IEICE TRANSACTIONS

on Information and Systems

VOL.E90-D
NO.1
JANUARY 2007

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

IEICE TRANS. INF. & SYST., VOL.E90–D, NO.1 JANUARY 2007

LETTER

Two-Band Excitation for HMM-Based Speech Synthesis

Sang-Jin KIM[†a)], *Student Member and* **Minsoo HAHN**[†], *Nonmember*

SUMMARY This letter describes a two-band excitation model for HMM-based speech synthesis. The HMM-based speech synthesis system generates speech from the HMM training data of the spectral and excitation parameters. Synthesized speech has a typical quality of "vocoded sound" mostly because of the simple excitation model with the voiced/unvoiced selection. In this letter, two-band excitation based on the harmonic plus noise speech model is proposed for generating the mixed excitation source. With this model, we can generate the mixed excitation more accurately and reduce the memory for the trained excitation data as well.

key words: *HMM-based speech synthesis, excitation model, maximum voiced frequency*

1. Introduction

The HMM-based approach to speech synthesis uses the statistical HMMs to model the spectra and the excitation parameters of speech. It can produce speech with various voice characteristics by using speaker interpolation, adaptation or an eigenvoice technique [1]. This approach was originally proposed by Tokuda et al. [1], [2] and extended by Yoshimura et al. [3], [4]. Recently, the HMM-based speech synthesis technique has been reported for other languages [5]–[8], although it was originally developed to support Japanese.

The HMM-based speech synthesis is based on a source-filter model, so it is necessary to generate the excitation source signal. If the simple excitation model with the voiced/unvoiced selection is adopted, it is inevitable to get a typical quality of "vocoded sound".

Yoshimura et al. incorporated the mixed excitation model used in the mixed excitation linear predictive (MELP) vocoder, which was originally developed for narrowband speech [9], and reported improved speech quality [4]. However, the mixed excitation model extended for the wideband speech has coarse frequency analysis bands. The multi-band excitation in the MBE vocoder might be the more accurate model [10], but it is not suitable for the HMM-based speech synthesis because the number of frequency bands varies depending on the fundamental frequency. The two-band speech model can be considered as the simplified version of the MBE speech model. It assumes that voiced and unvoiced characteristics can be mixed in one speech frame, and their regions are divided into two bands [11]. In this letter, the two-band excitation model is suggested for the HMM-based speech synthesis.

The rest of this letter is organized as follows. Section 2 describes a shortcoming of the mixed excitation model briefly, and Sect. 3 introduces the suggested two-band excitation model for HMM-based speech synthesis. Experimental results are discussed in Sect. 4 while the conclusion is given in the final section.

2. Shortcoming of Mixed Excitation Model

The MELP vocoder is proposed by McCree et al. for low bit rate narrowband speech coding at 2.4 kbps [9], and has been chosen for the U.S. military standard speech coder in 1996 [12]. This vocoder has the following added capabilities: mixed pulse and noise excitation, the position jitter to reduce the tonal sound quality, adaptive spectral enhancement, and the pulse dispersion filter to match the filtered synthetic and natural speech waveform. More details are described in [9]. The MELP has been applied to the wideband speech coder [13], and Yoshimura et al. incorporated into the excitation model of HMM-based wideband speech synthesis [4]. However, the number of frequency bands for the band-pass voicing analysis is same as that of the narrowband MELP. The 5 bands of 0–0.5, 0.5–1, 1–2, 2–3, 3–4 kHz are simply extended to that of 0–1, 1–2, 2–4, 4–6, 6–8 kHz as shown in Fig. 1. These bands are not optimal even though the buzziness in the synthesized speech is surely reduced.

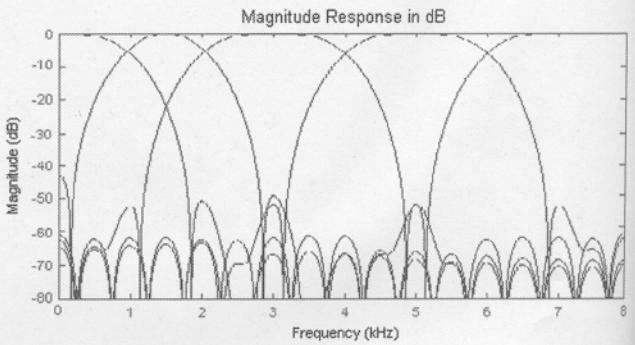

Fig. 1 Filterbanks in the mixed excitation model for the wideband speech.

Manuscript received June 5, 2006.
Manuscript revised August 21, 2006.

[†]The authors are with Multimedia Group, Information and Communications Univ., 119 Munjiro, Yuseong-gu, Daejeon, 305–732, Korea.
a) E-mail: sangjin@icu.ac.kr

DOI: 10.1093/ietisy/e90–d.1.378