A VERY LOW BIT RATE SPEECH CODER USING HMM WITH SPEAKER ADAPTATION

Takashi Masuko†, *Keiichi Tokuda*††, *and Takao Kobayashi*†

†Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 JAPAN
††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN

ABSTRACT

This paper describes a speaker adaptation technique for a phonetic vocoder based on HMM. In the vocoder, the encoder performs phoneme recognition and transmits phoneme indexes and state durations to the decoder, and the decoder synthesizes speech using HMM-based speech synthesis technique. One of the main problems of this vocoder is that the voice characteristics of synthetic speech depend on HMMs used in the decoder, and are therefore fixed regardless of a variety of input speakers. To overcome this problem, we adapt HMMs to input speech by transmitting transfer vectors, information on mismatch between the input speech and HMMs. The results of the subjective tests show that the performance of the proposed vocoder without quantization of transfer vectors is comparable to that of a speaker dependent vocoder.

1. INTRODUCTION

To code speech at rates on the order of 100 bit/s, phonetic or segment vocoders are the most popular techniques [1]-[6]. These coders decompose speech into a sequence of speech units (i.e., phonetic units or acoustically derived segment units), 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. We have also proposed a phonetic vocoder based on HMM (hidden Markov model) [7], in which 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 [8]-[10].

Speaker recognizability of the coded speech is one of the main problems for phonetic vocoders. Since the voice characteristics of coded speech depend on the synthesis units used in the decoder, some types of speaker adaptation is required for speaker independent coding. One possible method is to select the most suitable codebook from multiple speaker dependent codebooks, and another is to adapt the codebook to the input speech [5]. In this paper, we propose a new adaptation technique to realize a speaker independent version of the phonetic vocoder based on HMM which we have already proposed in [7].

In the following, we summarize the HMM-based speech synthesis technique, and describe the speaker adaptation in Section 2 and 3, respectively. The results of subjective evaluation tests are also shown in Section 4.

2. PHONETIC VOCODER BASED ON HMM

A block diagram of the proposed speech coder is illustrated in Fig.1. In this coder, speech spectra are consistently represented by mel-cepstral coefficients obtained by a mel-cepstral analysis technique [11],[12], and the sequence of mel-cepstral coefficient vectors for each speech unit is modeled by phoneme HMM. The encoder performs phoneme recognition which adopts advanced techniques used in the area of speech recognition, and phoneme indexes and state durations transmitted to the decoder by using entropy coding and vector quantization. Pitch information is also transmitted to the decoder. 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 [12]. In the following, we summarize two principal technique used in the coder: a mel-cepstrum-based vocoding and HMM-based speech parameter generation.

2.1. Vocoding Technique Based on Mel-Cepstrum

We model speech spectrum $H(e^{j\omega})$ by the mel-cepstral coefficients $c(m), 0 \leq m \leq M$ [12] as follows:

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

where

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

When the sampling frequency is 10kHz, the phase characteristics $\tilde{\omega}$ of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ with $\alpha = 0.35$ gives a good approximation to the mel frequency scale. Taking the gain factor K outside from $H(z)$, we can rewrite (1) as

$$H(z) = K \cdot D(z). \tag{3}$$

The coefficients $c(m)$ is determined by minimizing

$$\varepsilon = E[e^2(n)] \tag{4}$$

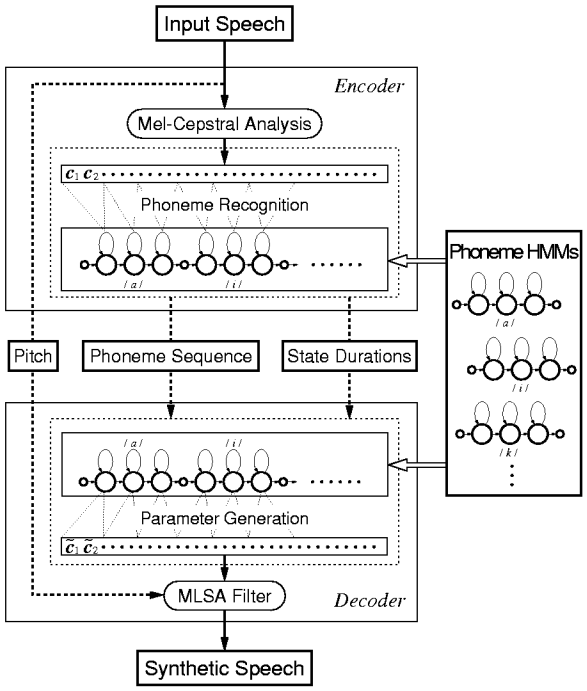

Figure 1: A very low bit rate speech coder based on HMM.

with respect to $c(m)$ where $e(n)$ is the output of the inverse filter $1/D(z)$.

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 approximates $H(z)$ with sufficient accuracy. The MLSA filter is an IIR filter which has a special structure shown in [12], and its stability is guaranteed for speech sounds. Thus, by using the MLSA filter, synthetic speech is obtained directly from the mel-cepstral coefficients.

2.2. Parameter Generation from HMM

Let $\boldsymbol{c}_t$ be the mel-cepstral coefficient vector 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, are calculated as follows:

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\,\boldsymbol{c}_{t+\tau} \tag{5}$$

$$\Delta^2\boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\,\boldsymbol{c}_{t+\tau}. \tag{6}$$

We assume that the speech parameter vector $\boldsymbol{o}_t$ at frame t consists of the static feature vector $\boldsymbol{c}_t$ and the dynamic feature vectors $\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$, 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 the constraints (5) and (6). Here we assume that the output distribution of each state is a single Gaussian distribution. Thus the logarithm of $P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda)$ can be written as

$$\log P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \lambda) = -\frac{1}{2}(\boldsymbol{O}-\boldsymbol{M})'\boldsymbol{U}^{-1}(\boldsymbol{O}-\boldsymbol{M}) - \frac{1}{2}\log|\boldsymbol{U}| + \textit{Const.} \tag{7}$$

where

$$\boldsymbol{M} = \left[\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}'\right]' \tag{8}$$

$$\boldsymbol{U} = \mathrm{diag}\left[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \ldots, \boldsymbol{U}_{q_T}\right], \tag{9}$$

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 speech parameter vector sequence is determined by the mean vectors, independently of the covariances $\boldsymbol{U}$.

On the other hand, under the constraints (5) and (6), 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, and as a result, the synthetic speech is quite smooth and natural sounding.

3. ADAPTATION TO INPUT SPEECH

The HMM-based phonetic vocoder described above has a problem that the voice characteristics of synthetic speech are fixed regardless of a variety of input speakers, because the voice characteristics of synthetic speech depend on HMMs used in the decoder. Moreover, mismatch between training and input speakers may cause much recognition error, resulting degradation of quality of the coded speech which does not occur in the speaker dependent

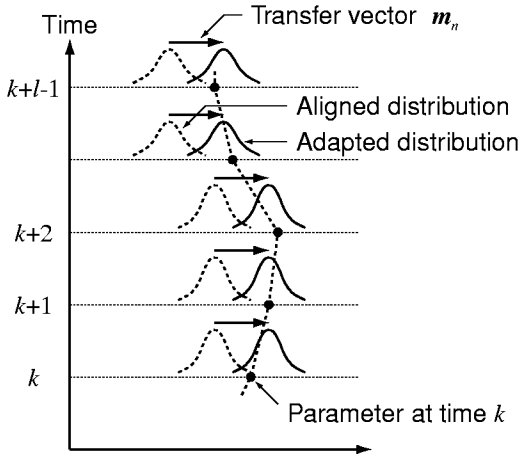

Figure 2: Adaptation of output distributions.

coder. To overcome this problem, we propose a technique for adapting phoneme HMMs used in the decoder by transmitting additional information on input speech.

To adapt HMMs to input speech, we move the output distributions of HMMs by adding a transfer vector to mean vectors so as to fit the distributions to the input parameter vectors (Fig.2).

In the encoder, first, the input parameter sequence $\boldsymbol{O}$ is divided into short segments. The lengths of segments may be fixed or variable. Then, from the input speech parameter sequence and the corresponding state sequence obtained by phoneme recognition, the transfer vector is determined for each segment by maximizing the output probability of the input sequence. Let $\boldsymbol{O}_n = (\boldsymbol{o}_k, \ldots, \boldsymbol{o}_{k+l-1})$ be the input parameter sequence of the n-th segment which starts at time $t = k$ and continues l frames, and $\boldsymbol{Q}_n = (q_k, \ldots, q_{k+l-1})$ be the corresponding state sequence of HMM λ obtained by the Viterbi alignment. Here we assume that the output distribution of each state is a single Gaussian distribution for convenience. Then the transfer vector $\boldsymbol{m}_n$ is determined by maximizing the output probability $P(\boldsymbol{O}_n|\boldsymbol{Q}_n, \boldsymbol{m}_n, \lambda)$ with respect to $\boldsymbol{m}_n$. We can write $\log P(\boldsymbol{O}_n|\boldsymbol{Q}_n, \boldsymbol{m}_n, \lambda)$ in the form

$$\begin{aligned}&\log P(\boldsymbol{O}_n|\boldsymbol{Q}_n, \boldsymbol{m}_n, \lambda)\\ &= -\frac{1}{2}\sum_{t=k}^{k+l-1}\{\boldsymbol{o}_t - (\boldsymbol{\mu}_{q_t} + \boldsymbol{m}_n)\}'\boldsymbol{U}_{q_t}^{-1}\{\boldsymbol{o}_t - (\boldsymbol{\mu}_{q_t} + \boldsymbol{m}_n)\}\\ &\quad + Const.,\end{aligned} \tag{10}$$

where $\boldsymbol{\mu}_{q_t}$ and $\boldsymbol{U}_{q_t}$ are the mean vector and the covariance matrix of the output distribution of the state q_t, respectively. Differentiating $\log P(\boldsymbol{O}_n|\boldsymbol{Q}_n, \boldsymbol{m}_n, \lambda)$ with respect to $\boldsymbol{m}_n$ and setting the result to zero, the transfer vector $\boldsymbol{m}_n$ which maximizes Eq.(10) is obtained as follows:

$$\boldsymbol{m}_n = \left(\sum_{t=k}^{k+l-1}\boldsymbol{U}_{q_t}^{-1}\right)^{-1}\cdot\left(\sum_{t=k}^{k+l-1}\boldsymbol{U}_{q_t}^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{q_t})\right). \tag{11}$$

The transfer vector $\boldsymbol{m}_n$ is quantized and its VQ index is transmitted to the decoder. In the decoder (Fig.3), the decoded transfer vector $\tilde{\boldsymbol{m}}_n$ is added to the mean vectors in the corresponding segment, and a sequence of mel-cepstral coefficient vectors is obtained by the parameter generation algorithm described in 2.2, using

$$\begin{aligned}\tilde{\boldsymbol{M}} = \big[&(\boldsymbol{\mu}_{q_1} + \tilde{\boldsymbol{m}}_1)', \ldots,\\ &\underbrace{(\boldsymbol{\mu}_{q_k} + \tilde{\boldsymbol{m}}_n)', \ldots, (\boldsymbol{\mu}_{q_{k+l-1}} + \tilde{\boldsymbol{m}}_n)'}_{n\text{-th segment}},\\ &\ldots, (\boldsymbol{\mu}_{q_T} + \tilde{\boldsymbol{m}}_N)'\big]'\end{aligned} \tag{12}$$

instead of (8), where N is the number of segments in the sentence.

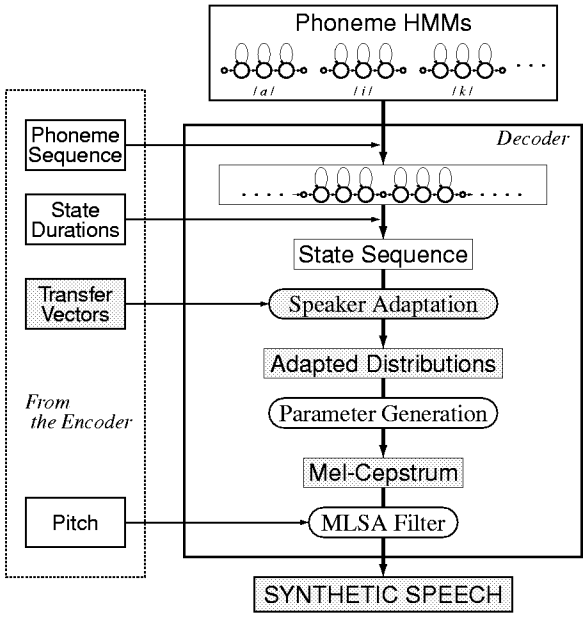

Figure 3: The decoder with speaker adaptation.

4. EXPERIMENTS

To evaluate the speech quality of the proposed speech coder, we conducted DMOS tests.

We trained a set of speaker independent (SI) models using 1,500 sentences in the ATR Japanese speech database uttered by 10 male speakers, and two sets of speaker dependent (SD) models using phonetically balanced 450 sentences uttered by male speakers MHT and MYI to compare with adapted models. It is noted that neither MHT nor MYI were included in the speakers of the training data of SI models. Speech signals were sampled at 10kHz and windowed by a 25.6ms Blackman window with a 5ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis technique. The feature vectors consisted of 16 mel-cepstral coefficients including the 0th coefficient, and their delta and delta-delta coefficients.

We used 5-state left-to-right triphone models with no skip. Each state was modeled by a single Gaussian distribution with the diagonal covariance. Total of 49 phoneme models including a silent model were prepared. Decision-tree based model clustering was applied to each set of triphone models under the same ending conditions, and the resultant sets of tied triphone SI, MHT, and MYI models has 2,213, 2,262 and 1,709 distributions, respectively.

Transfer vectors were obtained for every 100ms (20 frames) and quantized with 10 bits, i.e., 100 bits per a second were used for adaptation of the models. The VQ codebook was trained by the LBG algorithm using the training data of the SI models.

Test utterances were 4 sentences which are not included in the training data. Subjects were 6 males. In the experiments, state durations and pitch were not quantized.

Fig.4 and 5 show the results of subjective tests for speakers MHT and MYI, respectively. From these figures, it can be seen that the proposed adaptation technique significantly improves the performance of the SI coder, and the quality of coded speech from

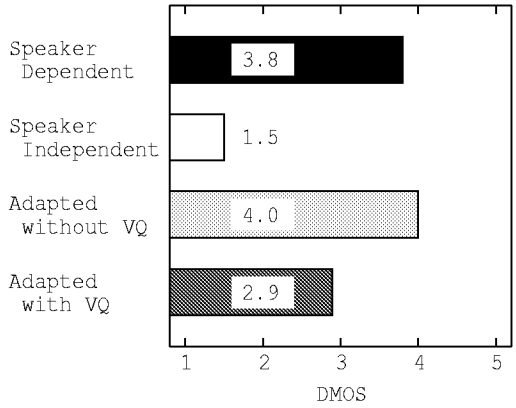

Figure 4: DMOS score for speaker MHT.

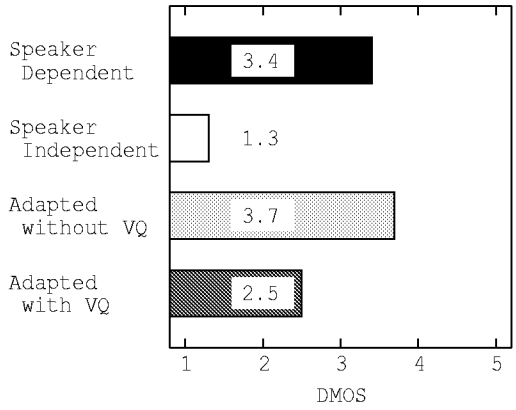

Figure 5: DMOS score for speaker MYI.

adapted models without quantization of transfer vectors was comparable to the SD coders. From informal listening tests, we observed that the adaptation technique improves not only speaker recognizability but also intelligibility.

In the experiments, quantization of transfer vectors degrades the performance of the coder. However, by taking account of correlation between successive transfer vectors, the performance of the coder with quantization will be improved. Furthermore, by using regression matrices instead of transfer vectors, we will be able to improve the performance of the corder.

5. CONCLUSION

In this paper, we have proposed a speaker adaptation technique for a phonetic vocoder based on HMM, and shown that we can improve the performance of the vocoder for speaker independent use. Further research should be concentrated on developing a better adaptation technique to improve speaker recognizability and intelligibility without significant increase of the bit rate.

6. ACKNOWLEDGMENT

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.

7. REFERENCES

1. S. Roucos, R. M. Scshwartz and J. Makhoul, "A segment vocoder at 150 b/s," in *Proc. ICASSP'83*, pp.61–64, 1983.
2. F. K. Soong, "A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis," in *Proc. ICASSP'89*, pp.584–587, 1989.
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. 1988.
4. Y. Hirata and S. Nakagawa, "A 100bit/s speech coding using a speech recognition technique," in *Proc. EUROSPEECH'89*, pp.290–293, 1989.
5. C. M. Ribeiro and I. M. Trancoso, "Phonetic vocoding with speaker adaptation," in *Proc. EUROSPEECH'97*, pp.1291–1294, 1997.
6. M. Ismail and K. Ponting, "Between recognition and synthesis — 300 bits/second speech coding," in *Proc. EUROSPEECH'97*, pp.441–444, 1997.
7. K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi and T. Kitamura, "A very low bit rate speech coder using HMM-based speech recognition/synthesis," in *Proc. ICASSP'98*, pp.609–612, 1998.
8. K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP'95*, pp.660–663, 1995.
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*, pp.757–760, 1995.
10. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP'96*, pp.389–392, 1996.
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. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP'92*, pp.137–140, 1992.