IMPROVING THE PERFORMANCE OF HMM-BASED VERY LOW BIT RATE SPEECH CODING

Takahiro Hoshiya †, *Shinji Sako* †, *Heiga Zen* †, *Keiichi Tokuda* †, *Takashi Masuko* ‡, *Takao Kobayashi* ‡, *and Tadashi Kitamura* †

†Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan
‡Interdisciplinary Graduate School of Science and Engineering Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {hoshiya,sako,zen,tokuda,kitamura}@ics.nitech.ac.jp
{masuko,Takao.Kobayashi}@ip.titech.ac.jp

ABSTRACT

In this paper, we define an F0 quantization scheme for a very low bit rate speech coder based on HMM (Hidden Markov Model). In the coding system, the encoder carries out phoneme recognition, and transmits phoneme indices, state durations and F0 information to the decoder. In the decoder, phoneme HMMs are concatenated according to the phoneme indices, and a sequence of mel-cepstral coefficient vectors is generated from the concatenated HMM. Finally we obtain synthetic speech by using the MLSA (Mel Log Spectrum Approximation) filter according to the mel-cepstral coefficients and F0 information. In addition to the F0 quantization, we investigate encoding methods for other parameters to reduce the bit rate, yet keeping the subjective speech quality. A subjective listening test shows that the performance of the proposed coder at about 100∼150 bit/s is superior to a VQ-based vocoder at 600 bit/s(mel-cepstrum: 6 bit/frame×50 frame/s, F0: 6 bit/frame×50 frame/s).

1. INTRODUCTION

Phonetic and segment vocoders are the most popular techniques to code speech at rates on the order of 100 bit/s [1]-[6]. We also have proposed a phonetic vocoder [7] based on HMM (Hidden Markov Model), in which speech spectra are consistently represented by mel-cepstral coefficients obtained by a mel-cepstral analysis technique [8], 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 indices and state durations to the decoder by using entropy coding and vector quantization. In the decoder, phoneme HMMs are concatenated according to the phoneme indices, 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 output probability of the sequence of mel-cepstral coefficient vectors is maximized for the concatenated HMM and the state sequence [9], [10]. Finally speech signal is synthesized by the MLSA (Mel Log Spectrum Approximation) filter according to the obtained mel-cepstral coefficients [8].

In this paper, we define an F0 quantization scheme for HMM-based speech coding since we did not quantize F0 in the previous work of [7]. The proposed quantization scheme can be viewed as a vector quantization version of MSD-HMM [11] for pitch pattern modeling. In addition, we investigate the encoding scheme for phoneme indices and state durations, and optimal setting of adjustable parameters.

In the following, we describe the HMM-based speech coder in Section 2, the results of a subjective evaluation test in Section 3, and the concluding reworks in the final section.

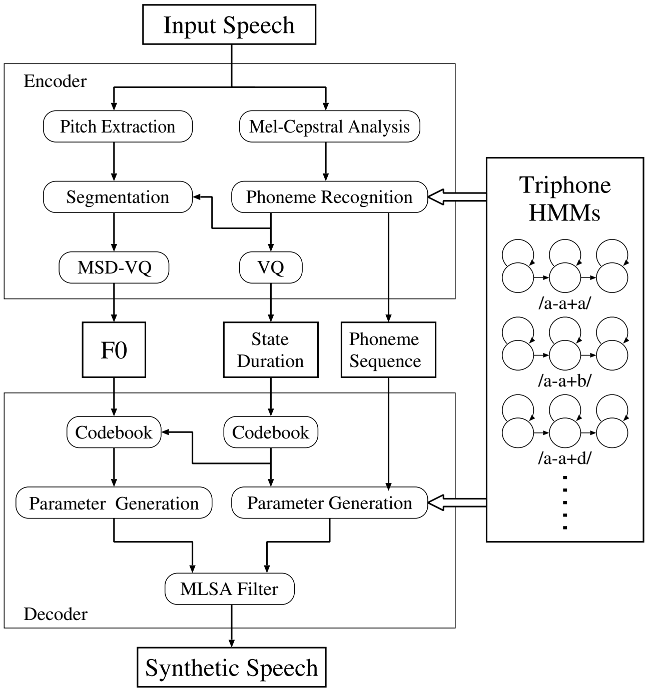

Fig. 1. A very low bit rate speech coder based on HMM.

2. VERY LOW BIT RATE SPEECH CODER BASED ON HMM

Fig. 1 illustrates the very low bit rate speech coder. 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]. In the decoder, triphone HMMs corresponding to the transmitted phoneme indices are concatenated, and from the obtained HMM a sequence of mel-cepstral coefficient vectors is generated using the algorithm described in Section 2.5. F0 values are also transmitted in a manner similar to mel-cepstral coefficients, explained in Section 2.4 and 2.5. By exciting a speech synthesis filter: the MLSA filter by pulse train or white noise generated according to the F0 information, speech signal is synthesized based on the generated mel-cepstral coefficients.

2.1. Mel-Cepstral Analysis

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

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

For a given input speech sequence $\boldsymbol{x} = [x(0),\, x(1),\, \cdots,\, x(N-1)]'$ assumed to be zero-mean Gaussian. we obtain mel-cepstral coefficients $\boldsymbol{c} = [c(0),\, c(1),\, \cdots,\, c(M)]'$ that maximize $P(\boldsymbol{x} \mid \boldsymbol{c})$. It can be shown that $P(\boldsymbol{x} \mid \boldsymbol{c})$ is convex with respect to $\boldsymbol{c}$, accordingly the minimization problem can be solved efficiently by an iterative technique described in [8], [12].

In the previous work [7], we used the sampling frequency of 10 kHz, and accordingly, we set $\alpha = 0.35$ since the all-pass transfer function in (2) with $\alpha = 0.35$ gives a good approximation to the mel-frequency scale when the sampling frequency is 10 kHz. In this work, we use the sampling frequency of 16 kHz because in the phonetic vocoder, the sampling frequency does not affect the bit rate. Accordingly, we use $\alpha = 0.42$ appropriate for 16 kHz sampling. The parameter M also does not affect the bit rate. Thus, we set $M = 25$, whereas the previous coder used $M = 12$.

2.2. Speech Recognition and Phoneme Index Coding

We used phonetically balanced 450 sentences uttered by a male speaker MHT in the ATR Japanese speech database for training phoneme HMMs. Speech signals were sampled at 16 kHz and windowed by a 25 ms Blackman window with a 5 ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis technique. The feature vectors consisted of 26 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. A total of 37 phonemes and 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 phoneme recognition rate for the test data was 82.89 % (95.42 % when insertion errors are ignored). It is noted that the test data includes 28.1 % of silence region.

The speech recognizer of the encoder uses the phoneme pair constraints in Japanese language. The phoneme sequence obtained by the phoneme recognizer is transmitted using entropy coding. In the previous work, we used the phoneme bigram probabilities to design Huffman codes. In this work, the histogram of mora[1] occurrence was measured from the phoneme recognition results for the training data, and the Huffman codes based on the occurrence probability distribution of morae was used.

2.3. State Duration Coding

State durations of each phoneme are regarded as a three-dimensional vector, and vector-quantized. Each phoneme has its own codebook. The codebooks are trained by the LBG algorithm based on state durations obtained by phoneme recognition for the training data. Furthermore, the VQ indices are transmitted by using Huffman coding. The Huffman codes are also designed for each phoneme separately.

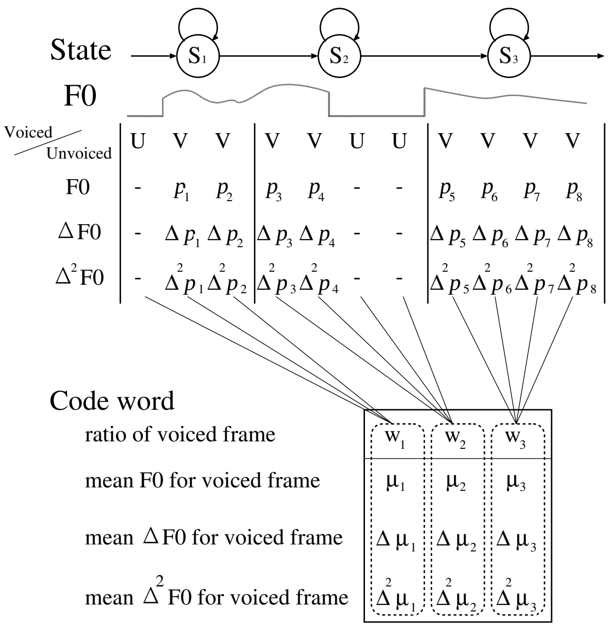
State
S1
S2
S3
F0
Voiced
Unvoiced
U V V | V V U U | V V V V
F0
- p_1 p_2 | p_3 p_4 - - | p_5 p_6 p_7 p_8
ΔF0
- Δp_1 Δp_2 | Δp_3 Δp_4 - - | Δp_5 Δp_6 Δp_7 Δp_8
Δ^2F0
- $\Delta^2 p_1$ $\Delta^2 p_2$ | $\Delta^2 p_3$ $\Delta^2 p_4$ - - | $\Delta^2 p_5$ $\Delta^2 p_6$ $\Delta^2 p_7$ $\Delta^2 p_8$
Code word
ratio of voiced frame
mean F0 for voiced frame
mean ΔF0 for voiced frame
mean Δ^2 F0 for voiced frame
w_1 w_2 w_3
μ_1 μ_2 μ_3
$\Delta\mu_1$ $\Delta\mu_2$ $\Delta\mu_3$
$\Delta^2\mu_1$ $\Delta^2\mu_2$ $\Delta^2\mu_3$

Fig. 2. MSD-VQ for F0 coding

2.4. F0 Coding

The F0 sequence is segmented according to state durations obtained by phoneme recognition. We define MSD-VQ for efficiently coding F0 observations composed of a one-dimensional continuous value in the voiced region and a discreet symbol in the unvoiced region. This can be viewed as a vector-quantization version of MSD-HMM [11]. A feature vector consists of four elements: UV, log F0, Δlog F0, Δ^2log F0. The distortion of F0 observations in each phoneme is measured with a codeword shown in Fig. 2. We define the distortion between observation vector $\boldsymbol{O}$ and codeword $\boldsymbol{\lambda}$ as follows:

$$d(\boldsymbol{O}, \boldsymbol{\lambda}) = \sum_{i=1}^{3} \sum_{t=1}^{T_i} d(\boldsymbol{O}_{it}, \boldsymbol{\lambda}) \tag{3}$$

$$d(\boldsymbol{O}_{it}, \boldsymbol{\lambda}) = \begin{cases} -\log\, w_i + d'(\boldsymbol{O}_{it}, \boldsymbol{\lambda}), & UV_{it} = V \\ -\log\,(1 - w_i), & UV_{it} = U \end{cases} \tag{4}$$

$$\begin{aligned} d'(\boldsymbol{O}_{it}, \boldsymbol{\lambda}) &= (p_{it} - \mu_i)^2 + (\Delta p_{it} - \Delta\mu_i)^2 \\ &\quad + (\Delta^2 p_{it} - \Delta^2\mu_i)^2 \end{aligned} \tag{5}$$

where

$$\boldsymbol{O} = [\boldsymbol{O}_{11}, \cdots, \boldsymbol{O}_{1T_1}, \boldsymbol{O}_{21}, \cdots, \boldsymbol{O}_{2T_2}, \boldsymbol{O}_{31}, \cdots, \boldsymbol{O}_{3T_3}] \tag{6}$$

$$\boldsymbol{O}_{it} = [UV_{it}, \boldsymbol{P}_{it}] \tag{7}$$

$$\boldsymbol{P}_{it} = [p_{it}, \Delta p_{it}, \Delta^2 p_{it}] \tag{8}$$

$$\boldsymbol{\lambda} = [w_1, \mu_1, \Delta\mu_1, \Delta^2\mu_1, \cdots, w_3, \mu_3, \Delta\mu_3, \Delta^2\mu_3] \tag{9}$$

w_i is occurrence probability of voiced frames in the state i, and $\mu_i, \Delta\mu_i, \Delta^2\mu_i$ are the mean of log F0, Δ log F0, Δ^2 log F0 in the voiced region of the state i, respectively.

[1] mora: a syllable-like speech unit in Japanese

The codebook is trained by the LBG-like algorithm based on the distortion measure. We use a unique codebook for every phoneme. Furthermore, the VQ indices are transmitted by using Huffman coding.

2.5. 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, are calculated as follows:

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\, \boldsymbol{c}_{t+\tau} \tag{10}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\, \boldsymbol{c}_{t+\tau}. \tag{11}$$

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 $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_T\}$, we obtain a sequence of mel-cepstral coefficient vectors $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \cdots, \boldsymbol{c}_T']'$ by maximizing $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \cdots, \boldsymbol{o}_T']'$ under constraints (10) and (11). The output distribution of each state is assumed to be a single Gaussian distribution. Thus the logarithm of $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ can be written as

$$\log P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda) = -\frac{1}{2}(\boldsymbol{O} - \boldsymbol{M})' \boldsymbol{U}^{-1} (\boldsymbol{O} - \boldsymbol{M}) - \frac{1}{2} \log |\boldsymbol{U}| + \mathit{Const.} \tag{12}$$

where

$$\boldsymbol{M} = \left[\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \cdots, \boldsymbol{\mu}_{q_T}'\right]' \tag{13}$$

$$\boldsymbol{U} = \mathrm{diag}\left[\boldsymbol{U}_{q_1}, \boldsymbol{U}_{q_2}, \cdots, \boldsymbol{U}_{q_T}\right], \tag{14}$$

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 $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ 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 (10) and (11), the sequence of mel-cepstral coefficient vectors $\boldsymbol{C}$ is determined by a set of linear equations $\partial \log P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)/\partial \boldsymbol{C} = \boldsymbol{0}$, which can easily be solved by a fast algorithm derived in [10], [13]. 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. In this work, not only the sequence of mel-cepstrum but also the sequence of F0 is generated in the same manner. It is noted that each frame is decided as "voiced" when corresponding w_i is greater than 0.5.

2.6. Speech Synthesis Filter Based on Mel-Cepstrum

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 that has a special structure, shown in [8], and its stability is guaranteed for speech sounds.

Table 1. Codebook size

	TEST1	TEST2	TEST3	TEST4
F0	3 bit	3 bit	4 bit	4 bit
State Durations	4 bit	5 bit	4 bit	5 bit

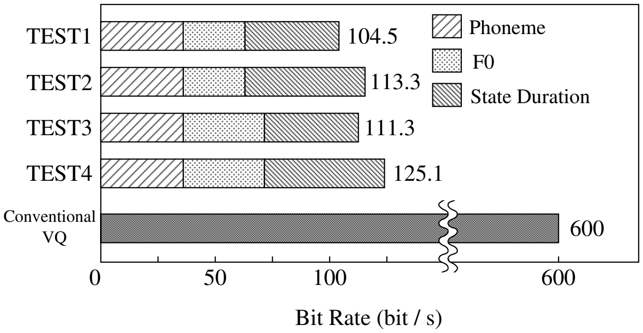

Fig. 3. Bit rates for the proposed and conventional coders.

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.

3. EXPERIMENTS

To evaluate the speech quality, we conducted a MOS test. Test utterances were ten sentences that are not included in the training data. There were twelve subjects who listened to ten sentences that are not included in the training data.

For the test, we prepared four different combinations of codebook sizes for F0 and state duration as shown in Table 1. Note that each codebook size is the maximum number of codewords, and some phonemes use fewer codewords depending on the training data amount. They were compared to conventional VQ: the mel-cepstral vocoder with vector quantization of F0 and mel-cepstral coefficients. The bit rates are shown in Fig. 3. In this regard, conventional VQ does not use Huffman coding. The MOS values are shown in Fig. 4. Fig. 5 exemplifies F0 patterns for (a) original speech, (b) conventional VQ, (c) TEST2, and (d) TEST4.

Fig. 4 and 3 show that the proposed coder achieves better performance than the conventional VQ at 600 bit/s (mel-cepstral: 6 bit/frame×50 frame/s, F0: 6 bit/frame×50 frame/s), and the proposed coder with higher bit rate performs better performance. Fig. 5 shows that the MSD-VQ is effective for quantization for very low bit rate speech coding: while the codebook size is small, F0 pattern is reproduced accurately. It is noted that voiced/unvoiced information is also transmitted accurately by the proposed scheme ((c) TEST2, (d) TEST4).

4. CONCLUSION

We proposed a new F0 coding scheme: MSD-VQ and improved the performance of an HMM-based very low bit rate speech coder. It has been shown that the performance of the proposed coder at about 110 bit/s is superior to that of the VQ-based vocoder at 600 bit/s (mel-cepstral: 6 bit/frame×50 frame/s, F0: 6 bit/frame×50 frame/s) in terms of subjective quality measured by MOS. We as-

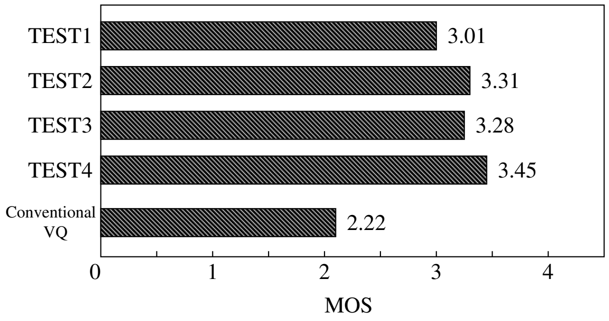

Fig. 4. Subjective performance for the proposed and conventional speech coder measured by MOS.

certained that MSD-VQ was effective for F0 coding in very low bit rate speech coding.

We expect that further improvement of the phoneme recognizer will increase the performance of the proposed coder. In the future, we will implement the mixed excitation model and postfilter [14] to improve synthetic speech quality.

5. REFERENCES

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

[2] F. K. Soong, "A phonetically labeled acoustic segment (PLAS) approach to speech analysis-synthesis," in *Proc. ICASSP*, 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, 1989, pp.1437–1444.

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

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

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

[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 technique," in *Proc. ICASSP*, 1998.

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

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp.2347–2350.

[10] K. Tokuda, Takayoshi Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000.

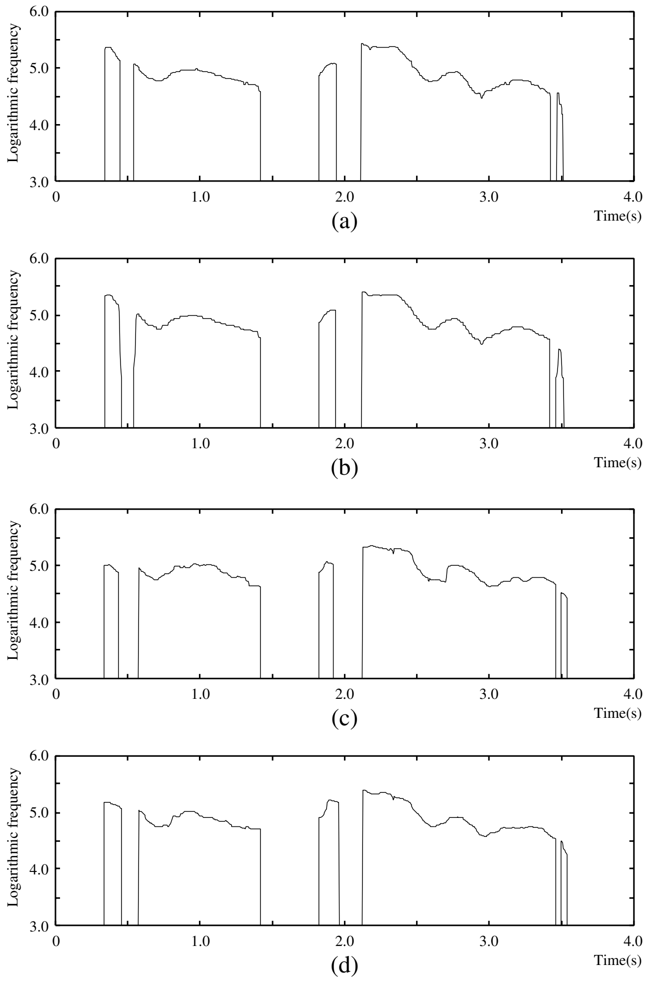

Fig. 5. F0 patterns for (a) original, (d) conventional VQ (300 bit/s), (c) TEST2 (27.0 bit/s), and (d) TEST4 (35.8 bit/s)

[11] K. Tokuda, T. Masuko, N. Mizutani and T. Kobayashi, "Multi-space probability distribution HMM," *Trans. IEICE*, 2002, vol. E85-D, no. 3, pp.455–464.

[12] 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, 1991, pp.1240–1248.(in Japanese).

[13] 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*, 1995, pp.757–760.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis," in *Proc. EUROSPEECH*, 2001, pp.2263–2266.