HMM-based speech synthesis using sub-band basis spectrum model

Yamato Ohtani, Masatsune Tamura, Masahiro Morita, Takehiko Kagoshima, Masami Akamine

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation, Japan
yamato.ohtani@toshiba.co.jp, masatsune.tamura@toshiba.co.jp, masahiro.morita@toshiba.co.jp,
takehiko.kagoshima@toshiba.co.jp, masa.akamine@toshiba.co.jp

Abstract
In this paper, we propose HMM-based text-to-speech (TTS) using sub-band basis spectrum model (SBM). SBM can represent vocal tract spectra and phase characteristics by a linear combination of sub-band basis vectors. Some reports suggest that analysis-synthesized speech based on SBM is close to natural speech and SBM can perform effectively in TTS. Therefore, the SBM framework is expected to have good effects on HMM-based TTS by improving speech quality. Subjective experimental results show that the proposed method improves speech quality in some conditions.

Index Terms: speech synthesis, hidden Markov model, sub-band basis spectrum model, phase feature

1. Introduction
A stochastic text-to-speech (TTS) system based on hidden Markov model (HMM) [1][2] has been studied widely. HMM-based TTS simultaneously models acoustic features such as vocal tract spectrum and source information such as fundamental frequency (F0). It generates acoustic feature sequences from phonetic and linguistic information of an input text, and then generates a waveform from the acoustic feature sequences using a vocoder system. In general, speech quality of HMM-based TTS depends on qualities of acoustic features and an excitation model. A simple excitation model generates buzzy sound, and insufficient quality of acoustic features results in muffled speech. Therefore, there is a need to introduce high-quality acoustic features and excitation model into HMM-based TTS systems in order to improve synthesized speech quality.

In many HMM-based TTS systems, mel-cepstrum (MCEP) [3] or mel-scaled line spectral pair (MLSP) [4] is often used as a vocal tract parameter. Lower dimensions of MCEP represent macro structure of the vocal tract spectrum, and higher parts include micro structure. Although MCEP is easy to apply to speaker adaptation techniques [5], the synthetic speech quality by HMM-based TTS using this feature tends to be muffled. This is because high-dimensional coefficients of MCEP tend to be reduced in modeling, and then the micro structure of the vocal tract spectrum disappears. On the other hand, MLSP is one of the speech parameters based on linear prediction analysis. MLSP has better quantization and interpolation characteristics than MCEP. The clarity of the synthesized speech based on MLSP is also better than that of the synthesized speech based on MCEP because spectral peaks of MLSP are sharper than those of MCEP. However, HMM-based TTS with MLSP destabilizes the synthesis filter more often than one with MCEP because of disordering and too small intervals between adjacent MLSP coefficients. Although some HMM-based TTS systems [6][7] have improved these qualities by using STRAIGHT [8], basic problems remain.

The traditional HMM-based TTS employs a simple excitation model in which source signals are generated by selecting pulse trains for voiced segments and white noise signals for unvoiced segments. The excitation model is too simple to represent actual source signals. To improve the excitation model, NIT’s HMM-based TTS [6] has introduced aperiodic component [9], which represents a ratio between periodic component and noise component in any frequency band. Its excitation model is realized by a weighted sum of phase-manipulated pulse train and a white noise sequence based on aperiodicity in voiced frames. Although this approach improved synthesized speech quality, it is not considered other factors included in the glottal signal. To alleviate this mismatch to the glottal signal, several approaches were proposed. Maia et al. introduced a residual model represented by voiced and unvoiced filters extracted from a residual signal [10]. This method has overcome the buzzy sound included in the synthesized speech. However, harsh sounds are caused to a slight extent. In the HMM-based TTS system proposed by Raitio et al. [11], voiced components of the mixed excitation signals are generated using actual glottal source signals selected from a large library. While this approach is reasonable, a large amount of memory may be required to keep the glottal source library.

In this paper, we propose introduction of the sub-band basis spectrum model (SBM) [12] to HMM-based TTS. SBM can represent vocal tract spectra and phase characteristics by a linear combination of sub-band basis vectors, which are given by approximating the result of applying a sparse coding method [13] to speech spectra. SBM is satisfied with the following characteristics: 1) the number of dimensions is constant and each dimension represents a specific acoustic space, 2) frequency domain processing can be easily applied, and 3) Euclidean distance of the parameter is equivalent to perceptual weighted spectral distance. According to [12], analysis-synthesized speech based on SBM is close to
natural speech. In addition, SBM has already been applied to a speaker adaptation method for a unit selection-based speech synthesis system [14] and has been reported to perform well while keeping high speech quality. Therefore, SBM is expected to improve the speech quality of HMM-based TTS. Subjective evaluations suggest that the proposed approach has improved speech quality in some conditions.

The rest of this paper is organized as follows. Section 2 describes the SBM framework. In section 3, we propose the HMM-based TTS system using SBM parameter. Section 4 shows results of experimental evaluations. Finally, we concluded this paper in section 5.

2. Sub-band basis spectrum model

2.1. Overview of sub-band basis spectrum model framework

Figure 1 shows an overview of the SBM process for speech spectrum vector of the $t$ frame $x_t = [x_t(1), x_t(2), \cdots, x_t(K)]^T$ ($T$: transposition) is analyzed with sub-band basis vectors and then a weight vector is extracted. The obtained weight vector is a SBM parameter vector. In decoding, speech spectrum $x_t$ is rebuilt by using weight parameters and the sub-band basis vectors as follows:

$$x_t = \Phi c_t, \quad (1)$$

where, $c_t = [c_t(0), c_t(1), \cdots, c_t(N-1)]^T$ and $\Phi = [\Phi_0, \Phi_1, \cdots, \Phi_{N-1}]$ represents weight parameter vector and sub-band basis vectors, respectively.

2.2. Sub-band basis vector

In the SBM framework, sub-band basis vectors have the following characteristics: 1) there is one frequency band in each basis vector, 2) the frequency band is asymmetrical unimodal in shape, 3) a frequency band overlaps other adjacent frequency bands and 4) peak frequencies less than $\pi/2$ [rad] conform to mel-scale and the others keep linear scale. Figure 2 shows some sub-band basis vectors. Sub-band basis vectors satisfy characteristics 1), 2) and 3). It can be seen that the band width included in a sub-band basis vector is different from others. The frequency scale described in the characteristic 4) is formulated as follows:

$$\tilde{\Omega}(n) = \begin{cases} \frac{\Omega_n + 2 \tan^{-1} \frac{\alpha \sin \Omega_n}{1 - \alpha \cos \Omega_n}}{N - N_w}, & 0 \leq n < N_w \\ \frac{n - N_w \pi + \pi}{2N_w}, & N_w \leq n < N, \end{cases} \quad (2)$$

where, $\alpha$, $\Omega(n)$ [rad], $\Omega_n$ [rad] and $N_w$ denote a warping parameter, the peak frequency of the $n$th basis vector, $n\pi/N_w$ [rad] the dimension index that satisfies $\tilde{\Omega}(N_w) = \pi/2$ [rad], respectively. Using the defined frequency scale, $n$th sub-band basis vector $\phi_n = [\phi_n(0), \phi_n(1), \cdots, \phi_n(K)]$ is written by

$$\phi_n(k) = \begin{cases} 0.5 - 0.5 \cos \left( \frac{k - \Omega(n)}{\Omega(n) - \tilde{\Omega}(n + 1) - \tilde{\Omega}(n)} \pi \right), & \tilde{\Omega}(n) - 1 < k \leq \tilde{\Omega}(n) \\ 0.5 - 0.5 \cos \left( \frac{k - \tilde{\Omega}(n)}{\tilde{\Omega}(n + 1) - \tilde{\Omega}(n)} \pi + \pi \right), & \tilde{\Omega}(n) \leq k < \tilde{\Omega}(n + 1) \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

2.3. SBM parameter estimation

SBM weight parameters $\hat{c}_t$ related to a speech spectrum $x_t$ are estimated by minimizing error between the original spectrum and the decoded one as follows:

$$\hat{c}_t = \arg \min |x_t - \Phi c_t|. \quad (4)$$

According to [12], vocal tract SBM weight parameters (VTSP), which are extracted from logarithmic vocal tract spectra analyzed by pitch-synchronous Fourier transform, are obtained by non-negative least squares [15]. On the other hand, phase SBM weight parameters (PSP) related to time-scaled unwrapped phase spectra are estimated by basic least squares.

3. Introduction of SBM parameters

3.1. SBM weight parameter for HMM-based TTS

The preliminary result shows that it is difficult to apply original SBM weight parameters to HMM-based TTS directly. Therefore, it is necessary to convert each SBM weight parameter into the proper one for HMM-based TTS.

Figure 1: Overview of sub-band basis spectrum model

Figure 2: Parts of sub-band basis vectors (Number of basis vectors is 20).
3.1. VTSP for HMM-based TTS

In a preliminary test, we confirmed that HMM-based TTS using original VTSP had very low quality. This is because VTSPs are correlated to one another and the basic HMM-based TTS cannot be trained appropriately. In order to alleviate correlations among VTSP, a discrete cosine transform (DCT), which is one of the orthogonal transforms, is applied to them. Here, the VTSP converted by DCT (DCT-VTSP) is similar to MCEP. In MCEP, the number of dimensions is generally reduced by higher-dimensional coefficients. Therefore, MCEP could lose necessary factors. By contrast, DCT-VTSP is produced from a spectrum parameter to which a dimension compression has been applied. Thus, unlike MCEP, the proposed parameter is able to keep mandatory factors.

3.1.2. PSP for HMM-based TTS

We confirmed in a preliminary test that HMMs could not model the original PSP with too large variances, properly. To introduce phase parameters to HMM-based TTS, instead of time-scaled phase spectra, we use the following complex number representation:

$$\xi_t(\omega_k) = \cos \theta_t(\omega_k) + j \sin \theta_t(\omega_k), \quad (5)$$

where, \(j\) is imaginary unit and \(\theta_t(\omega_k)\) [rad] denotes the phase of \(t^{th}\) frame on frequency \(\omega_k\) [rad]. SBM analysis is applied to the real part and the imaginary part in eq. (5) respectively and we use a set of them as a phase parameter. We call this parameter complex PSP (CPSP) in this paper. Although the number of CPSP components is twice large as that of the conventional ones, the range of CPSP is limited to from \(-1\) to \(1\), and it is also unnecessary to consider phase wrapping. This allows us to introduce phase information to HMM-based TTS easily.

3.2. HMM-based TTS with SBM weight parameter

Figure 3 shows the proposed HMM-based TTS system. In the training process, context features, log-F0, banded aperiodic component (BAP), DCT-VTSP and CPSP are extracted from Speech DB. Then these parameters are modeled by context-dependent hidden semi-Markov models (HSMM) [16]. In the context clustering, decision trees are built based on the maximum likelihood criterion [2]. In this paper, the decision tree for the spectrum parameter is shared also for the phase parameter.

In the synthesis process, context-dependent HSMMs are selected by contexts included in an input text and then each parameter sequence is generated based on the maximum likelihood criterion [2]. A synthesized waveform is generated by the process shown in Figure 4. In the mixed excitation generation, CPSP sequence is decoded using sub-band basis vectors. Multi-pulse trains are built based on the log-F0 and decoded phase sequence. Then, a mixed excitation is generated by weighted sum of multi-pulse train and white noise based on BAP. In filtering, DCT-VTSP sequence is converted into VTSP by inverse DCT and decoded to vocal tract spectral sequence by sub-band basis vectors. The synthesized waveform is produced by filtering the mixed excitation with the decoded spectral sequence.

4. Experimental evaluation

4.1. Experimental conditions

We subjectively evaluated speech quality of the HMM-based TTS system using the proposed parameters and together with the one using the conventional parameters as shown in Table 1. In the evaluation, we employed one Japanese male database and two different databases of the same Japanese female speaker with different numbers of utterances. They respectively consisted of 853, 572 and 2000 utterances. The sampling frequency was 22.05 [kHz]. The speech waveforms were analyzed by a pitch-synchronous Fourier transform with 1024 points. MLSP, MCEP and DCT-VTSP were extracted from pitch-synchronous spectra, and the respective parameters included delta coefficients. BAP was calculated with the pitch-scaled harmonic filter [18] and consisted of 5-dimensional components (0–1.4, 1.4–3.1, 3.1–5.5, 5.5–8.3 and 8.3–11 [kHz]) and their delta components. A log-F0 vector was composed of a static component, their delta and delta-delta. CPSP was composed of 50-dimensional real and imaginary values, and their deltas.

This evaluation was carried out on our crowdsourcing system. 12 test sentences were used in this evaluation. The number of listeners was around 25 in each evaluation. Each listener evaluated speech quality of each target speaker by the 5-level mean opinion score (1: bad
Table 1: Speech synthesis systems for evaluation

<table>
<thead>
<tr>
<th>Label</th>
<th>System</th>
<th>Spectral type</th>
<th>CPSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>L40N</td>
<td>HMM</td>
<td>39-dims MLSP</td>
<td>no use</td>
</tr>
<tr>
<td>L40P</td>
<td>HMM</td>
<td>39-dims MLSP</td>
<td>use</td>
</tr>
<tr>
<td>C40N</td>
<td>HMM</td>
<td>39-dims MCEP</td>
<td>no use</td>
</tr>
<tr>
<td>C40P</td>
<td>HMM</td>
<td>39-dims MCEP</td>
<td>use</td>
</tr>
<tr>
<td>C50N</td>
<td>HMM</td>
<td>49-dims MCEP</td>
<td>no use</td>
</tr>
<tr>
<td>C50P</td>
<td>HMM</td>
<td>49-dims MCEP</td>
<td>use</td>
</tr>
<tr>
<td>S50N</td>
<td>HMM</td>
<td>49-dims DCT-VTSP</td>
<td>no use</td>
</tr>
<tr>
<td>S50P</td>
<td>HMM</td>
<td>49-dims DCT-VTSP</td>
<td>use</td>
</tr>
</tbody>
</table>

Figure 5: Results of subjective evaluation

and 5: excellent).

4.2. Subjective evaluations

Figure 5 shows results of the subjective evaluations. In the result of the male speaker, the proposed systems such as S50N and S50P achieved better speech quality than C40N, C50N and C50P, while L40N had the best score. And, C40P was better than C40N.

In the case of the female speaker 1, S50N and S50P were better than any other HMM-based TTS system. In this result, the systems with the phase parameter tended to improve its voice quality slightly compared with those without phase parameter.

In the result of the female speaker 2, the proposed system S50P had the best speech quality. In addition, phase parameter was able to improve speech quality slightly. However, with regard to the speech quality of HMM-based TTSs, there were no significant differences.

The results of the subjective evaluations suggest that, compared with the conventional HMM-based speech systems, the HMM-based TTS using the SBM parameter may improve speech quality.

5. Conclusions

This paper has proposed an HMM-based TTS system using the sub-band basis spectrum model (SBM). The SBM could parameterize not only vocal tract spectrum but also the phase spectrum. Subjective experimental results have shown that, compared with the conventional systems, the proposed systems conditionally achieved better speech quality and are promising for realization of natural synthesized speech.

6. References