IMPROVING HMM BASED SPEECH SYNTHESIS BY REDUCING OVER-SMOOTHING PROBLEMS

Meng Zhang(1) Jianhua Tao(2) Huibin Jia(3) Xia Wang(4)

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

ABSTRACT

Although Hidden Markov Model based speech synthesis has been proved to have good performance, there are still some factors which degrade the quality of synthesized speech: vocoder, model accuracy and over-smoothing. This paper analyzes these factors separately. Modifications for removing different factors are proposed. Experimental results show that over-smoothing in frequency domain mainly affect the quality of synthesized speech whereas over-smoothing in time domain can nearly be ignored. Time domain over-smoothing is generally caused by model structure accuracy problem and frequency domain over-smoothing is caused by training algorithm accuracy problem. Currently used model structure is capable of representing speech without quality degradation. ML-estimation based parameter training algorithm causes distortion of perception in speech synthesis. Modification for improving parameter training algorithm is more likely to improve the synthesizing performance.

Index Terms—Hidden Markov Model, speech synthesis

1. INTRODUCTION

In the last few years, a kind of statistical parametric speech synthesis method has been proposed and given widely attention [1]. In this method, hidden Markov Model (HMM) is used for modeling spectrum, pitch and duration of speech simultaneously in a unified framework [2]. Smoothed parameter sequences are generated from HMMs with considering dynamic features for synthetic speech [3].

HMM based speech synthesis (HTS) has been proved to have good performance [4][5][6], and several advantages. For instance, it can generate flexible characteristics of synthetic voice which can be easily modified. HTS systems are almost language-independent with small footprint comparing to unit selection synthesis systems. However, they still have some drawbacks with buzzy and muffled synthesized speech sounds. In [7], A. Black has made a detailed analysis on the advantages and disadvantages of HTS and pointed out that the quality of synthetic speech is degraded by three main factors: vocoder, modeling accuracy, and over-smoothing.

Many research efforts have been carried out to upgrade the quality of HTS. To improve model structure accuracy, some methods such as hidden semi-Markov models (HSMMS) [8], and trajectory HMMs [9] have been investigated. New training frameworks have also been derived, e.g. minimum generation error (MGE) criterion [10]. Other works focus on the over-smoothing problem widely considered caused by parameter generation algorithm which can generate smooth speech parameter trajectories, such as conditional speech parameter generation algorithm [11] and speech parameter generation algorithm considering global variance [12].

Although there have been a lot of works proposed to improve the quality of synthetic speech of HTS in many ways, the correlation between these factors and the performance of HTS is still not very clear. We still don’t know how these factors affect the quality of synthesized speech.

Quality degradation after voicing process is a problem that every parameter generation based speech synthesis methods need to overcome. The alleviation of this problem depends on the progress of speech coding technique. This factor will not be considered in this paper. As for the modeling accuracy and the over-smoothing, the first one is more like a reason and the second one is more like a result. Poor modeling accuracy may cause over-smoothed model parameters, and further result in quality degradation of synthesized speech. In our work, we analyze modeling accuracy by separating it into two parts: model structure accuracy and training algorithm accuracy. And the over-smoothing is also classified into two types: the over-smoothing in time domain and the over-smoothing in frequency domain [13].

The paper uses two methods to reduce the influence from the model structure accuracy and training algorithm accuracy respectively. By comparing the results, we try to investigate which factors are important to influence the performance of HTS. With the experiments, we finally find that over-smoothing in frequency domain is the main factor...
which influences the quality of synthesized speech and it is generally caused by training algorithm accuracy problem.

The rest part of this paper is organized as follows: Section 2 analyzes factors which degrade the quality of synthesized speech. Then two methods for reducing the influences from model structure accuracy and training algorithm accuracy are proposed. In section 3, we make some experiments for two methods. By comparing the synthetic results of different type of modifications, the influence of each factor on quality degradation is studied. Finally, the conclusion of this paper is presented in Section 4.

2. RUDUCING OVER-SMOOTHING PROBLEMS

2.1. Over-smoothing problems and modeling accuracy

In our former work, we have classified the over-smoothing problems of HMM into two types according to their presentation: the over-smoothing in time domain and the over-smoothing in frequency domain [13], as shown in Figure 1 and Figure 2. By comparing the LSF trajectories of synthesized speech and the real speech, the frequency domain over-smoothing makes distortion of the formant structure within each frame of synthetic speech. The over-smoothing in frequency domain is mainly caused by inappropriate parameter training algorithm which leads to spectrum distortion. The most frequently used training algorithm for HMM is ML-estimation which is only able to generate parameters with mean and covariance of model states.

The time domain over-smoothing may make the synthetic speech lost detailed spectrum presentation for phoneme structures. In HMM based speech synthesis, model structure of hidden Markov model is usually three or five effective states, left to right with no skip [5][6]. The over-smoothing in time domain is mainly caused by such limited model structure and parameter generation algorithm taken account of constraints between statistic and dynamic features, which make the trajectory of spectral parameter less expressive in time domain.

2.2. Modification of HMMs

To analysis which modeling accuracy gives more influence on the over-smoothing problems and the general quality of synthesized speech, we use two methods by reducing the influence of model structure accuracy and training accuracy respectively. By comparing the results of their synthesized results, we will know the relationship between the speech quality and the different factors.

Fig.1 the over-smoothing problem in frequency domain

(a) real spectrum (b) synthesized spectrum

Fig.2 the over-smoothing problem in time domain

(a) real LSP trajectory (b) synthesized LSP trajectory

2.2.1. Combine continuous HMMs with discrete HMMs

In this part of work, we are trying to reduce the training accuracy problems by combining the continuous HMMs with discrete HMMs [13]. In this method, we replace the Gaussian function in continuous HMM with the output probability of codevector in discrete HMMs, which resolves the over-smoothing problem in frequency domain. Then, the continuous spectrum parameters are represented as discrete codevector indexes by vector quantization (VQ). 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. More detailed description of this work can be found in [13].

In this paper, we are trying to find the influence of training algorithm, the spectrum distortion from the codevector has to be avoided. Then, we construct a very large vodewords which contains all speech frames of full training corpus. We replace the mean of each state generated by continuous HMM by nearest spectrum from real speech. The method is illustrated in Figure 3. We believe there are minimum training accuracy problem in this synthetic speech which is marked as MS_speech for the later experiments.

2.2.2. Increase the amount of HMM states

If we can construct a model which is able to contain detailed spectrum distribution for each frame, there will be
no problem for the over-smoothing in time domain. But unfortunately, we only can use some limited HMM states to model a phoneme. Although the unit selection method can solve the time domain over-smoothing problem, we have to use a very large speech corpus. The only method we can do is that we can increase the amount states for HMM to reduce the influence caused by the model structure accuracy. The typical state amount of HMM for Chinese phonemes is three or five. The first one is usually used for single vowels while the second is mainly used for compound vowels. This kind of model structure works very well for most speech recognition systems, however the accuracy is not enough to simulate the detailed spectrum distribution of the phoneme in time domain. To get the influence of the model structure accuracy, we increase the state amount to five (for single vowels) and seven (for compound vowels). In the later experiments, we are trying find if there is any difference of the synthesized results between the traditional model (with three HMM sates) and this revised model (with five and seven states). We mark the results from traditional HMM as HTS_speech, and the results from the revised model as TA_speech.

3. EXPERIMENTS FOR STUDYING INFLUENCE ON QUALITY DEGRADATION

3.1. Experiment setup

For analyzing relationship between quality of synthetic speech and each factor, four sets of speech are generated: RL_speech (real speech), MS_speech, TA_speech and HTS_speech (HTS generated speech). Each set have 42 test sentences. Sentences in each set have parallel context and prosody information but different speech quality. 10 subjects participated in the test. They were presented a pair of two set of parallel speech in random order and asked which one has better quality. In order to avoid vocoder factor in subjective test, all real speech (RL_speech) are also represented through vocoding process without changing any spectral feature.

3.2. Experimental results

Preference scores are shown in Figure 4. From the results we can see that, four sets of speech are generally divided into three groups: (1) RL_speech; (2) MS_speech; (3) TA_speech, HTS_speech. Group (1) has much higher preference score than group (2) and (3). In group (3), TA_speech is nearly the same with HTS_speech.

Figure 5 shows the LSF trajectory of MS_speech and TA_speech. Simultaneously considering Figure 4 and Figure 1, what interesting is by comparing to RL_speech and HTS_speech, MS_speech is only a little over-smoothing in frequency domain but more in time domain; whereas TA_speech is only a little over-smoothing in time domain but much in frequency domain.

3.3. Discussion

Figure 5 shows that MS speech only has a little over-smoothing in frequency domain but much in time domain comparing to HTS_speech and RL_speech. This implies that frequency domain over-smoothing is mainly caused by training algorithm accuracy problem since MS_speech alleviated this factor. Similarly, according to comparison between TA_speech and HTS_speech, time domain over-smoothing is mainly caused by model structure accuracy problem. This conclusion agrees with the intuition.
Fig. 6 Illustration of the influence affecting quality of synthesized speech

Further more, the results of preference score shows that although model structure accuracy problem causes time domain over-smoothing, it nearly doesn’t affect the quality of synthesized speech. The main reason for quality degradation is over-smoothing in frequency domain which is induced by training algorithm accuracy problem. We can present the final conclusion by Figure 6.

Model parameter training algorithm for HMM used in HTS is based on ML-estimation, which is applied in speech recognition originally. For speech recognition, parameters of HMM need to be “medial” enough to represent data distribution of unit such as phoneme. It is unnecessary to consider perception or quality of model parameters. But for speech synthesis, too “medial” may not be good for quality of synthesized speech but induce distortion in perception. Parameters for synthesis sometimes need to be “individual” for generating expressive speech. This may be the reason that ML-estimation based training algorithm causes frequency domain over-smoothing and degrade the quality of synthesized speech. The improvement made by Minimum generation error (MGE) criterion training for HTS [10] which modifies the training algorithm is another proof of this conclusion. On the other hand, this conclusion is also an explanation for the success of MGE.

The high preference score of MS_speech comparing to TA_speech demonstrates that the model structure generally used in HTS is capable of representing speech without quality degradation. Currently, the parameter generation algorithm ignores the dependency between each dimension of spectral parameter so it may introduce spectral distortion and affect the speech quality.

4. CONCLUSION

This paper analyzes the factors which affect the quality of synthesized speech of HTS: vocoder, modeling accuracy and over-smoothing. Modeling accuracy is separated into two parts: model structure accuracy and training algorithm accuracy which respectively caused time domain over-smoothing and frequency domain over-smoothing. Modifications for reducing different factors are proposed. Experimental results shows that over-smoothing in frequency domain mainly affects the quality of synthesized speech whereas over-smoothing in time domain can nearly be ignored. Currently used model structure is capable of representing speech without quality degradation. ML-estimation based training algorithm causes distortion in perception for synthesizing speech. Modification for improving parameter training algorithm is more likely to improve the synthesizing performance. These conclusions can guide further modifications and improvements for HTS.

5. ACKNOWLEDGEMENTS

The work was supported by the National Natural Science Foundation of China (No. 60575032) and 863 Programs (No. 2006AA01Z138, No. 2006AA01Z194)

6. REFERENCES