INTERSPEECH 2009 BRIGHTON

Voiced/Unvoiced Decision Algorithm for HMM-based Speech Synthesis

Shiyin Kang[1], *Zhiwei Shuang*[2], *Quansheng Duan*[1], *Yong Qin* [2], *Lianhong Cai*[1]

[1] Key Laboratory of Pervasive Computing, Ministry of Education
[1] Tsinghua National Laboratory for Information Science and Technology
[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2] IBM China Research Lab, Beijing, China

{kangshiyin,ddandre32}@gmail.com, {shuangzw,qinyong}@cn.ibm.com,
clh-dcs@mail.tsinghua.edu.cn

Abstract

This paper introduces a novel method to improve the U/V decision method in HMM-based speech synthesis. In the conventional method, the U/V decision of each state is independently made, and a state in the middle of a vowel may be decided as unvoiced. In this paper, we propose to utilize the constraints of natural speech to improve the U/V decision inside a unit, such as syllable or phone. We use a GMM-based U/V change time model to select the best U/V change time in one unit, and refine the U/V decision of all states in that unit based on the selected change time. The result of a perceptual evaluation demonstrates that the proposed method can significantly improve the naturalness of the synthetic speech.

Index Terms: speech synthesis, unvoiced/voiced determine, HMM, GMM

1. Introduction

HMM-based speech synthesis technology has developed rapidly in recent years [1,2]. Compared to the traditional large-corpus-based unit selection synthesis method, HMM-based speech synthesis has a lower resource cost, and less relevance to speakers and languages. In addition, the characteristics of the synthetic speech, such as pitch and prosody, are easy to adjust due to the use of vocoder-based synthesizer.

Conventional HMM-based speech synthesis can be divided into 2 stages. In the training stage, spectrum and F0 parameter sequence are extracted from the corpus. Then the Hidden Markov Models (HMM) of the speech units are built from the parameter sequences and their dynamic characteristics. Finally, a decision tree is built to cluster the HMMs [3]. In the synthesis stage, context information is generated by the text analysis procedure. HMM sequence is predicted by the decision tree based on the context information. Then the speech parameter sequence is generated based on the predicted models. Finally, the speech waveform is synthesized from the speech parameter sequence by a vocoder.

Currently, the Gaussian parameters of neighboring states are estimated independently. However, an abrupt change of speech parameters in neighboring states may make the synthesized speech sounds unnatural. To make the synthesized speech more natural, dynamic features are introduced to improve the continuity of speech parameters [4].

However, in current HMM based TTS solution [5,6], the Unvoiced/Voiced (U/V) decision of each state is still independently made based on the multi-space distribution (MSD) [8] of F0 parameters of that state. Such independent U/V decision may cause unnatural transition between voiced segments and unvoiced segments in the synthetic speech, which can greatly hurt the quality of synthetic speech.

In this paper, we propose to utilize the constraints of natural speech to improve the U/V decision. As shown in Figure 1, a Gaussian Mixture Model (GMM) is introduced to model the U/V boundary of the speech unit in the training stage, and to help predicting the U/V decision of HMM states in the synthesis stages.

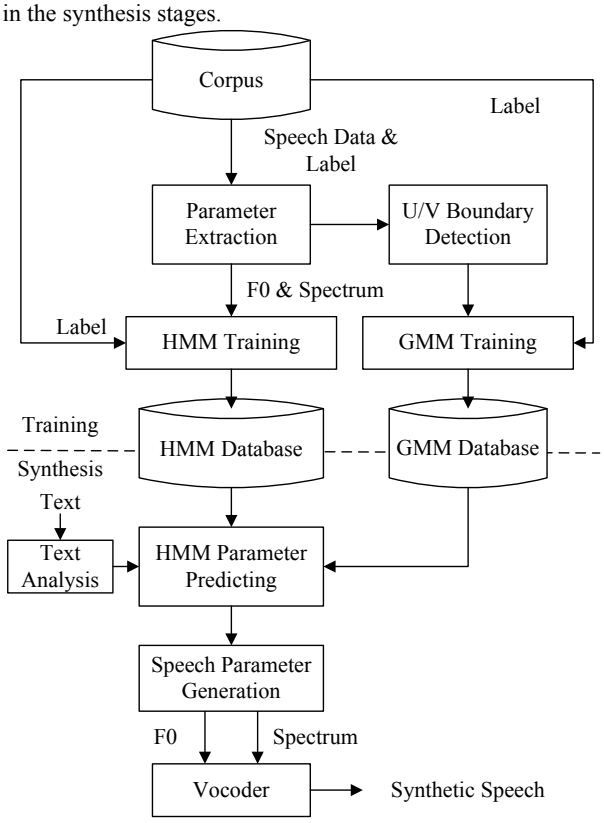

Figure 1: *HMM-based TTS using U/V decision algorithm*

The rest of this paper is organized as follows. The conventional U/V decision method and its problems are described in Section 2. The proposed UV decision method is described in Section 3. Results of the statistics of synthesis error and the MOS test are shown in Section 4. Concluding remarks are presented in the final section.

2. Conventional Method

In the current HMM TTS system, the Unvoiced/Voiced (U/V) decision of each state is independently made based on the multi-space distribution (MSD) of F0 parameters of that state.

The MSD of F0 parameters of one state is estimated by traversing the decision tree by the contextual features till a leaf node. Because of some pitch detection error or some bad pronounced vowel; one leaf of the state belong to a vowel may even contain more unvoiced occurrences than voiced occurrences. Then, if choosing that leaf, that state will be decided as unvoiced. If that unvoiced state happened to occur in the middle of a vowel with voiced neighboring states, it will sound very unnatural.

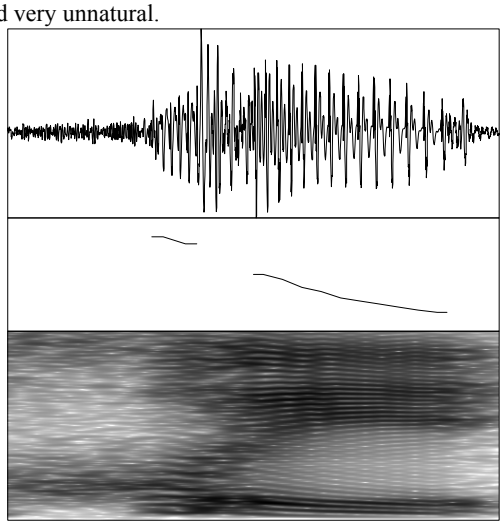

Figure 2: *Mandarin syllable "Hui4" with a bad "ui".*

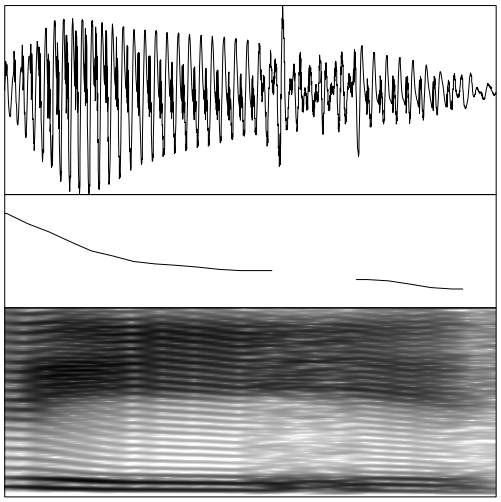

Figure 3: *Mandarin syllable "Li3" with a bad "i".*

Figure 2 is a synthesized sample of Mandarin syllable "Hui4", which should be constructed by an unvoiced consonant "h" and a voiced vowel "ui". In natural Mandarin syllable, there should be no unvoiced part in the vowel. However, because of the voiced/unvoiced decision error, an unvoiced state occurs inside the vowel. Thus, the vowel sounds very dry and hoarse, which greatly hurt the overall quality of synthesized speech.

Such problems can occur more often in the Mandarin syllables of the third tone. Because the pitch of these syllables can be very low, pitch detection algorithm may often fail. An example is shown in Figure 3. An unexpected unvoiced part appears in the mandarin syllable "Li3", which should contain a voiced consonant "l" and a voiced vowel "i".

We also meet similar problem in the English HMM-based TTS system. Figure 4 shows an example of a synthesized English word "considering". The last part of this word "-ring" should be all voiced. However, most part of the voiced "-ring" in the word is synthesized as unvoiced, which make the word sounds hoarse.

Figure 5 shows another example. An unexpected unvoiced part appears in the first part of the vowel "o", which ruins the whole phrase.

To solve this problem, we utilize the constraints of natural speech to improve the Voiced/Unvoiced decision.

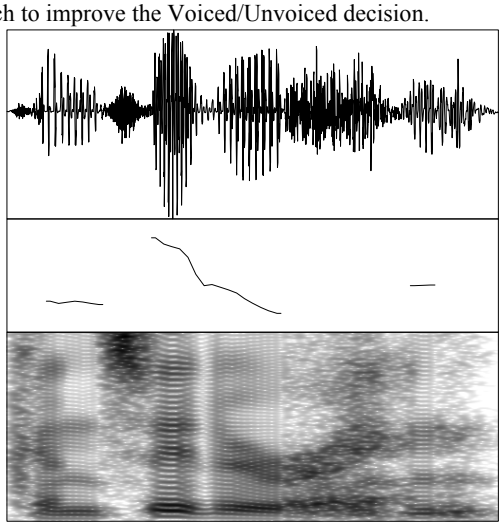

Figure 4: *English word "considering" with bad a "ring".*

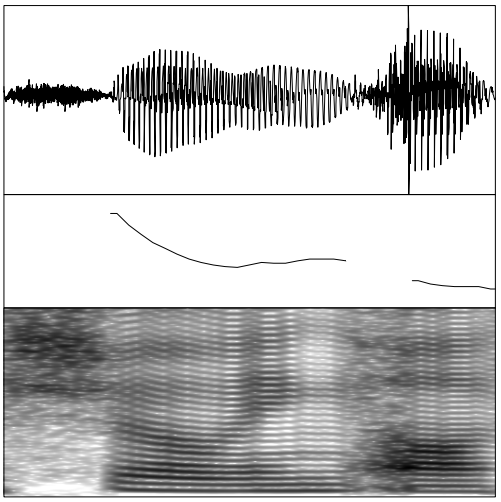

Figure 5: *English phrase "showing of" with a bad "of".*

3. Proposed U/V Decision Method

3.1. Unvoiced/Voiced Modeling

Because of the Consonant-vowel structure of Mandarin syllable, there should be no more than one Voiced/Unvoiced changing point in the whole syllable. We utilize this feature to avoid inappropriate unvoiced decision in the vowel part. This approach can be applied to other basic units if there should be no more than one Voiced/Unvoiced changing point in the whole unit i.e. phones.

First, we calculate the unvoiced percentage in the whole syllable. Then for each tonal syllable, we model the distribution of unvoiced percentage by GMM models. Then when synthesizing one syllable, we will check all the V/UV

changing point introduced by current HMM U/V decision, and select the changing point with the highest probability according to the GMM models of that syllable. The states before the selected changing point will be set as unvoiced, while the states after the selected changing points will be set as voiced. As to our experiments, this U/V decision syllable level refinement can greatly improve the quality of synthetic speech. We will introduce the details in the below.

Because of potential pitch detection errors and syllable alignment errors, we model the unvoiced percentage of a syllable instead of the position of the first unvoiced to voiced changing position.

The unvoiced percentage of a syllable is calculated using F0 parameter as

$$x = \frac{s_{unvoiced}}{n_{end} - n_{start} + 1} \tag{1}$$

where n_{start}, n_{end} are the indexes of first and last frame of the syllable respectively, $s_{unvoiced}$ is the total number of the unvoiced frames in the syllable.

The GMM-based unvoiced percentage model for every different pinyin is built using EM algorithm. The probability density function (PDF) is written as

$$p(x \mid \lambda) = \sum_{k=1}^{K} \omega_k \mathrm{N}(x \mid \mu_k, \Sigma_k) \tag{2}$$

where ω_k is the weight of the k-th Gaussian component; $\mathrm{N}(x \mid \mu_k, \Sigma_k)$ is the PDF of Gaussian.

$$\mathrm{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi \mid \Sigma_k \mid}} e^{-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)} \tag{3}$$

In the HMM based TTS system , the duration of each state is predicted using the context information. For an n-state syllable HMM, we use t_0 to denote the starting time of the syllable; and use $t_1, t_2, \ldots, t_n$ to denote the ending time of each state. Then, $t_n - t_0$ is predicted syllable duration.

Using the conventional algorithm, m candidates of the U/V changing time can be obtained as

$$t_{i_1}, t_{i_2}, \ldots, t_{i_m}, (0 \le i_1 < i_2 < \cdots < i_m \le n) \tag{4}$$

The indexes $i_1, i_2, \ldots, i_m$ satisfy the following condition.

$$\begin{cases} p_{i_j}(voiced) < threshold_{voiced} \\ p_{i_j+1}(voiced) \ge threshold_{voiced} \end{cases}, (1 \le j \le m) \tag{5}$$

where $p_i(voiced)$ is the voiced probability of state i. The voiced probabilities of state 0 and state n+1 are defined by

$$\begin{aligned} p_0(voiced) &= 0 \\ p_{n+1}(voiced) &= 1 \end{aligned} \tag{6}$$

The problem for U/V decision is to obtain the optimal U/V changing time $t_{i_{J^*}}$ of the syllable, which maximizes $p(x \mid \lambda)$ with respect to j,

$$J^* = \arg\max_{1 \le j \le m} p\left(\frac{t_{i_j} - t_0}{t_n - t_0} \mid \lambda\right) \tag{7}$$

According to the optimal U/V changing time $t_{i_{J^*}}$, we will set the interval $[t_0, t_{i_{J^*}}]$ as unvoiced part, and set the interval $[t_{i_{J^*}}, t_n]$ as the voiced part.

3.2. U/V Decision Comparison

Figure 6 shows an example in which the F0 predicting is refined by the GMM-based U/V decision algorithm. In the conventional method, F0 is predicted respectively in each HMM state (which is separated by the thin dashed line). The 6th and 10th state are predicted as unvoiced parts, and the F0 curve is cut into 2 parts. The synthetic voice sounds hoarse and not clear.

As described above, the proposed method first builds a GMM-based change point model for the syllable "hui" In the training stage. As shown in the second part of Figure 6, the probability density is represented by the thick solid line. It contains 4 Gaussian components, which are represented by the thin dotted line. The histogram of the training data is also shown on the background as a reference. In the synthesis stage, 2 change point candidates (which is represented by the thick dotted line) are compared according to the change point model. The change at the end of state 4 gets a higher probability. Then, all the states after state 4 are set to voiced for synthesis. As shown in the 3rd part of Figure 6, the F0 curve is continuous in the vowel part of syllable and the synthesized speech sounds much better.

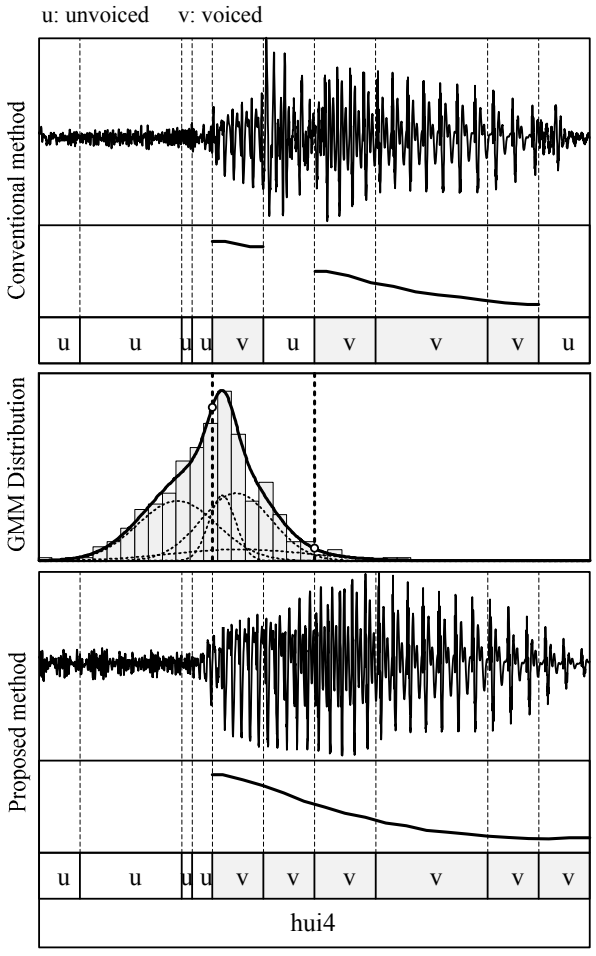

Figure 6: *GMM-based U/V decision for F0 predicting.*

4. Experimental evaluation

4.1. Experimental Conditions

A manually checked female mandarin speaker's corpus is used for both methods. The corpus has 1000 phonetically balanced sentences, which contains 19712 syllables. The total speech length is 1.45 hours. The speech signals were sampled at a rate of 16 kHz and quantified to 16-bit. The frame length was set to 25ms, and the frame shift was set to 5ms.

The speech parameter contained 0th through 24th MGC [8] coefficients, log-scaled F0, including their delta and delta-delta features. The speech parameter was modeled by 10-state left-to-right HMM. The context-dependent HMM was constructed using a decision-tree based context clustering technique based on MDL [9] criterion.

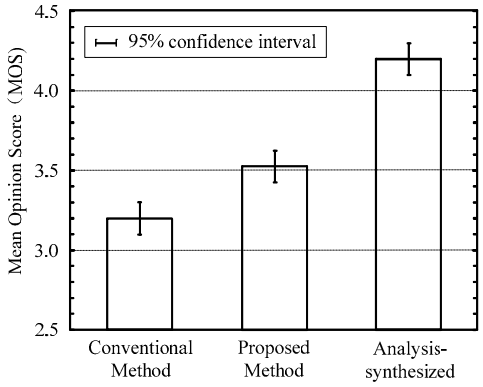

Figure 7: *MOS for synthetic voice.*

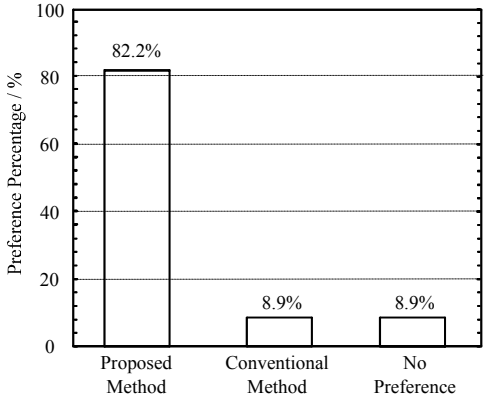

Figure 8: *MOS for synthetic voice.*

4.2. MOS Test

A Mean Opinion Score (MOS) Test was performed on the naturalness of synthetic speech to evaluate the effectiveness of the proposed method. The following 3 voices were evaluated: 1) synthetic speech using the conventional method; 2) synthetic speech using the proposed method; 3) analysis-synthesized speech, which means speech parameters extracted from natural speech were directly used for synthesis without the HMM training procedure.

Nine experienced native listeners participated in the test. Each listener evaluated 10 sentences for each method. The sentences were randomly selected from 30 testing data.

Figure 7 shows the result of the test. The proposed method performs better compared to the conventional one. The score rise from 3.2±0.1 to 3.5±0.1. However the propose method is still worse the analysis-synthesized one, which shows that further improvement is needed for HMM-based TTS.

4.3. Preference Test

Along with the MOS test, a preference test also performed to compare the naturalness between the proposed method and conventional one.

Figure 8 shows the preference scores. The proposed method provided a higher performance. The proposed method got 82.2%, while the conventional method got 8.9%. The other 8.9% was no preference between these 2methods.

5. Conclusions

This paper proposes an improved unvoiced/voiced (U/V) decision algorithm for HMM-based speech synthesis. The original U/V decision of each state is independently made, and a state in the middle of a voiced vowel may be decided as unvoiced. In this paper, we propose to utilize the constraints of natural speech to improve the U/V decision. A GMM-based U/V change time model is used to select the best U/V change time in the unit, and the U/V decisions of all states inside the unit will be refined according to selected U/V change time. Evaluation result shows that the proposed method can significantly improve the naturalness of the synthetic speech.

6. Acknowledgements

This work was partly supported by the SUR project of IBM China Research Lab. This work was also supported by the National Natural Science Foundation of China (60805008, 90820304), the National Basic Research Program of China (973 Program) (No.2006CB303101) and the National High Technology Research and Development Program ("863" Program) of China (No. 2007AA01Z198).

7. References

[1] Tokuda, K., Zen, H. and Black, A. W., "An HMM-based speech synthesis system applied to English", Proc. of IEEE Workshop on Speech Synthesis, 2002.

[2] Ling, Z., Wu, Y. and Wang, Y., "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method ", Proc. of ICSLP Satellite Workshop Blizzard Challenge, 2006.

[3] Yoshimura, T., Tokuda, K. and Masuko, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", Proc. of Eurospeech, vol. 5 2347-2350, 1999.

[4] Masuko, T., Tokuda, K. and Kobayashi, T., "Speech synthesis from HMMs using dynamic features", Proc. of ICASSP, 389-392, 1996.

[5] Dines, J. and Sridharan, S., "Trainable speech synthesis with trended hidden Markov models", Proc. of ICASSP, 833-837, 2001.

[6] Toda, T. and Tokuda, K., "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis", Proc. of Eurospeech, 2801-2804, 2001.

[7] Tokuda, K., Mausko, T. and Miyazaki, N., "Multi-space probability distribution HMM ", IEICE Trans. Inf. & Syst., vol.E85-D, no.3: 455-464, 2002.

[8] Fukada, T., Tokuda, K., Kobayashi, T. and Imai, I., "An adaptive algorithm for mel-cepstral analysis of speech", in Proc. ICASSP, 137-140, 1992.

[9] Shinoda, K. and Watanabe, T., "MDL-based context-dependent subword modeling for speech recognition", J. Acoust. Soc. Jpn., 21(2), 79-86, 2000.