Mixed Excitation for HMM-based Speech Synthesis

Takayoshi Yoshimura\(^1\), Keiichi Tokuda\(^1\), Takashi Masuko\(^2\), Takao Kobayashi\(^1\) and Tadashi Kitamura\(^2\)

\(^1\)Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan
\(^2\)Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan

{yossie,tokuda,kitamura}@ics.nitech.ac.jp, {masuko,Takao.Kobayashi}@ip.titech.ac.jp

Abstract
This paper describes improvements on the excitation model of an HMM-based text-to-speech system. In our previous work, natural sounding speech can be synthesized from trained HMMs. However, it has a typical quality of “vocoded speech” since the system uses a traditional excitation model with either a periodic impulse train or white noise. In this paper, in order to reduce the synthetic quality, a mixed excitation model used in MELP is incorporated into the system. Excitation parameters used in mixed excitation are modeled by HMMs, and generated from HMMs by a parameter generation algorithm in the synthesis phase. The result of a listening test shows that the mixed excitation model significantly improves quality of synthesized speech as compared with the traditional excitation model.

1. Introduction
We have proposed an HMM-based text-to-speech (TTS) system [1] (Fig. 1), in which spectral and excitation parameters are extracted from speech database and modeled by context dependent HMMs. In the synthesis part, spectral and excitation parameters are generated from HMM by using a speech parameter generation algorithm [2]. By filtering the excitation, a synthesis filter controlled by the spectral parameter generates speech. The system has the following features.

(1) Smooth and natural sounding speech can be synthesized from HMMs.
(2) The voice characteristics can be changed.
(3) It is “trainable”.

In the parameter generation algorithm, by taking account of statistics of both static and dynamic feature coefficients, the dynamics of the generated speech parameter sequence are constrained to be realistic. As for (2), by transforming HMM parameters appropriately, voice characteristics of synthesized speech can be changed since the system generates speech from the HMMs. In fact, we have shown that we can change voice characteristics of synthesized speech by applying a speaker adaptation technique [3] or a speaker interpolation technique [4]. As for (3), the system can be automatically constructed by embedded training of HMMs using only transcription and speech data without label boundaries.

In the previous work [1], natural sounding speech can be synthesized from trained HMMs. However, synthesized speech has a typical quality of “vocoded speech” since the HMM-based TTS system used a traditional excitation model with either a periodic impulse train or white noise shown in Fig. 2. To overcome this problem, the excitation model should be replaced with more precise one.

For low bit rate narrowband speech coding at 2.4kbit/s, the mixed excitation linear predictive (MELP) vocoder has been proposed [5]. In order to reduce the synthetic quality and mimic the characteristics of natural human speech, this vocoder has the following capabilities:

- mixed pulse and noise excitation
- periodic or aperiodic pulses

Figure 1: The scheme of HMM-based TTS system.
The mixed excitation is implemented using a multi-band mixing model, and can reduce the buzz of synthesized speech. Furthermore, aperiodic pulses and pulse dispersion filter reduce some of the harsh or tonal sound quality of synthesized speech. In recent years, the mixed excitation model of MELP has been applied not only to narrow-band speech coding but also to wideband speech coder [6] and speech synthesis system [7].

In this paper, mixed excitation model which is similar to the excitation model used in MELP is incorporated into the TTS system. Excitation parameters, i.e., pitch, bandpass voicing strengths and Fourier magnitudes, are modeled by HMMs, and generated from trained HMMs in synthesis phase.

The rest of this paper is organized as follows. The next section describes the mixed excitation model. The section 3 describes the HMM-based TTS system with mixed excitation model. Experimental results are given in the section 4, and concluding remarks and our plans for future work are given in the final section.

2. Mixed excitation

2.1. Analysis phase

In order to realize the mixed excitation model in the system, the following excitation parameters are extracted from speech data.

- pitch
- bandpass voicing strengths
- Fourier magnitudes

In bandpass voicing analysis, the speech signal is filtered into five frequency bands, with passbands of 0–1000, 1000–2000, 2000–4000, 4000–6000, 6000–8000Hz [6]. Note that the TTS system deals with 16kHz sampling speech. The voicing strength in each band is estimated using normalized correlation coefficients around the pitch lag. The correlation coefficient at delay \( t \) is defined by

\[
c_t = \frac{\sum_{n=0}^{N-1} s_n s_{n+t}}{\sqrt{\sum_{n=0}^{N-1} s_n^2} \sqrt{\sum_{n=0}^{N-1} s_{n+t}^2}},
\]

where \( s_n \) and \( N \) represent the speech signal at sample \( n \) and the size of pitch analysis window, respectively. The Fourier magnitudes of the first ten pitch harmonics are measured from a residual signal obtained by inverse filtering.

2.2. Synthesis phase

A block diagram of the mixed excitation generation and speech synthesis filtering is shown in Fig. 3.

The bandpass filters for pulse train and white noise are determined from generated bandpass voicing strength. The bandpass filter for pulse train is given by the sum of all the bandpass filter coefficients for the voiced frequency bands, while the bandpass filter for white noise is given by the sum of the bandpass filter coefficients for the unvoiced bands. The excitation is generated as the sum of the filtered pulse and noise excitations. The pulse excitation is calculated from Fourier magnitudes using an inverse DFT of one pitch period in length. The pitch used here is adjusted by varying 25% of its position according to the periodic/aperiodic flag decided from the bandpass voicing strength. By the aperiodic pulses, the system mimics the erratic glottal pulses and reduces the tonal noise. The noise excitation is generated by a uniform random number generator. The obtained pulse and noise excitations are filtered and added together.

By exciting the MLSA filter [8], synthesized speech is generated from the mel-cepstral coefficients, directly. Finally, the obtained speech is filtered by a pulse dispersion filter which is a 130-th order FIR filter derived from a
3. Text-to-speech synthesis with mixed excitation

3.1. Feature vector

The structure of the feature vector is shown in Fig. 4. The feature vector consists of spectral and excitation parameters.

Mel-cepstral coefficients including zero-th coefficient and their delta and delta-delta coefficients are used as spectral parameters. By using a mel-cepstral analysis technique [9] of an order of 24, mel-cepstral coefficients are obtained from speech signal windowed by a 25-ms Blackman window with a 5-ms shift.

Excitation parameters include pitch represented by log fundamental frequency \((\log f_0)\), five bandpass voicing strengths, Fourier magnitudes of the first ten pitch harmonics, and their delta and delta-delta parameters.

3.2. Context dependent model

Feature vectors are modeled by 5-state left-to-right HMMs. Each state of an HMM has four streams for mel-cepstrum, pitch, bandpass voicing strengths and Fourier magnitudes, respectively. In each state, mel-cepstrum, bandpass voicing strengths and Fourier magnitudes are modeled by single diagonal Gaussian distributions, and pitch is modeled by the multi-space probability distribution [10].

Feature vectors are modeled by context dependent HMM taking account of contextual factors which affect spectral parameter and excitation parameter such as phone identity factors, stress-related factors and locational factors. Details of the contextual factors are shown in [1]. The trained context dependent HMMs are clustered using a tree-based context clustering technique based on MDL principle [11]. Since each of mel-cepstrum, pitch, bandpass voicing strength, Fourier magnitude and duration has its own influential contextual factors, the distributions for each speech parameter are clustered independently, where state occupation statistics used for clustering are calculated from only the streams of mel-cepstrum and pitch.

3.3. HMM-based TTS system

A context dependent label sequence is obtained by text analysis of input text, and a sentence HMM is constructed by concatenating context dependent phoneme HMMs according to the obtained label sequence. By using a speech parameter generation algorithm [12], mel-cepstrum, pitch, bandpass voicing strength and Fourier magnitude are generated from the sentence HMM taking account of their respective dynamic feature statistics. Speech is synthesized from the obtained spectral and excitation parameters as described in section 2.2.

4. Experiments

4.1. Excitation generation

Excitation parameters were generated from an HMM set trained using phonetically balanced 450 sentences of ATR Japanese speech database. The resulting decision trees for mel-cepstrum, pitch, bandpass voicing strength, Fourier magnitude and state duration models had 934, 1055, 1651, 3745 and 1016 leaves in total, respectively. Examples of traditional excitation and mixed excitation are shown in Fig. 5, where the pulse dispersion filter was applied to mixed excitation. From the figure, it can be observed that the voiced fricative consonant “z” has both the periodic and aperiodic characteristics in the mixed excitation.
4.2. Subjective evaluation

The TTS system with mixed excitation model was evaluated. We compared traditional excitation and mixed excitation by a pair comparison test. In addition, effects of the Fourier magnitudes, aperiodic pulses and the pulse dispersion filter were evaluated.

The following five excitation models were compared:

- TE: traditional excitation
- ME: mixed excitation
- FM: ME + Fourier magnitudes
- AP: FM + aperiodic pulses
- PD: AP + pulse dispersion filter

The model TE is the traditional excitation model which generates either periodic pulse train or white noise. Each of models ME, FM, AP and PD is the mixed excitation model. In the model ME, pulse train was not calculated from Fourier magnitude, and the aperiodic pulse and the pulse dispersion filter were not applied. In the model FM, pulse excitation was calculated from Fourier magnitude. The model AP used aperiodic pulses, and the model PD used the pulse dispersion filter additionally. Eight subjects tested the five kinds of synthesized speech. Eight sentences were selected at random for each subjects from 53 sentences which were not included in the training data. Figure 6 shows preference scores. It can be seen that the mixed excitation model significantly improved the quality of synthetic speech. Although no additional gain was obtained by using Fourier magnitudes and aperiodic pulses, the additional use of pulse dispersion filter achieved further improvement in speech quality.

5. Conclusions

In this paper, we have described an HMM-based speech synthesis system with a mixed excitation model. The pair comparison test has shown that the quality of synthesized speech is significantly improved by using mixed excitation model. In addition, the speech quality is further improved by the pulse dispersion filter.

Our TTS system can adopt other high-precision excitation models since MSD-HMM can deal with vectors having variable dimensionality. The improvement on the presented mixed excitation model will allow us to synthesize higher quality speech. The future work also includes the use of other high-precision excitation models. The synthesized speech generated by the latest system can be found at http://kt-lab.ics.nitech.ac.jp/~yossie/TTS/.

6. References