Phoneme Variation based Synthesized Speech discrimination for Speaker Verification

LianWu Chen[1, 2], Wu Guo[1], Yan Song[1], LiRong Dai[1]

[1]*Department of Electronics Engineering and Information Science, University of Science and Technology of China, Hefei, China*
[2]*ATG Sound Research, Dolby Laboratories, Beijing, China*

ABSTRACT

How to discriminate the synthesized speech from the natural speech for speaker verification is addressed in this paper. With the development of HMM-based speech synthesis, it is easy to obtain high quality synthesized speech which sounds like target speaker, the robustness of synthesized speech become important for speaker verification. In this paper, a method based on the phoneme variation is proposed to discriminate synthesized speech from natural speech, which could be used as front-end module to detect the synthesized speech for speaker verification system. The experimental results show the effectiveness of the proposed method.

Index Terms— *phoneme variation, synthesized speech discrimination, speaker verification*

1. INTRODUCTION

In the last decades, there has been a great advance in the field of text-independent speaker verification. The Gaussian Mixture Model-Universal Background Model (GMM-UBM) proposed by D. A. Reynolds [1] has become the state of the art algorithm. Moreover, the Gaussian Supervectors with Support Vector Machines (GSV-SVM) [2] and the Joint Factor Analysis (JFA) [3] have further improved the performance of speaker verification system. Today, the speaker verification technology has come into practical usage, and it is widely used in security departments and forensic systems. Most of the researches in speaker verification have been focusing on the natural speech, that is to say, the speech used in model training and testing are both spoken by humans. However, impostors using synthesized speech have barely been taken into consideration.

At the same time, the speech synthesis has made impressive progress in the last decade. With the development of HMM-based speech synthesis (HTS) [4], statistical parametric speech synthesis has been widely developed into a mainstream method for generating natural sounding synthesized speech with high flexibility. Furthermore, it is easy to build a speaker-dependent HTS model through model adaptation using only a few utterances of the target speaker [5]. Since the impostors can produce speech similar with the target speaker through the HTS technology, synthesized speech becomes a serious threat to the speaker verification system.

In the pilot study of this topic [6], the state of the art speaker verification system usually accepts the synthesized speech as the speech spoken by target speaker. In our previously research [7], we found that the Equal Error Rate (EER) of GMM-UBM system for natural speech is 0% in a corpus of 15 speakers, while the False Acceptance Rate (FAR) for synthesized speech is 99.2% using the same threshold of the natural speech. Similar results can also be found in [8] [9]. To improve the robustness of speaker verification against synthesized speech, Q. Jin used the phonetic feature to construct the speaker verification system [8], PL. De Leon proposed to detect synthesized speech based on Relative Phase Shift (RPS) features [9], and the higher order of Mel-cepstrum (MCEP) was used to discriminate synthesized speech from natural speech in our previously work [7].

Since the decision tree-based context clustering or similar process is an essential step for building the HMM-based parametric speech synthesis system, it is reasonable to assume that the synthesized speech has less variation than the natural speech. In this paper, a method based on the phoneme variation is proposed to discriminate synthesized speech from natural speech, where the variation is measured by the MCEP distance between different realizations of same phone. The experimental results show that the synthesized speech could be discriminated from the natural speech based on the proposed method.

The paper is organized as follows. In section 2, we describe the HTS and experiment setup briefly, and the weakness of speaker verification against synthesized speech is investigated in section 3. In section 4, we introduce the synthesized speech discrimination system based on phoneme variation. The experimental results are shown in section 5. Finally, the conclusions are given in section 6.

2. HTS SYSTEM AND EXPERIMENTAL SETUP

In this paper, the HTS system produced by iFlyTek Corporation [10] is used to generate the synthesized speech for experiments.

2.1. HTS system

In HTS model training stage, the 72 dimensional MCEP coefficients (24-dimensional static feature and 48-dimensional dynamic feature) are used as spectral feature in our system. The coefficients are extracted using STRAIGHT algorithm. Speech is segmented by a 25ms Blackman window with 5 ms frame shift. The context dependent tri-phone HMMs are used for this experiment, each HMM has a 5-state left-to-right structure without skip, and each stream-output distribution for spectral features is a single Gaussian distribution with diagonal covariance matrix. Stream-level parameter sharing structure is built using the decision tree-based context clustering based on the minimum description length (MDL) criterion.

In the speech parameter generating stage, traditional method allows the trajectory close to mean vector sequence of the HMM. Although this method reasonably reduces the generation error, it always makes the global variance much smaller than the natural speech. An improved method generates the trajectory with appropriate global variance [11]. In this paper, we try both of these two methods to investigate the influence of global variance in our experiments.

2.2. Database and experimental setup

The database used in experiments is the 863 Putonghua (Mandarin) corpus [12] designed with phonetic balance. 20 speakers (gender balance) from the corpus are selected for our experiments. Each speaker has 521 utterances of natural speech with duration form 1 second to 10 seconds, amount to 50 minutes per speaker. The phonetic balance and amounts of the duration make it be a good candidate for constructing a HMM-based parametric speech synthesis system.

We construct HMM-based parametric speech synthesis system for each speaker. For each natural speech, we generate two kinds of synthesized speech of the same context. One is the synthesized speech without considering global variance; the other is the synthesized speech with global variance. That is, for each speaker, we have three kinds of speech of same context: natural speech (Nat), synthesized speech without global variance (Syn_NoGV) and synthesized speech with global variance (Syn_GV).

3. WEAKNESS OF SPEAKER VERIFICATION AGAINST SYNTHESIZED SPEECH

The GMM-UBM based speaker verification system is used in this paper. In the system, the 39-dimemsional PLP parameter (13-dimensional static feature as well as its first and second derivatives) is adopted as the acoustic feature.

To evaluate the performance of the speaker verification system against synthesized speech, the nature speech model for each speaker is trained first. Then 20 utterances from each kind of speech are selected for testing. The duration of utterances selected here is about 10 seconds. For the speaker verification task, we treat the natural speech from the claimed speaker as target, and the speech from the non-target speaker and all the synthetic speech as impostor, even the synthetic speech from the target speaker.

In the experiments, the speaker verification system is first tested by natural speech, the equal error rate (EER) is 1.8% with the decision threshold properly set. Then the system is tested by the synthesized speech using the same threshold, the false accept rate (FAR) is 96.5% for synthesized speech without global variance and 93.4% for synthesized speech with global variance. These results indicate that the traditional PLP and GMM-UBM based speaker verification is working for natural speech, but still has problem for detecting the synthesized speech impostors.

4. PHONEME VARIATION BASED SYNTHESIZED SPEECH DISCRIMINATION SYSTEM

Human pronounces speech through vocal organs. As the variation such as emotion and breathing, it is impossible for human to speak the same context twice without any difference. However, the HMM-based parametric speech synthesis system produces speech through text analysis. If we synthesize the speech twice using the same text, we can generally get the same speech. Besides, when the same phone is synthesized twice, the same state model in the decision tree is likely to be selected because of the decision tree-based context clustering. That may cause small variation between different realizations of same phone. Therefore, we consider using phoneme variation for discriminating synthesized speech from natural speech. The variation is measured by acoustic feature distance between different realizations of same phone. For two realizations $\{\boldsymbol{X}, \boldsymbol{Y}\}$ of same phone in a test utterance, the dynamic time warping (DTW) algorithm can be applied to calculate the distance between them.

DTW is used to measure the similarity between two sequences which may vary in time or speed. The DTW distance between two frame series $\boldsymbol{X}= [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$ and $\boldsymbol{Y}= [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M]$ is $D(\boldsymbol{X},\boldsymbol{Y})$, M and N are the number of frames. In general, N is not equal to M because of timing variability in speech. The asymmetric score D is given by

$$D(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{M} d(\boldsymbol{y}_i,\boldsymbol{x}_{j(i)}) \quad (1)$$

Where $d(\ )$ is the distance metric between two acoustic features, and Euclidean distance is always adopted in DTW. The template indices $j(i)$ is decided during the dynamic warping process. DTW algorithm performs a constrained, piecewise linear warping of one (or both) time axis(es) to align the two sequences while minimizing the distance D.

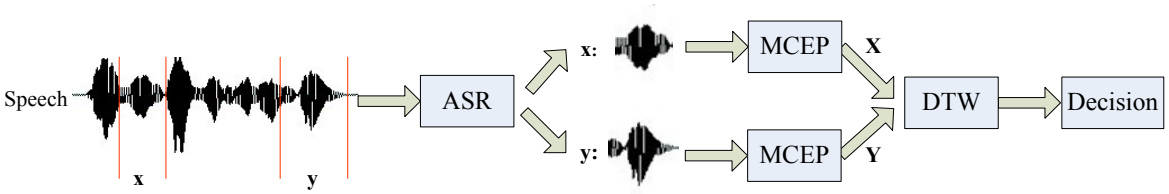

Figure 1: Synthesized speech discrimination system based on the phoneme variation

The flow chart of the synthesized speech discrimination system is shown in Figure 1. HTK tool's Hvite [13] is first used to generate phonetic transcripts of the test utterance. Then we get the occurrence information for various kinds of phone. For the phone which occur more than once in the test utterance, the MCEP feature is extracted and the relative DTW distance is calculated through formula (1). Finally, the distance D is used as the score to discriminate synthesized speech. Note that the system relies on the constraint that the test speech should contain different realizations of same phone. The phone mentioned above will be any kinds of phone with certain the context feature, such as mono-phone, bi-phone, tri-phone etc.

5. EXPERIMENTAL RESULTS

In this section, the performance of synthesized speech discrimination system based on phoneme variation is investigated. The related experimental results are presented as following.

5.1. Tri-phone

In the first experiment, we investigate the tri-phone variation for natural speech and synthesized speech. For example, to investigate the variation of the tri-phone "d-e+zh" for natural speech, we first select the utterances from natural speech which contain "d-e+zh", and then the MCEP distance is calculated between different realizations of "d-e+zh" for each speaker. The resulting distances of all speaker are seem as the variation of the tri-phone "d-e+zh" for natural speech. As the numerous kinds of tri-phone, it is impossible to investigate the variation for each kind of tri-phone. So all the tri-phone variation scores are combined together and a fixed threshold is setup to detect synthetic speech.

The MCEP distance distribution between different realizations of same tri-phone is illustrated in Figure 2. There are three distributions in the figure. The blue line shows the tri-phone variation of natural speech, while the red dash line shows the tri-phone variation of synthesized speech without global variance and the black dot line is for the synthesized speech with global variance. The distributions contain the phoneme variation of all the 20 speakers in our corpus. We could find that the phoneme variation of synthesized speech is smaller than that of natural speech due to the decision tree-based context clustering in

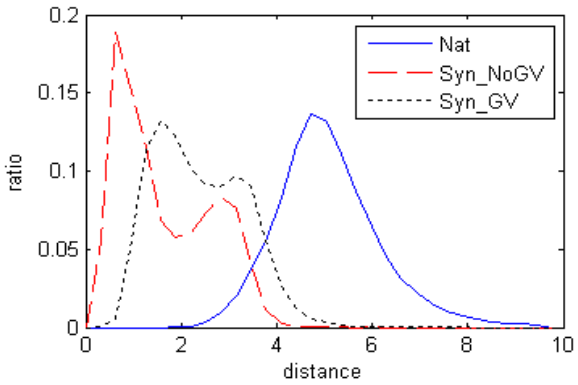

Figure 2: Variation of tri-phone

Table 1: EER of synthesized speech discrimination system based on tri-phone, voiced tri-phone and unvoiced tri-phone respectively.

System	EER for Nat against Syn_NoGV	EER for Nat against Syn_GV
tri-phone	4.09%	8.38%
voiced tri-phone	1.99%	6.82%
unvoiced tri-phone	1.25%	4.46%

the HMM-based speech synthesis. Besides, the global variance algorithm increases the phoneme variation of synthesized speech, which makes it more difficult to discriminate the synthesized speech from nature speech. The EER of the discrimination system is shown in Table 1. It could be found that the EER for natural speech against synthesized speech without global variance is 4.09%, while the EER for natural speech against synthesized speech with global variance is 8.38%.

5.2. Voiced tri-phone and unvoiced tri-phone

In Figure 2, the distance distribution for synthesized speech is bimodal, which is caused by the voiced and unvoiced difference as Figure 3 shows. For all kinds of speech, the unvoiced tri-phone has larger variation than the voiced tri-phone statistically in Figure 3, so it is intuitively to setup discrimination system for voiced and unvoiced tri-phone individually. The relative EER is shown in Table 1. We find that the performance of discrimination system based on voiced or unvoiced tri-phone variation respectively is better than the system based on tri-phone variation that contains both voiced and unvoiced tri-phone variation.

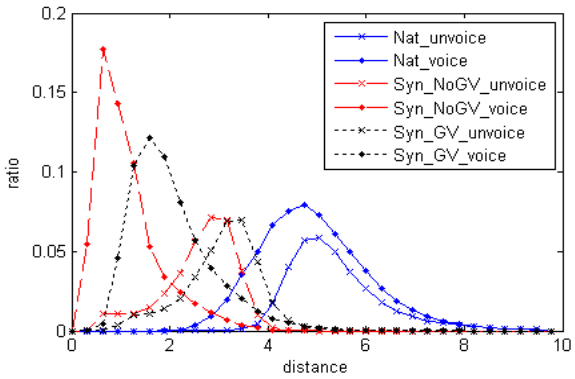

Figure 3: Variation of voiced tri-phone and unvoiced tri-phone

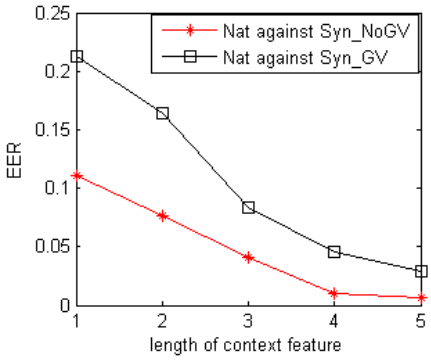

Figure 4: EER of synthesized speech discrimination system for different length of context feature

5.3. Length of context feature

To investigate the relation between the length of context feature and the performance of the synthesized speech discrimination system, five kinds of context features is investigated here: Mono-phone, Bi-phone, Tri-phone, Quad-phone, Quin-phone (the relative length of context feature is from 1 to 5). The performance of the discrimination system for different kinds of context features is showed in Figure 4. It could be found that the EER of discrimination system will decrease as the length of context feature increasing. This is reasonable since the state unit for parameter synthesis is selected according to the questions of decision tree. It is more possible to select the same state model for the phone as the restrain become stronger, which leads to less variation between different realizations of the phone.

6. CONCLUSIONS

In this paper, a phoneme variation based synthesized speech discrimination system is proposed, where the variation is measured by the MCEP distance between different realizations of same phone. Experimental results show that the natural speech has larger variation than synthesized speech, which demonstrate the effectiveness of the method.

Since the performance of proposed method is highly depend on the content information of test utterance (type of the phone that occur more than once in the test utterance), the performance of speaker verification system which combine the proposed method is not given in this paper. Despite of this, it is foreseeable that the robustness of speaker verification system against the synthesized speech will be improved by combining the proposed method if the text utterance is properly designed, since the synthesized speech discrimination system and the speaker verification system are almost decoupled.

7. ACKNOWLEDGEMENTS

This work is partly supported by Nature Science Foundation of China (Grant No.60970161), Chinese Universities Scientific Fund (Grant No.Wk2100060008), and Fundamental Research Funds for Central Universities.

8. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, 2000, vol. 10, pp.19-41.

[2] W. M. Campbell, D. E. Sturim, D. A. Reynolds, “Support Vector Machines using GMM Supervectors for Speaker Veriffication”, IEEE Signal Processing Letters, vol. 13,no. 5, pp. 308–311, 2006.

[3] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” IEEE Transations on Audio, Speech and Language Processing, vol. 16, no. 5, pp. 980-988, July 2008.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech Parameter Generation Algorithms for HMM-based Speech Synthesis,” Proc. ICASSP 2000, pp.1315-1318.

[5] J. Yamagishi, T. Kobayashi, N. Yuji,*et.al*, “Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm,” IEEE Transactions on Audio, Speech and Language Processing, vol. 17(1), pp.68-83, January 2009.

[6] T. Masuko, K. Tokuda, T. Kobayashi, “Imposture Using Synthesized Speech Against Speaker Verification Based on Spectrum and Pitch, ” Proc. ICSLP 2000, vol.2, pp. 302-305.

[7] L. W. Chen, W. Guo, L. R. Dai, “Speaker Verification against Synthesized Speech,” Proc. ISCSLP 2010, pp. 309-312.

[8] Q. Jin, A. R. Toth, A. W. Black and T. Schultz, “Is Voice Transformation a Threat to Speaker Identification?” Proc. ICASSP 2008, pp. 4845-4848.

[9] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, “Detection of synthesized speech for the problem of imposture,” Proc. ICASSP 2011, pp. 4844–4847.

[10] 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 ICSLP Satellite Workshop, Blizzard Challenge, 2006.

[11] 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.

[12] Y. Q. Zu, “Sentences design for speech synthesis and speech recognition database by phonetic rules,” Proc. Eurospeech, 1997, pp. 743-746.

[13] S. Young, D. Ollason, V. Valtchev, P. Woodland, “The HTK Book,” Cambridge University, 2006.