	0.01

Table 1. The MSE between retrieved LSP codevectors and real LSP vectors in different VQ stage

HMMs, and MSD-HMMs are constructed to obtain the output probability of codevector, gain trajectory, pitch contour and formant trajectory, respectively.

3.1.1 Vector quantization and discrete HMM training

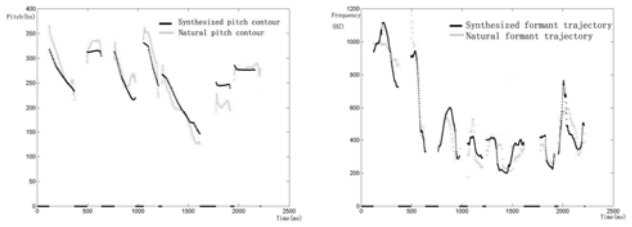
In the vector quantization procedure, there are some computing complexity problems on the direct use of a large single codebook. Therefore, a two-stage vector quantization method is adopted [7] to resolve it. In the method, the second stage encoder encodes the error between the original vector and the reconstruction generated by first encoder. Table 1 shows the performance of the two-stage vector quantization.

Based on the constructed codebook, the speech signal is represented by the index of its best match codevector. The training procedure of the method is quiet similar with the standard discrete HMM training method described in HTK [10], except that more contextual features are involved. After the training procedure, the output probabilities of codevectors are obtained, which is one of the most important criterions for the codevector selection.

It should be mentioned that all codevectors are selected from real speech so they keep entire details of real spectral envelope. Therefore, by the replacement of spectral envelope represented by continuous Gaussian distribution with that represented by such codevectors, the over-smoothing problem in frequency domain can be resolved.

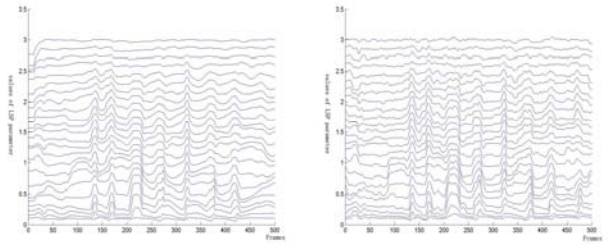
3.1.2 Continuous HMM and MSD-HMM training

Another part of LPC-based spectrum, the gain trajectory, is modeled using the standard continuous HMM. However, because there are no definitions of pitch contour and formant trajectory in unvoiced part of speech, it's hard to apply the standard HMM directly. To resolve that problem,



(a) Pitch contour (b) First formant trajectory

Fig.4. A comparison of generated pitch contour and formant trajectories with real ones (the second formant trajectory is not listed because of space limitation)



(a) Conventional HTS (b) New HTS

Fig.5. A comparison of generated LSP trajectories by two methods (it should be mentioned that the selected LSP trajectory by New HTS is smoothed by STRAIGHT filter)

a multi-space probability distribution HMM (MSD-HMM) which can cope with sequence of observation vectors with variable dimension including zero-dimensional observations (discrete symbols) is adopted to model the pitch contour and formant trajectory [9]. In fact, the MSD-HMM can be considered as a mixture of continuous HMM and discrete HMM.

By using corresponding algorithm for speech parameter generation, we generate pitch contours and formant trajectories which are close to corresponding parameters of natural speech, as shown in Fig 4. Because listener is more sensitive to low formants, only first and second formant frequencies are finally used in the model.

3.2. The synthesis procedure

To find the most appropriate codevector sequence, the well-studied unit selection framework widely used for concatenation-based speech synthesis is adopted. The main difference is that, in concatenation-based TTS, considered units are phones, diphones, syllables or even phrases, whereas in current task, the base unit are set to be a single codevector, i.e., a single speech frame.

The selected codevector sequence should meet following requirements: first, the selected codevector should be consistent with current phone identity, which makes the synthesized speech sound intelligible. Second, the corresponding formant trajectory should be clear and stable. Finally, to carry enough detailed information, the generated spectrum parameter trajectory must not be excessively smoothed. Based on these criterions, several cost functions are designed to direct the codevector selection.

3.2.1 The state output probability of discrete HMM

Codevectors whose output probabilities are higher than a set threshold will be selected as candidates and the probability value itself is one of cost functions in the codevector selection. With that only criterion, we can output the codevector sequence with maximum output probabilities, but the synthesized result is not so perfect, only sound intelligent but not clear.

$$\mathbf{Cost_1 = 1 - output\_probability}$$

3.2.2 The predicted formant trajectory by MSD-HMM

Based on theories in speech coding, the formant position is one of the key parameters for the voice quality of synthesized speech. In current task, the formant trajectory generated by MSD-HMM is the ideal formant value that the synthesized speech should match. The difference between that ideal value and current candidate's real value is another cost function which makes the synthesized speech sound clear and transparent. Just as the MSD-HMM training procedure, only first and second formants are considered.

$$\mathbf{Cost_2 = formant\_difference}$$

3.2.3 The concatenation probability

The concatenation probability describes the probability of two neighbor candidates, which can be obtained by statistics analysis on large corpus. The effect of this cost function is similar with the dynamic features used in conventional HTS system.

$$\mathbf{Cost_3 = 1 - concatenation\_probability}$$

Based on these well-designed criterions, an appropriate codevector sequence can be selected and the LSP trajectory is reconstructed by retrieved codevectors. Fig 5 shows an example of selected LSP trajectory. Compared to the result generated by conventional HTS, it keeps more detailed information. In addition, the corresponding formant trajectory is clear and stable, which is very important to high voice quality. Therefore, the over-smoothing problem in time domain is better resolved.

Finally, with the gain trajectory generated by continuous HMM and the pitch contour generated by MSD-HMM, transparent speech is synthesized by STRAIGHT filter [13].

4. EXPERIMENT AND EVALUATION

4.1. Experimental conditions

In our experiment, the training data consists of 2000 phonetically balanced Mandarin sentences. Speech signal are sampled at 16000 Hz and they are windowed by a 25ms hanning window with a 5 ms shift. In the two-stage vector quantization procedure, the codebook size in our system is

set as 14 bit (7 bits are allocated to each stage), which means 2^{14} codevectors are included. With the setting, the voice quality of reconstructed speech will not degrade so much and the convergence time of HMMs is acceptable.

In the discrete HMM training part, the spectrum parameter consists of 24-order LSP coefficients obtained from the smoothed spectrum analyzed by STRAIGHT. The LPC-based spectrum gain is constructed using continuous HMM and the pitch contour and formant trajectory are constructed using MSD-HMM. In all cases, the 5-state left-to-right HMMs with single diagonal Gaussian output distributions are used. The contextual features for contextual HMM modeling and tree-based clustering are designed considering the Mandarin characteristics.

4.2 Perceptual evaluation

In the paper, a mean opinion score (MOS) test on the voice quality of synthetic speech is performed to demonstrate the effectiveness of the presented novel system. Speech synthesized by our method, the conventional method and the direct analysis-synthesized method are compared in terms of voice quality. In the test, 100 test sentences, which are not included in the training data, are synthesized by using these three methods. Then, 15 persons, all of them under-graduate university students, are asked to provide a rating for the speech quality form 1.0 to 5.0.

Fig 5 shows the MOS result for the three kinds of voices. Although the voice quality of our HTS system is still worse than the analysis-synthesized speech quality, it has been much improved compared with the conventional HTS system.

	Conventional HTS	Novel HTS	Analysis-synthesized
MOS	3.3	4.0	4.6

Table 2. The MOS result

5. CONCLUSION

The paper presents a novel HTS system which makes use of both continuous HMMs and discrete HMMs to resolve two over-smoothing problems in current HTS system. From the experiment result, the synthesized speech becomes clearer and the muffled phenomenon is alleviated.

However, there are still several disadvantages in this novel HTS system. First, the storage requirement is increased because of the storage of large number of codevectors. Second, because the replacement of continuous Gaussian distribution with discrete codevectors' output probability distribution, some self-adaptation algorithm [11] [12] may not be applied to the new system directly. Future work will focus on the resolution of these problems.

6. ACKNOWLEDGEMENTS

The authors would like to thank Professor Fuyuan Mo for his many constructive suggestions. The work was supported by the National Natural Science Foundation of China (No. 60575032) and the 863 Program (No. 2006AA01Z138)

7. 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, 1999, pp. 2347–2350.

[2] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in Proc. of ICASSP, 1995, pp. 660–663.

[3] Heiga Zen and Tomoki Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005", in Proc. of InterSpeech 2005, pp. 93-96

[4] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in Proc. of Eurospeech, 2005.

[5] YiJian Wu and Renhua Wang, "Minimun Generation error training for HMM-based speech synthesis", in Proc of ICASSP 2006, France

[6] Jian Yu, Wanzhi Zhang, Jianhua Tao, "A New Pitch Generation Model Based on Internal Dependence of Pitch Contour for Mandarin TTS System", ICASSP 2006, Toulouse, France

[7] Venkatesh Krishnan, David V. Anderson, Kwan K. Truong, "Optimal Multistage Vector Quantization of LPC Parameters Over Noisy Channels", IEEE Transaction on Speech and Audio Processing, VOL. 12, No.1 , Jan 2004

[8] David Sundermann, Harald Hoge, Antonio Bonafonte, Hermann Ney, Alan Black, Shri Narayanan, "Text-Independent Voice Conversion Based on Unit Selection", ICASSP 2006, France

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling." Proc. ICASSP, pp. 229–232, Phoenix, U.S.A., May 1999.

[10] Steve Young, "The HTK book", at website http://htk.eng.cam.ac.uk/

[11] Junichi Yamagishi, Katsumi Ogata, Yuji Nakano, Juri Isogai, Takao Kobayashi, "HSMM-Based Model Adaptation Algorithms for Average-Voice-Based Speech Synthesis", ICASSP 2006, France

[12] 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, May 2001, pp. 805-808

[13] H.Kawahara, I. Masuda-Katsuse and A. deCheveigne, "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