ABSTRACT
This paper presents a very low bit-rate F0 coding technique for speaker-dependent phonetic vocoder based on hidden Markov model (HMM) using quantized F0 context. In the proposed technique, the input F0 sequence is converted into F0 symbol sequence at a phoneme level using scalar quantization. The quantized F0 symbols are used in the decoding process as the prosodic context for the HMM-based speech synthesis. The synthetic speech is generated from the context-dependent labels and input speaker's pre-trained HMMs by using the HMM-based parameter generation algorithm. By taking account account of preceding and succeeding phonemes and F0 symbols as the contextual factors, we can generate smooth F0 trajectory similar to that of the original with only a small number of quantization bits. Experimental results demonstrate that the proposed technique can generate F0 contour with acceptable quality even when the bit-rate is less than 50 bps.

Index Terms—phonetic vocoder, HMM-based speech synthesis, very low bit-rate speech coding, F0 context, multi-space distribution HMM

1. INTRODUCTION
For the purpose of very low bit-rate speech coding with less than several hundreds bps, a phonetic vocoder (e.g., [1, 2]) is one of the most popular techniques. In the phonetic vocoder, input speech is decomposed into frame sequences of spectral feature and the fundamental frequency (F0). The spectral feature sequence is encoded into a phoneme symbol sequence using a speech recognizer and is transmitted with durations. The F0 sequence is also encoded, then transmitted separately.

Recently, several F0 coding techniques have been proposed for very low bit-rate coders. In [3], piecewise linear approximation [4] was used, which is similar to polygon approximation [5] in image coding. However, the decoded F0 contour was linear within each segment and F0 variations in segment boundaries were not smooth. To alleviate the problem, an alternative technique [6] was proposed using vector quantization based on multi-space probability distribution (MSD-VQ). In this technique, F0 values are modeled with the MSD where the observation space of F0 features is represented by a union of voiced and unvoiced spaces. Although this approach is feasible to statistically treat F0 values, codebooks are separately trained for respective phonemes and codewords depend only on current phonemes.

In this paper, we propose a novel F0 coding technique based on MSD-HMM [7] for further improvement in F0 coding performance. We employ quantized F0 context which has been originally proposed for unsupervised F0 modeling [8] in HMM-based speech synthesis. In the encoding process, an input F0 sequence is converted into an F0 symbol sequence at a phone level. We quantize the average log F0 value of each phone unit into a discrete level. The F0 symbol sequence represents a rough shape of the original F0 contour. From the experimental results of the previous study [8], we found that these symbols can be used as a prosodic context for HMM-based speech synthesis. After phoneme recognition and F0 quantization, the phoneme and F0 symbol sequences are transmitted to the decoder. For speech synthesis, context-dependent labels are created from the transmitted symbols. Synthetic speech is then generated using an HMM-based parameter generation algorithm [9].

2. PHONETIC VOCODER BASED ON MSD-HMM

2.1. HMM-based speech synthesis

Here, we briefly review the HMM-based speech synthesis, which plays a primary role in the decoding process. In speech recognition, only spectral features such as mel frequency cepstral coefficients (MFCCs) are modeled with HMM. However, we need to model and generate F0 features as well as spectral ones for speech synthesis. For this purpose, multi-space distribution HMM (MSD-HMM) [7] can be used to model not only voiced frames but also unvoiced frames. The model training procedure is based on a maximum likelihood (ML) estimation of model parameters using an expectation-maximization (EM) algorithm and is almost the same as that for speech recognition. When synthesizing speech in text-to-speech application, phonetically and prosodically context-dependent labels are created from a given text using text analysis. Then, spectral and F0 feature sequences are generated using an HMM-based parameter generation algorithm [10], which is also based on an ML criterion.

2.2. Overview of the proposed coder

A block diagram of the proposed speech coder is illustrated in Fig. 1. In the encoder, spectral and F0 feature sequences are extracted from the input speech. For spectral features, we use MFCCs, which are widely used in state-of-the-art automatic speech recognition systems. A phoneme sequence with durations is obtained from the MFCC sequence of the input speech using phoneme recognizer with input speaker’s pre-trained HMMs. The F0 sequence is converted into an F0 symbol sequence using F0 quantization, which is described in Sect. 2.3. Then, a phonetically and prosodically context-dependent label sequence is created from phoneme, duration, and F0 symbol sequences in the same manner as [8]. In the decoder, synthetic speech is generated from the label sequence using HMM-based speech synthesis as described in Sect. 2.1.
2.3. F0 encoding using phone-level F0 quantization

In the F0 encoding, the extracted F0 contour is converted to an F0 symbol sequence at a phoneme level. Each F0 symbol is obtained by roughly quantizing the average log F0 value of each phone. The resultant F0 symbol sequence represents the outline of the original F0 contour. In our previous studies on unsupervised F0 modeling [8], we showed that these F0 symbols can be used as a prosodic context for HMM-based speech synthesis.

In the F0 quantization, we conduct scalar quantization based on the statistical parameters of the input speaker’s F0 values. Specifically, we assume that the log F0 values of the input speaker follow a normal distribution. Before quantizing F0, we calculate global mean $\mu$ and variance $\sigma^2$ of log F0 values for the input speaker’s training data. We calculate the mean $\bar{F}_p$ of the log F0 values for each phone $p$, where the phone boundaries obtained in the phoneme recognition are used. Then, an F0 symbol $s_p$ is obtained by quantizing $\bar{F}_p$ into a discrete value out of $2^N - 1$ symbols as follows:

$$ s_p = Q(\bar{F}_p), \quad s_p \in \{1, \ldots, 2^N - 1\}, $$

where $Q(\cdot)$ denotes an operation of scalar quantization, and $N$ is the number of quantization bits. Additional one symbol is assigned when the phone segment is unvoiced. An example of the distribution of log F0 values for a male speaker included in ATR Japanese speech database is shown in Fig. 2. Since we have empirically found that there are only a few log F0 values lying outside the interval $[-3\sigma, 3\sigma]$, we set the $2^N - 1$ points which equally divide the interval $[-3\sigma, 3\sigma]$ as the codewords. An example of three-bit F0 quantization is shown in Fig 3.

The quantized F0 symbols are then transmitted to the decoder with phoneme labels and durations. Although entropy coding can be used for transmitting these symbols with durations in a manner similar to that in [6], we here focus on the F0 coding and do not use such additional techniques to reduce bit-rates to reveal the intrinsic nature of the proposed approach.

2.4. F0 decoding from context-dependent MSD-HMMs

In the decoder, a context-dependent label sequence for speech synthesis is created from phoneme, duration, and F0 symbol sequences. For the labels, preceding, current, and succeeding phonemes and F0 symbols are used as the contextual factors for spectral and F0 features. We train MSD-HMMs using the input speaker’s speech data and label information in advance. We use STRAIGHT analysis [11] for speech feature extraction, and extract spectral envelope, F0, and aperiodicity features. The spectral envelope is then converted to mel-cepstral coefficients using a recursion formula. The aperiodicity feature is also converted to average values for five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. Labels for training data are also created using F0 quantization as described in Sect. 2.3. For the given label sequence, mel-cepstrum, F0, and aperiodicity features are generated from MSD-HMMs using a parameter generation algorithm based on the ML criterion [9]. Finally, the speech waveform is synthesized using STRAIGHT synthesis. By using the F0 symbols of preceding and succeeding phones as the F0 context in the synthesis labels, we can efficiently save the number of quantization bits compared to the case where we use only the current F0 symbols.
Table 1. Bit-rates [bits/sec] of two techniques.

<table>
<thead>
<tr>
<th>number of quantization bits</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame-based</td>
<td>200</td>
<td>400</td>
<td>600</td>
<td>800</td>
<td>1K</td>
</tr>
<tr>
<td>MSD-HMM</td>
<td>16</td>
<td>31</td>
<td>47</td>
<td>63</td>
<td>79</td>
</tr>
</tbody>
</table>

3. EXPERIMENTS

3.1. Experimental conditions

In the following experiments, we used reading style speech of a male speaker MHT included in ATR Japanese speech database set B. The speaker uttered 503 phonetically balanced Japanese sentences. We used 450 sentences for model training and the remaining 53 sentences for evaluation. The phoneme recognition rate including insertion error for the test data was 82.1%. The average phoneme rate computed from recognition results for the test data was 13.8 phonemes/sec. Speech signals were sampled at a rate of 16KHz, and the interval of frame shift was 5-ms (200 frames/sec). In the encoder, we used 26 MFCCs including static and delta parameters, where the log energy was normalized. In the decoder, we used a feature vector consisted of 39 mel-cepstral coefficients including the zeroth coefficient, log F0, 5-band aperiodicity values, and their delta and delta-delta coefficients. The total number of dimension was 138.

We used 3-state left-to-right with no skip topology both in HMM and MSD-HMM. The output distribution in each state was modeled with a single Gaussian density function, and covariance matrices of these models were assumed to be diagonal. In context clustering for parameter tying, a decision tree was automatically constructed based on the minimum description length (MDL) criterion [12]. We used questions on F0 contexts for the clustering of spectral features as well as F0 features. To mitigate perceptual degradation of synthetic speech quality caused by over-smoothing, we used a parameter generation algorithm considering GV [13] only in the subjective evaluation.

3.2. Objective evaluation

We first evaluated performance of the proposed F0 coding technique for different numbers of quantization bits from 1 to 5. For the objective measure of F0 distortion, we used the root mean square (RMS) error of log F0 between original and synthetic speech samples. For comparison, we also evaluated simple frame-based coding. The frame-based coding was conducted using scalar quantization, where the same codewords were used as the proposed technique. In the frame-based coding, we smoothed decoded F0 contours using moving average with five preceding and succeeding frames to alleviate discontinuity between frames. The bit-rates for the two techniques with respective quantization bits are listed in Table 1, where the average bit-rate \( r \) [bits/sec] is given by

\[
r = n \times r_p,
\]

where \( n \) is the number of quantization bits and \( r_p \) is average number of phonemes per second for each speaker.

Figure 4 plots the average RMS error of log F0 of 53 test sentences. The distortion in the proposed technique with only one-bit quantization was significantly lower than the frame-based quantization. It is noted that the total bit-rate of the frame-based coding is much higher than that in the proposed technique.

3.3. Subjective evaluation

To evaluate perceptual quality of F0 coding, we conducted subjective evaluation by a degradation MOS (DMOS) test. Synthetic speech samples were generated using the decoded F0 of two techniques and original spectral and aperiodicity features. This means that we focused on the degradation of F0 through speech coding in this experiment. Vocoder speech was used as the reference. Seven participants evaluated perceptual similarity of decoded speech samples to the reference speech. Eight sentences were chosen randomly for each participant from the 53 test sentences of six speakers, and participants rated degradation on a five-point scale, i.e., 1 for very annoying, 2 for annoying, 3 for slightly annoying, 4 for audible but not annoying, and 5 for inaudible. The scores are shown in Fig. 5 with confidence intervals of 95%. Although the bit-rate of the proposed technique is much smaller than that of the frame-based, F0 coding accuracy of our technique is comparable to that of the frame-based coding in three- and four-bit quantization. Moreover, the proposed technique is significantly better than the frame-based coding in two-bit quantization. Figure 6 shows examples of generated F0 contours with different numbers of quantization bits.

Next, we evaluated overall subjective quality of phonetic vocoders using spectral and aperiodicity features generated from MSD-HMM. Seven participants evaluated perceptual quality of decoded speech samples. Eight sentences were chosen randomly for each participant from the 53 test sentences of six speakers, and participants rated speech quality on a five-point scale, i.e., 1 for bad, 2 for poor, 3 for fair, 4 for good, and 5 for excellent. Fig. 7 shows the scores with confidence intervals of 95%. From the result, it is seen that the proposed technique outperforms the frame-based coding in
all numbers of quantization bits. This indicates that simultaneous modeling of spectral and F0 features is very important in very low bit-rate speech coding. In addition, the quality of the decoded speech with our technique seems to be acceptable even when the bit-rate of F0 coding is less than 50 bps.

4. CONCLUSIONS

In this paper, we have proposed a very low bit-rate F0 coding technique for the HMM-based phonetic vocoder. The proposed technique utilizes roughly quantized F0 symbols as a prosodic context for the HMM-based speech synthesis in the decoding process. By taking account of preceding and succeeding F0 symbols as contexts, the decoded F0 contour is very smooth and similar to that of the original even when the number of quantization bits is very small such as three. Experimental results demonstrated that our F0 coding technique achieved approximately 31 bps with acceptable speech quality. In future work, we will address a speaker-independent version of the proposed technique.

5. ACKNOWLEDGMENTS

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063 and 21800020.

6. REFERENCES