운율 경계 정보를 이용한 HMM 기반의 한국어 음성합성 시스템

주영선, 정치상, 강홍구
연세대학교 전기전자공학과 디지털신호처리 연구실
disfruta@dsp.yonsei.ac.kr, jtoctos@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

An HMM-based Korean TTS synthesis system using phrase information

Young-Seon Joo, Chi-Sang Jung, Hong-Goo Kang
Yonsei University

Abstract

In this paper, phrase boundaries in sentence are predicted and a phrase break information is applied to an HMM-based Korean Text-to-Speech synthesis system. Synthesis with phrase break information increases a naturalness of the synthetic speech and an understanding of sentences. To predict these phrase boundaries, context-dependent information like forward/backward POS(Part-of-Speech) of eojeol, a position of eojeol in a sentence, length of eojeol, and presence or absence of punctuation marks are used. The experimental results show that the naturalness of synthetic speech with phrase break information increases.

1. Introduction

In a Text-to-Speech (TTS) system [1], naturalness and intelligibility of a synthesized speech are very important. The study to increase these has been kept and the one of primary synthesis method is a concatenation of acoustic units. TD-PSOLA [2] analyzes pitch-synchronously and synthesizes units in time domain. Although a modification of the prosody is easy, discontinuity problems appear at concatenation points which make synthetic sounds uncomfortable, and low quality of speech problems appear when the speech database size is small. To overcome these problems a statistical parametric speech synthesis system based on an hidden Markov model(HMM) has been used in a TTS system [3,4]. The system which to an HMM-based synthesis applied trains context-dependent HMMs of the speech database and generates speech from context-dependent HMMs of speech units. It synthesizes various models which are different depending on contexts without large database and solves mismatch problems in prosody and phase which appear in the concatenative synthesis methods.

As a quality of synthesized speech increases applicability of a TTS system also increases, like a navigation, an automated information service, or electronic dictionaries. Besides, some broadcasts using a TTS system. Since the only requirement to synthesize speech is a text, it's convenient for use. Broadcasting becomes possible to do without narrators or announcers and to select voice as female or child according to one’s preference. The important thing is a naturalness of synthetic speech. For naturalness not only synthesis techniques but also a phrase breaking are necessary. While speech breaks at every whitespaces without phrase boundaries prediction, prediction of phrase boundaries makes synthetic sounds natural and makes it easy to understand of sentence. A prediction algorithm to make sounds natural has been researched, and there are various methods, like an HMM-like POS sequence model [5], decision tree [6], CRF(Conditional Random Field) [7-8], and a conditional ME model [9].

This paper predicts the phrase break boundaries adopting a rule-based method with context-dependent factors, like punctuation marks, lengths of eojeol, a position of eojeol in phrase/sentence, and POS tags. A proposed method takes little time comparing to an HMM-like POS sequence model which uses a Viterbi beam search algorithm to find the best break level sequence. In the broadcasting and mobile environmental aspect, taking little times becomes a great advantage, that kinds of environment need speediness.

The rest of this paper is organized as follows. Section 2 summarizes an HMM-based TTS synthesis system and section 3 describes a prediction of phrase break boundaries and factors which are used in prediction. Section 4 describes an evaluation method and shows the experimental results. Experiments are performed with the three types of synthesized speech. Section 5 concludes this paper with suggestions for future works.

2. An HMM–based speech synthesis system

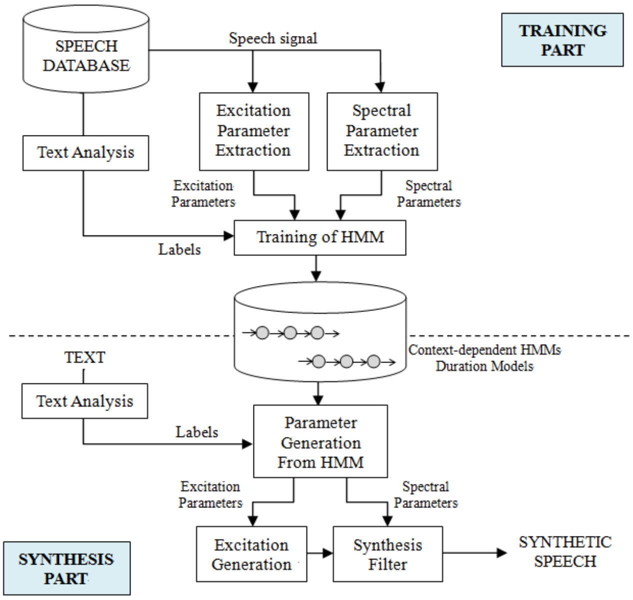

Figure 1. Overview of an HMM-based speech synthesis system

Figure 1. shows an overview of an HMM–based speech synthesis system(HTS)[10]. The system is composed of two, a training part and a synthesis part. In the training part, excitation parameters and spectral parameters are extracted from speech signal with context–dependent labels which input text is converted to using a text analyzer. With extracted parameters, system estimates HMM parameters. Then context–dependent HMMs of speech unit are modeled.

In the synthesis part, a speech signal corresponding to input labels which are also processed using a text analyzer as in the training part is synthesized. First procedure is the selection of the most similar HMMs to label. Second is a generation of excitation and spectral parameters HMMs. Since that parameters have information like pitch and spectral envelope, speech is synthesized finally.

3. The proposed phrase break analysis

The Korean language spaces out every eojeol. Although texts are designated by a whitespace, people read and break at phrase boundaries which they predicted. If they break at every whitespaces as designated in text, the naturalness of speech and the understanding of sentence are decreased. On the contrary to this, reading without breaks is also unnatural and is fast to understand. So, a breaking at phrase boundaries is important to synthesize natural and comprehensible sounds.

To increase a naturalness of synthesized speech, this paper predicts phrase boundaries and supplements phrase break information to labels. The phrase break strength is divided into 3 levels. ‘ Break 0’ means no break and ‘ break 2’ mostly appears at the end of sentence.

Table 1. Break level and duration

Break level	Break duration
Break 0	$0ms \le break < 50ms$
Break 1	$50ms \le break < 200ms$
Break 2	$500ms \le break$

Factors which are used in predicting phrase boundaries are as following. These factors are made to rules and then a rule–based phrase boundaries prediction is performed.

Table 2. Factors are used in prediction

- POS information (preceeding/succeeding)
- The presence or absence of punctuation mark
- The number of commas
- The distance of between punctuation marks
- The length of eojeol
- The position of eojeol in utterance
- The length of utterance

POS information

This paper divides Korean POS(Part–of–Speech) into nine POS and uses it of preceeding/succeeding eojeol of current whitespace. When a POS tag is the nominative or objective postposition, a probability of being a phrase boundary is high.

The information of punctuation marks

The whitespace which has a comma mainly becomes break point except some cases. Since comma is used to list words sometimes, consider the use of comma using a distance between commas and others.

Position and length

Eojeols which are closed to beginning or end of sentence have low probability to being a phrase boundary. In this case, use a position information of eojoel,

Phrase break is important in most cases, but it is unnecessary in short sentences. It is decided by length of sentence. Outside of that, length information also used to predict accurate phrase break boundaries.

4. Evaluation

The training and synthesis processes were performed using the HTS. To evaluate a naturalness of synthetic speech with phrase break information, a listening test was performed. The test used three types of speech. One is manually predicted, other is predicted using proposed method, and the other is without phrase break information. The total number of test sets is 4. And 7 peoples (5 males, 2 females) participated in this test. Figure 2 shows that the manually predicted one has the highest score, 3.45, a predicted one with phrase break information has the second highest score, 3.17, and the other which does not have phrase information has the lowest score, 3.13. It means that the proposed method increases a naturalness of synthetic speech.

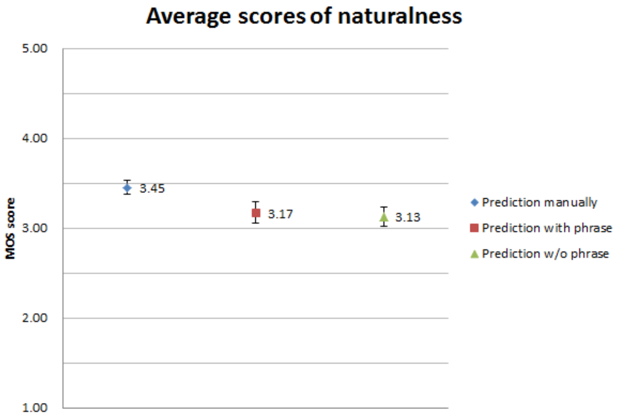

Figure 2. The average MOS scores of naturalness

Table 3. MOS Score

5	Very natural
4	Natural
3	Normal
2	Bad
1	Very bad

5. Conclusion

In this paper, phrase boundaries for break are predicted by rule which using context-dependent information. Synthetic speech breaks at not all whitespaces but predicted phrase boundaries. Evaluation shows that it makes sounds more natural and improves the comprehensibility of sentence. A manually predicted sentence had the highest score in MOS test. Since people have been trained to analyze the structure of a sentence when they read a text, it is a natural result. For future work, structure of sentence analysis is necessary to increase a naturalness of synthesized speech.

6. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y.Stylianou, and A. Syrdal, " *The AT&T next-gen TTS system,*" Citeseer, 1999.

[2] E. Moulines and F. Charpentier, " *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,*" Speech Commun., vol. 9, pp.453-467, Dec. 1990.

[3] A.W. Black, H. Zen, and K. Tokuda, " *Statistical parametric speech synthesis,*" in Proc. ICASSP, 2007, pp. 1229- 1232.

[4] J. Yu, M. Zhang, J. Tao, and X. Wang, " *A novel HMM-based TTS system using both continuous HMMs and discrete HMMs,*" in Proc. ICASSP, 2007, pp. 709- 712.

[5] Sanghun Kim, Youngjik Lee, and Keikichi Hirose, " *A new Korean corpus-based text-to-speech system,*" International Journal of Speech Technology 5, 105-116, 2002.

[6] P. Koehn, S. Abney, J. Hirschberg, and M. Collins, " *Improving intonational phrasing with syntactic information,*" In ICASSP, 2000.

[7] 김승원, 김병창, 정민우, 이근배, " *CRF 를 이용한 한국어 울율 경계 추정,*" 2005 년 제 17 회 한글 및 한국어 정보처리 학술대회, 2005.10,page(s):1-193.

[8] J. Lafferty, A. McCallum, and F. Pereira, " *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of International Conference on Machine Learning,*" ICML-01, pp.282-289, 2001.

[9] Y. Zheng, B. Kim, and G. G. Lee, " *Using multiple linguistic features for Mandarin phrase break prediction in maximum-entropy classification framework,*" Proceedings of the 8th international conference on spoken language processing(interspeech2004-ICSLP), 2004.

[10] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda, " *The HMM-based Speech Synthesis System(HTS) Version 2.0,*" in 6th ISCA Workshop on Speech Synthesis, 2007.