A VERY LOW BIT RATE SPEECH CODER USING HMM-BASED SPEECH RECOGNITION/SYNTHESIS TECHNIQUES

Keiichi Tokuda †, *Takashi Masuko* ‡, *Jun Hiroi* †, *Takao Kobayashi* ‡, *and Tadashi Kitamura* †

†Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan
‡Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 227 Japan
Email: {tokuda,jun,kitamura}@ics.nitech.ac.jp, {masuko,tkobayas}@pi.titech.ac.jp

ABSTRACT

This paper presents a very low bit rate speech coder based on HMM (Hidden Markov Model). The encoder carries out phoneme recognition, and transmits phoneme indexes, state durations and pitch information to the decoder. In the decoder, phoneme HMMs are concatenated according to the phoneme indexes, and a sequence of mel-cepstral coefficient vectors is generated from the concatenated HMM by using an ML-based speech parameter generation technique. Finally we obtain synthetic speech by exciting the MLSA (Mel Log Spectrum Approximation) filter, whose coefficients are given by mel-cepstral coefficients, according to the pitch information. A subjective listening test shows that the performance of the proposed coder at about 150 bit/s (for the test data including 26 % silence region) is comparable to a VQ-based vocoder at 400 bit/s (= 8 bit/frame × 50 frame/s) without pitch quantization for both coders.

1. INTRODUCTION

To code speech at rates on the order of 100 bit/s, phonetic and segment vocoders are the most popular techniques [1]-[6]. These coders decompose speech into a sequence of speech units (i.e., phonetic units and acoustically derived segment units, respectively) by using a speech recognition technique, and transmit the obtained unit indexes and unit durations. The decoders synthesize speech by concatenating typical instances of speech units according to the unit indexes and unit durations.

This paper presents a phonetic vocoder based on HMM (Hidden Markov Model), in which speech spectra are consistently represented by mel-cepstral coefficients obtained by a mel-cepstral analysis technique [7], and the sequence of mel-cepstral coefficient vectors for each speech unit is modeled by phoneme HMM. The encoder carries out phoneme recognition which adopts advanced techniques used in the area of speech recognition, and transmits phoneme indexes and state durations to the decoder by using entropy coding and vector quantization. Pitch information is also transmitted to the decoder whereas this paper excludes pitch quantization from consideration. In the decoder, phoneme HMMs are concatenated according to the phoneme indexes, and the state sequence is determined from the transmitted state durations. Then a sequence of mel-cepstral coefficient vectors is determined in such a way that the likelihood of the sequence of mel-cepstral coefficient vectors is maximized for the concatenated HMM and the state sequence [8]-[10]. Finally speech signal is synthesized by the MLSA (Mel Log Spectrum Approximation) filter according to the obtained mel-cepstral coefficients [7].

This work was partially supported by the Ministry of Education, Science and Culture of Japan, Grant-in-Aid for Encouragement of Young Scientists, 09750399, 1997.

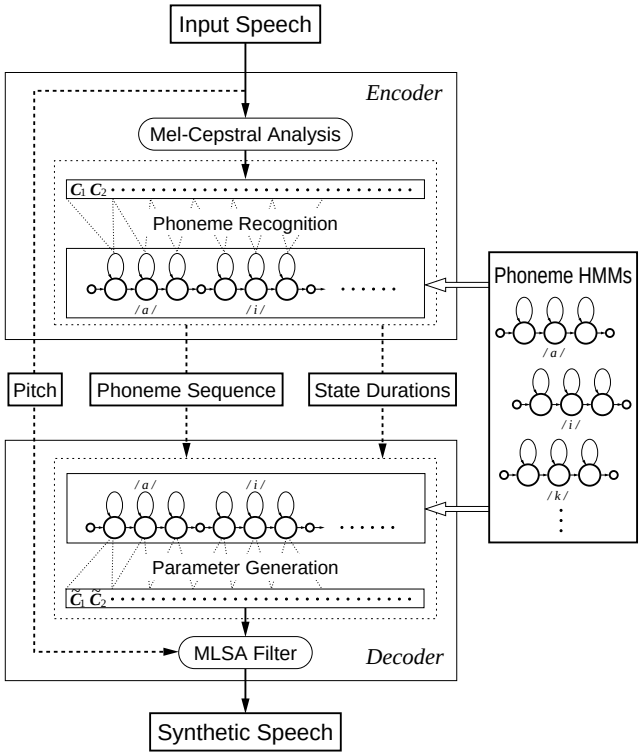

Figure 1: A very low bit rate speech coder based on HMM.

In the following, we summarize the HMM-based speech synthesis technique, and describe the proposed coder in Section 2 and 3, respectively. The results of a subjective evaluation test are shown in Section 4.

2. HMM-BASED SPEECH SYNTHESIS

A block diagram of the proposed speech coder is illustrated in Fig. 1. The encoder is equivalent to an HMM-based phoneme recognizer, and the decoder does the inverse operation of the encoder

using an HMM-based speech synthesis technique [10], which consists of two techniques: a mel-cepstrum-based vocoding [7] and HMM-based speech parameter generation [8], [9]. This section summarizes these two techniques.

2.1. Vocoding Technique Based on Mel-Cepstrum

Since we model speech spectrum using $M+1$ mel-cepstral coefficients, i.e., frequency-transformed cepstral coefficients, the minimum phase synthesis filter can be written as

$$D(z) = \exp \sum_{m=0}^{M} c(m)\, \tilde{z}^{-m} \tag{1}$$

where $\tilde{z}^{-1}$ is an all-pass transfer function defined by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1. \tag{2}$$

When the sampling frequency is 10kHz, the phase characteristics of the all-pass transfer function with $\alpha = 0.35$ gives a good approximation to the mel frequency scale. For a given the input speech sequence assumed to be zero-mean Gaussian $\boldsymbol{x} = [x(0),\, x(1),\, \ldots,\, x(N-1)]'$, we obtain mel-cepstral coefficients $\boldsymbol{c} = [c(0),\, c(1),\, \ldots,\, c(M)]'$ which maximize $P(\boldsymbol{x} \,|\, \boldsymbol{c})$. It can be shown that $P(\boldsymbol{x} \,|\, \boldsymbol{c})$ is convex with respect to $\boldsymbol{c}$, accordingly the minimization problem can be solved efficiently by an iterative technique described in [11], [7]. It is noted that maximizing $P(\boldsymbol{x} \,|\, \boldsymbol{c})$ with respect to $\boldsymbol{c}$ corresponds to minimizing the energy of the inverse filter output under a constraint that the gain factor of $D(z)$ is unity, i.e., the impulse responses of $D(z)$ and $1/D(z)$ are unity at time 0.

To synthesize speech from the mel-cepstral coefficients, we have to realize the transfer function of (1), which is not a rational function. Fortunately the MLSA filter can approximate $D(z)$ with sufficient accuracy. The MLSA filter is an IIR filter which has a special structure shown in [7], and its stability is guaranteed for speech sounds. The coefficients of the MLSA filter can be obtained from the mel-cepstral coefficients with M multiply-add operations. Thus, by using the MLSA filter, we can synthesize speech easily from the mel-cepstral coefficients.

2.2. Speech Parameter Generation from HMM

Let $\boldsymbol{c}_t$ be the vector of mel-cepstral coefficients at frame t. Then the dynamic features $\Delta \boldsymbol{c}_t$ and $\Delta^2 \boldsymbol{c}_t$, i.e., delta and delta-delta mel-cepstral coefficients at frame t, respectively, were calculated as follows:

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\, \boldsymbol{c}_{t+\tau} \tag{3}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\, \boldsymbol{c}_{t+\tau}. \tag{4}$$

We assume that a speech parameter vector $\boldsymbol{o}_t$ at frame t consists of static and dynamic feature vectors, that is, $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta \boldsymbol{c}_t', \Delta^2 \boldsymbol{c}_t']'$, where $\cdot'$ denotes matrix transpose.

For a given continuous HMM λ and a state sequence $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$, we obtain a sequence of mel-cepstral coefficient vectors $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$ by maximizing $P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$ under constraints (3) and (4). The output distribution of each state is assumed to be a single Gaussian distribution. Thus the logarithm of $P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ can be written as

$$\begin{aligned} \log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda) &= -\frac{1}{2}(\boldsymbol{O} - \boldsymbol{M})' \boldsymbol{U}^{-1} (\boldsymbol{O} - \boldsymbol{M}) \\ &\quad - \frac{1}{2} \log |\boldsymbol{U}| + \textit{Const.} \end{aligned} \tag{5}$$

where

$$\boldsymbol{M} = \left[\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}'\right]' \tag{6}$$

$$\boldsymbol{U} = \mathrm{diag}\left[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}\right], \tag{7}$$

and $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{U}_{q_t}$ are the mean vector and the covariance matrix associated with state q_t, respectively. Without dynamic features (i.e., $\boldsymbol{o}_t = \boldsymbol{c}_t$), it is obvious that $P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)$ is maximized when $\boldsymbol{C} = \boldsymbol{M}$, that is, the sequence of mel-cepstral coefficient vectors is determined by the mean vectors, independently of the covariances $\boldsymbol{U}$.

On the other hand, under the constraints (3) and (4), the sequence of mel-cepstral coefficient vectors $\boldsymbol{C}$ is determined by a set of linear equations $\partial \log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda)/\partial \boldsymbol{C} = \boldsymbol{0}$, which can easily be solved by a fast algorithm derived in [8], [9]. It has been shown that the obtained mel-cepstral coefficient vectors reflect not only the means of static and dynamic feature vectors but also the covariances of those, resulting in a natural-sounding synthetic speech.

3. VERY LOW BIT RATE SPEECH CODER BASED ON HMM

In this section, each part of the speech coding system is described briefly.

3.1. Speech Recognition

We used phonetically balanced 503 sentences uttered by a male speaker MHT in the ATR Japanese speech database for training phoneme HMMs. Speech signals were sampled at 10kHz and windowed by a 25.6ms Hamming window with a 5ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis technique. The feature vectors consisted of 13 mel-cepstral coefficients including the 0th coefficient, and their delta and delta-delta coefficients.

We used 3-state left-to-right triphone models with no skip. Each state was modeled by a single Gaussian distribution with the diagonal covariance. Total of 34 phonemes and a silent models were prepared. Decision-tree based model clustering was applied to each set of triphone models, and the resultant set of tied triphone models has approximately 1,800 distributions.

The speech recognizer of the encoder uses the phoneme pair constraints in Japanese language. The phoneme recognition rate for the test data used in the subjective evaluation (Section 4) was 73.68 % (88.7 % when insertion errors are ignored). The average phoneme rate computed from the transcription data is about 9.5 phoneme/s while the average phoneme rate computed from the recognition results for the test data was 11.7 phoneme/s. It is noted that the test data includes 26 % of silence region.

3.2. Phoneme Index Coding

The phoneme sequence obtained by the phoneme recognizer is transmitted using entropy coding. The histograms of phonemes and phoneme pairs were measured from the phoneme recognition results for the training data. When the Huffman coding based on the occurrence probability distribution of phonemes was used, the bit rate of phoneme information for the test data was about 54 bit/s. Further, using the occurrence probability distribution of phoneme pairs (i.e., phoneme bigram probability) we obtained a bit rate of about 46 bit/s.

3.3. State Duration Coding

For transmitting state durations we examined the following three methods:

Method 1

The histogram of state durations for each phoneme was measured from the phoneme recognition results for the training data. State durations are transmitted by the Huffman coding based on the occurrence probability distribution of state duration for the corresponding phoneme.

Method 2

The histogram of phoneme durations for each phoneme was measured from the phoneme recognition results for the training data. Each phoneme duration is transmitted using the Huffman coding based on the occurrence probability distribution of the corresponding phoneme. In the decoder each phoneme duration is divided into three state durations using state duration densities associated with the corresponding phoneme HMM. The state durations are determined by a method based on the maximum likelihood criterion [8], that is,

$$d_k = m_k + \rho\sigma_k^2 \tag{8}$$

$$\rho = \left(T - \sum_{k=1}^{3} m_k\right) \Big/ \sum_{k=1}^{3} \sigma_k^2 \tag{9}$$

where T is phoneme duration, m_k, σ_k^2 are the mean and variance of the duration density associated with k-th state of the phoneme HMM, respectively. To obtain the state duration densities, the histogram of state durations was measured from the phoneme recognition results for the training data. Each state duration density was modeled by a single Gaussian distribution. Regarding state duration densities of a triphone HMM as a three-dimensional Gaussian, we applied decision-tree based model clustering to the three-dimensional Gaussians, and the resultant set of tied state duration models had approximately 1,600 distributions.

Method 3

State durations of each phoneme are regarded as a three-dimensional vector, and vector-quantized. The codebook is trained by the LBG algorithm based on state durations obtained by phoneme recognition for the training data. Three codebooks whose sizes are 8, 32 and 1024, respectively, were trained for the experiment of Section 4. Further the VQ indexes are transmitted by using the Huffman coding.

3.4. Speech Synthesis

In the decoder, triphone HMMs corresponding to the transmitted phoneme indexes are concatenated, and from the obtained HMM a sequence of mel-cepstral coefficient vectors is generated using the algorithm described in Section 2.2. By exciting the MLSA filter by pulse train or white noise generated according to the pitch information, speech signal is synthesized based on the generated mel-cepstral coefficients.

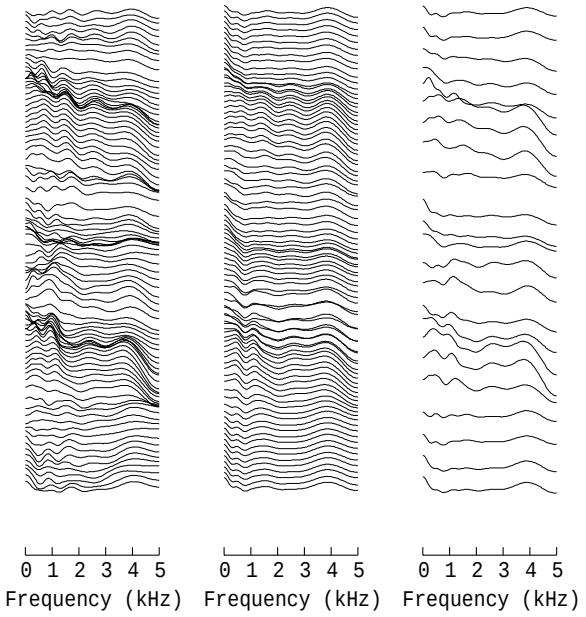

Figure 2: Spectra comparing original (left), proposed 160 bit/s (middle), and vector-quantized 400 bit/s (= 8 bit/frame × 50 frame/s) (right).

4. EXPERIMENTS

In preliminary experiments we observed the following:

1. In the case where both state durations and phoneme durations are not transmitted and the decoder determines state durations of each phoneme based on the state duration densities associated with each phoneme HMM, recognition errors not only have an impact on the subjective quality of the coded speech but degrade the intelligibility significantly.
2. When the unquantized state durations are transmitted, recognition errors do not have an impact on the subjective quality of the coded speech whereas the subjective quality is in proportional to recognition rate.

To evaluate the speech quality of the proposed speech coder, we conducted a DMOS test. Test utterances were eight senteces which are not included in the training data. Subjects were eight males. In this experiment, pitch was not quantized, and original pitch values were used with Viterbi alignment based on phoneme HMMs. Fig. 2 exemplifies spectra for original speech, those reconstructed by the proposed coder, and those vector-quantized. In Fig. 3, DMOS values for the proposed coder were compared to that for the mel-cepstral vocoder with vector quantization of mel-cepstral coefficients. The speech database used for training

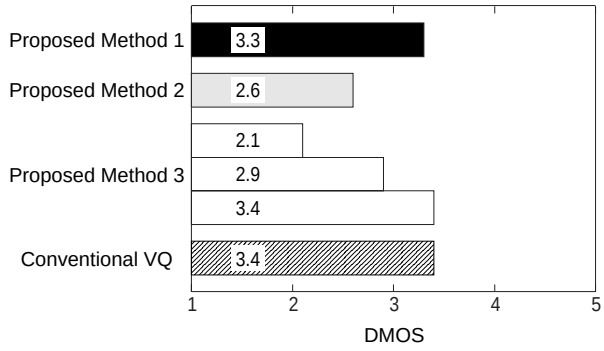

Figure 3: Subjective performance for the proposed and conventional vocoders measured by DMOS.

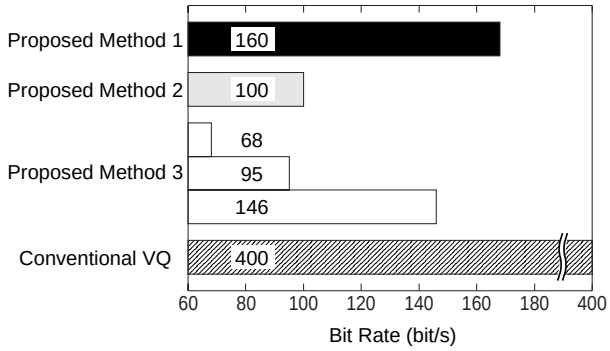

Figure 4: Bit rates for the proposed and conventional coders.

phoneme HMMs was also used for training the VQ codebook for the VQ-based vocoder. The bit rates for both coders are shown in Fig. 4. The proposed coder uses 46 bit/s for transmitting phoneme indexes, and the remaining bits are used for transmitting state or phoneme durations.

Fig. 3 and 4 show that the proposed coder with higher bite rate achieves better performance. This suggests that inaccurate reproduction of state durations degrades the coded speech quality. It can be seen from the figures that the performance of the proposed coder at about 150 bit/s is comparable to that of the VQ-based vocoder at 400 bit/s ($=$ 8 bit/frame $\times$ 50 frame/s) without pitch quantization for both coders. The proposed coder at about 70 bit/s degrades its speech quality while it still preserves the intelligibility of the coded speech.

The coding delay of the proposed coder can be summarized as follows. The delay which arises in the encoder depends on the search strategy of the phoneme recognizer. Generally it could be on the order of 100 ms. On the other hand, the delay corresponding to one phoneme duration; an average of about 100 ms, arises at the decoder since the decoder needs the next phoneme index to choose a triphone HMM. Additionally the speech parameter generation algorithm causes a delay of approximately 100 ms at the decoder.

5. CONCLUSION

We presented a new framework of a very low bit rate speech coder using HMM-based speech recognition and synthesis techniques. The HMM-based speech synthesis consists of a vocoding technique based on mel-cepstrum and an HMM-based speech parameter generation algorithm. It has been shown that the performance of the proposed coder at about 150 bit/s (for the test data including 26 % silence region) is comparable to that of a VQ-based vocoder at 400 bit/s ($=$ 8 bit/frame $\times$ 50 frame/s) in terms of subjective quality measured by DMOS without pitch quantization for both coders. The proposed coder at 70 bit/s can still preserve the intelligibility of the coded speech.

We expect that further improvement of the phoneme recognizer results in better performance of the proposed coder. The future work will be conducted towards constructing a speaker-independent version of the proposed coder using an HMM-based voice conversion technique [12] which adopts a speaker adaptation technique used in HMM-based speech recognition.

REFERENCES

[1] S. Roucos, R. M. Scshwartz and J. Makhoul, "A segment vocoder at 150 b/s," in *Proc. ICASSP-83*, 1983, pp.61–64.

[2] F. K. Soong, "A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis," in *Proc. ICASSP-89*, 1989, pp.584–587.

[3] Y. Shiraki and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 9, pp.1437–1444, Sep. 1989.

[4] Y. Hirata and S. Nakagawa, "A 100bit/s speech coding using a speech recognition technique," in *Proc. EUROSPEECH-89*, 1989, pp.290–293.

[5] C. M. Ribeiro and I. M. Trancoso, "Phonetic vocoding with speaker adaptation," in *Proc. EUROSPEECH-97*, 1997, pp.1291–1294.

[6] M. Ismail and K. Ponting, "Between recognition and synthesis —300 bits/second speech coding," in *Proc. EUROSPEECH-97*, 1997, pp.441–444.

[7] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, 1992, pp.137–140.

[8] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, 1995, pp.660–663.

[9] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. EUROSPEECH-95*, 1995, pp.757–760.

[10] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. ICASSP-96*, 1996, pp.389–392.

[11] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito and S. Imai, "Spectral estimation of speech based on mel-cepstral representation," *Trans. IEICE*, vol. J74-A, pp.1240–1248, Aug. 1991 (in Japanese).

[12] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP-97*, 1997, pp.1611–1614.