A DESIGN OF TRANSFORM CODER FOR BOTH SPEECH AND AUDIO SIGNALS AT 1 BIT/SAMPLE

Takehiro Moriya, Naoki Iwakami, Akio Jin, Kazunaga Ikeda and Satoshi Miki

3-9-11 Midori-cho Musashino, Tokyo, 180 Japan.
NTT Human Interface Labs.
e-mail : moriya@splab.hil.ntt.co.jp

ABSTRACT

This paper proposes a speech and audio coder which operates at 1 bit/sample, namely an 8 kbit/s coder for 8 kHz sampling or a 16 kbit/s coder for 16 kHz sampling. The basic structure is inherited from a TwinVQ (Transform domain Weighted Interleave Vector Quantization) high-quality audio coding scheme. Periodical component extraction scheme is newly added to the quantization of MDCT coefficients. This scheme is found to be effective for reducing distortion and improving robustness against channel errors. Qualities for music signals at 8 kbit/s are better than those of G.729 at the same bit rates, while they are worse for clean speech. Qualities at 16 kbit/s are comparable or better than those of G.722 at 48 kbit/s.

1. INTRODUCTION

Toll quality speech coding algorithms are available at the compression ratio of 1 bit/sample. Examples are a new standard at 8 kbit/s (ITU-T G.729) [1] for 8 kHz sample, and a newly proposed transform coder [2] at 16 kbit/s for wideband speech (16 kHz sample). They are useful for two-way telecommunication, since they can reproduce good quality speech signals with low delay and with a reasonably inexpensive fixed point DSP chip. They can, however, provide rather poor quality for complex music signals, since they are mainly designed for speech signals. In one-way multi-media applications, such as the internet, data storage, and digital broadcasting, a general speech and audio coding algorithm is useful even if it needs a longer delay and higher complexity at the encoder.

We propose a transform coder that operates at around 1 bit/sample (8 kbit/s for 8 kHz samples, 16 kbit/s for 16 kHz samples). The algorithm is based on the TwinVQ (Transform domain Weighted INterleave Vector Quantization) [3, 4], which has been originally designed for high-quality audio signals. The major difference is a periodic component extraction scheme.

2. CODING SCHEME

2.1. Basic structure

The structure of the proposed coder (encoder part only) is shown in Fig. 1. The corresponding waveform examples are shown in Fig. 2. The global structure belongs to a transform coder commonly used in audio coding [5, 6]. The input signal is transformed into the frequency domain through MDCT (Modified Discrete Cosine Transform) [7]. Before the MDCT, the input signal is classified into three modes with different transform window sizes (long/medium/short). In the long frame mode, transform size is equal to frame size, while transform operations are carried out twice in a frame with a half transform size in the medium frame mode, and eight times with one eighth transform size in the short frame mode.

The envelope of the MDCT coefficients is represented in a cascaded envelope estimation process by an LPC envelope and a Bark-scale envelope. At first, the amplitude envelope of the MDCT coefficients is approximated by LPC analysis applied to the time domain signal. The predictive coefficients are efficiently quantized through LSP (Line Spectrum Pair) parameters [8]. The MDCT coefficients are globally flattened in the frequency domain by this envelope.

Only in the case of the long frame mode, the periodical components of MDCT coefficients are extracted and vector quantized before Bark-scale envelope normalization. These periodical components are roughly due to the pitch period of the speech or simple musical tone.

Furthermore, the fine structure envelope of the MDCT coefficients is estimated in a frequency band proportional to a Bark-scale. The shapes of the envelope are quantized by interleaved weighted vector quantization. Reconstructed shapes are used for normalization of the MDCT coefficients.

At the final stage, the flattened MDCT coefficients are globally normalized in amplitude and the amplitude is quantized in a log scale. Then they are interleaved, divided into subvectors and vector quantized with a weighted distortion measure derived from the envelope.

At the decoder, output signal is generated by the reverse process at the encoder.

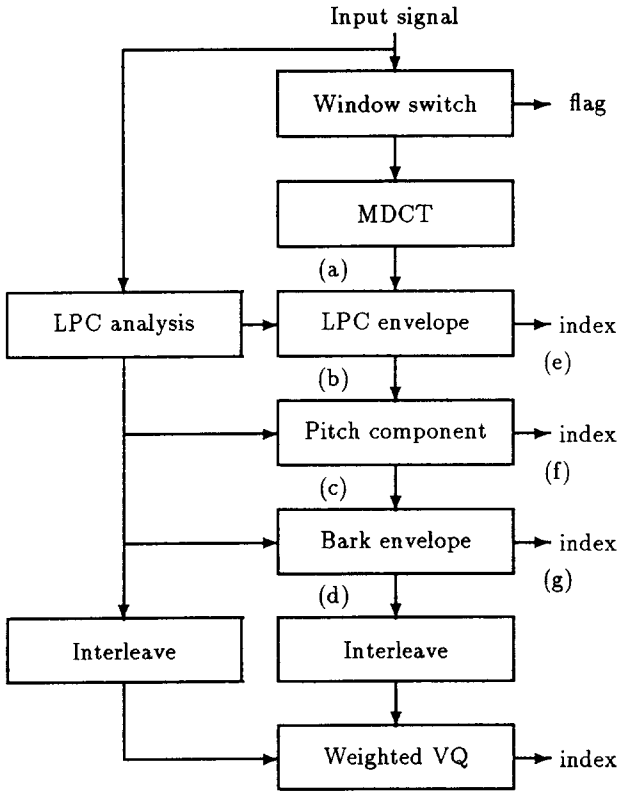

Figure 1. Block diagram of the proposing encoder. From (a) to (g) are corresponding to waveforms in Fig. 2.

2.2. Quantization of periodic components

Periodic component extraction has three steps, namely, the pitch period estimation process, extraction process, and quantization process.

In the pitch period extraction, MDCT coefficients are sampled at every fixed interval. The process searches for the best interval of MDCT coefficients, that is, the interval that gives the largest accumulated power of the extracted MDCT coefficients. Step size of the interval quantization is proportional to the log-scale in the frequency domain.

In the resampling process, a fixed number of components are extracted from the flattened MDCT coefficients from the lower half frequency components. Since the number of extracted coefficients is fixed independent of a pitch frequency, not only coefficients at peak values but also adjacent samples of the peak are extracted in case of higher pitch frequency (longer interval).

In the quantization process, the average amplitude of the packed periodic components is calculated and quantized prior to the shape quantization. The shape of the packed periodic components is quantized by interleaved weighted

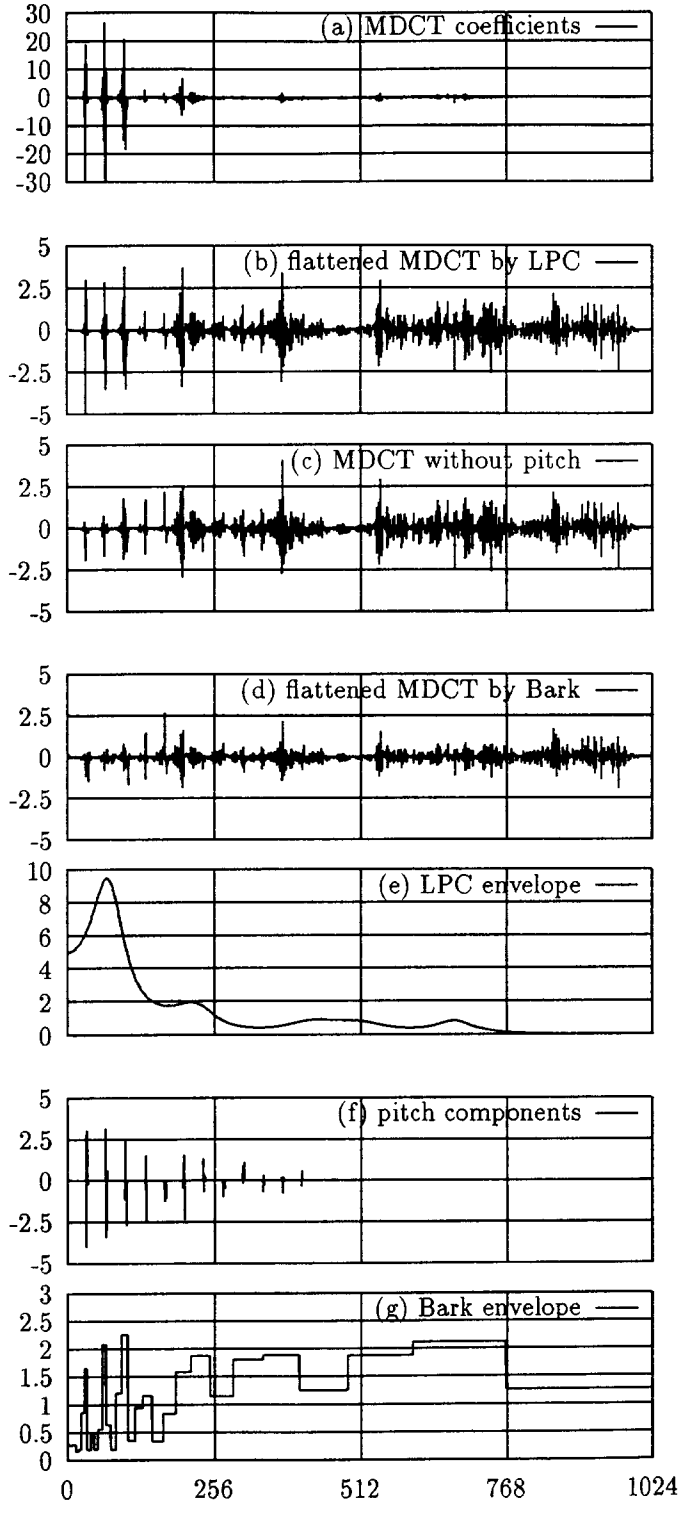

Figure 2. Examples of waveforms for 16 kbit/s at 16 kHz. Horizontal axis shows frequency and vertical axis shows linear amplitude.

vector quantization. This quantization scheme is identical to that used for the flattened MDCT coefficients.

2.3. Typical Implementation

Examples of the bit assignments are listed in Table 1. Vector quantization techniques are used in the LSP parameters, periodic pitch components, Bark-scale envelope, and normalized MDCT coefficients. A multi-stage split VQ with interframe MA (Moving Average) prediction [9] is used for the LSP parameter. For the Bark-scale envelope, an interframe MA prediction VQ is applied to the averaged values. For both the periodic components and the normalized MDCT coefficients, a 2-channel conjugate structure VQ [10] with weighted distortion measure is used. The reconstruction vector is generated by adding two code vectors from 6-bit codebooks each multiplied by one bit polarity, thus total bits become 14 bits as shown in the table 1.

Table 1. Example of bit assignment (l:long, m:medium, s:short)

Bitrate	[kbit/s]	8	16
Sampling	[kHz]	8	16
Frame size	[point]	512	1024
Medium frame size	[point]	256	512
Short frame size	[point]	64	128
Frame gain	[bit/fr]	8	8
LPC order		12	16
LSP quantization	[bit/fr]	15	19
Prediction flag	[bit/fr]	1	1
1st-stage VQ	[bit/fr]	6	6
2nd-stage split VQ	[bit/fr]	3	4
Number of split		3	3
Total bits for pitch component	[bit/fr]	34	65
Pitch frequency	[bit/fr]	8	9
Number of pitch subvectors		2	4
Dimension of pitch subvector		10	15
Quantization bits for pitch subvector	[bit]	14	14
Bark-scale envelope (l/m/s)	[bit/fr]	19/22/48	19/22/48
Number of subvectors (l/m/s)		31/33/29	65/69/65
Dimension of subvectors (l/m/s)		17/16/18	16/15/16
Quantization bits for subvector	[bit]	14	14

3. EVALUATION

3.1. Quality

The newly added process for periodic components can help to reduce quantization distortion and enables the design of a better unequal error protection scheme.

Weighted interleave vector quantization has the largest coding gain, when the averaged envelope values for each subvector are equal [12], since fixed bits are assigned to each subvector. If there are a lot of peak values in the MDCT coefficients, the averaged envelope values may be largely unbalanced, which will increase the distortion. Periodic component extraction can avoid such inbalance and is effective especially for input signals with rich tonal components. Due to the pitch component extraction, quantization distortion has been reduced by 2 dB for some input signals such as "harpsichord", while the distortion remain unchanged for other input signals.

According to an informal listening test, this scheme can provide reasonably good quality for a wide range of input signals, including speech and audio. The proposed scheme for 8 kbit/s can provide better quality than G.729 for general music signals and speech with background music, while G.729 is better for clean speech signals.

The qualities obtained by the 16 kbit/s coder are comparable to those obtained by G.722 at 48 kbit/s. For some music signals, the qualities of the proposed scheme is significantly better than that of G.722 at 48 kbit/s, while it is comparable or slightly worse for clean speech signals.

3.2. Unequal error protection

For applications with channel errors, unequal error protection is effective for waveform recovery [11]. Transmission data streams of TwinVQ are very convenient for designing an unequal error protection scheme, since they consist of main information and side information in fixed number of bits per frame. The main information includes a set of indices of weighted VQ for flattened MDCT coefficients. The side information includes codes for power, LSP parameter, window switch, pitch frequency, indices of VQ for Bark-scale envelope and pitch components. Bitrate of the side information is from 10 to 15% of the total and most of it is sensitive to channel errors while the main information is insensitive to errors.

The introduction of pitch component extraction has made the unequal error protection scheme more feasible, because extracting and protecting the larger values reduces the error sensitivities of the main information.

It should be noted that this scheme is inherently robust against channel errors, because it uses neither adaptive bit allocation, nor variable length coding, which are frequently used in traditional audio coders. In addition, it is robust against frame erasure, since all the interframe predictions

in the quantization are based on moving average and only transmitted code in the previous frame is used. This means that even if a whole frame information is lost, the complete waveform is recovered only one frame after the erasure. This scheme, therefore, can be applied to even mobile communication channels with heavily bursty noise.

3.3. Operation on bitstream

As already described, the coding scheme has a simple and fixed bitstream structure. Therefore it is easy to modify the bitstream. Since output signals are only dependent on three consecutive frame data including MDCT overlap, a part of frame data picked from the whole bitstream can generate a reasonable sound with quick playback and reverse playback.

It is also easy to control the quality by means of intentional scrambling process on the main data, and to include water marking information to the main data for the purpose of the protection the copyright of the music contents. These feature may be very useful in the multimedia communication environment.

3.4. Delay and complexity

Disadvantages of the coder are the delay and the encoder complexity. As you can see from the table 1, when the frame size is 64 ms arithmetic delay is 128 ms due to MDCT windowing. Vector quantization needs a large amount of computation for codebook searching, even if it uses suboptimal searches in the conjugate structure codebook. It is noted that the complexity of the decoder is reasonably small, which is more important than the encoder complexity.

The proposed scheme does not meet the requirement of the current ITU-T wideband speech/audio coding standard because of the above two drawbacks. However it may be useful for the MPEG standard or general multi-media applications.

4. CONCLUSION

We proposed a transform coding scheme that operates at around 1 bit/sample, namely, 8 kbit/s for 8 kHz sample and 16 kbit/s for 16 kHz sample. Qualities for music signals at 8 kbit/s were better than those of G.729 at the same bit rates, while they were worse for clean speech. Qualities at 16 kbit/s were comparable to those of G.722 at 48 kbit/s, but were significantly better for some music signals. Due to a newly introduced pitch component extraction scheme, quantization distortion is reduced and an error sensitivities of the main data stream become reduced. This helps to design more effective unequal error protection schemes and to design various modification schems of the main data stream for various application, such as fast playback, copyright protection. Although this coder needs higher delay and higher complexity of the encoder than the ITU-T speech coding standards require, it is useful for one-way multi-media applications, such as the internet, data storage, and broadcasting.

5. ACKNOWLEDGEMENT

The authors would like to express thanks to Dr. Nobuhiko Kitawaki and Takao Kaneko for their research guidance.

REFERENCES

[1] R. Salami, et.al., "Description of the proposed ITU-T 8 kbit/s Speech Coding Standard," Proc. IEEE Speech coding workshop, pp.3–4, 1995.

[2] J. H. Chen and D. Wang: "Transform Predictive Coding of Wide band Speech Signals," *Proc. ICASSP'96*, pp. 275–278, 1996.

[3] N. Iwakami, T. Moriya and S. Miki: "High-quality Audio Coding at less than 64 kbit/s by Using TwinVQ," *Proc. ICASSP'95*, pp. 937–940, 1995.

[4] T. Moriya, N. Iwakami, K. Ikeda and S. Miki: "Extension and Complexity Reduction of TwinVQ Audio Coder ," *Proc. ICASSP'96*, pp. 1029–1032, 1996.

[5] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio," J. Audio Eng. Soc., vol. 42, No. 10, pp. 780–792, 1994.

[6] D. Pan: "A Tutorial on MPEG/Audio Compression," IEEE Multimedia magazine, vol. 2, No. 2, pp. 60–74, 1995.

[7] J. Princen, A. Johnson and A. Bradley, "Adapative Transform Coding Incorporating Time Domain Aliasing Cancellation," Speech Commun., vol. 6, pp. 299–308, 1987.

[8] F. Itakura, T. Kobayashi and M. Honda, "A Hardware Implementation of a New Narrow to Medium Band Speech Coding," Proc. ICASSP'82, pp. 1964–1967, 1982.

[9] A. Kataoka, T. Moriya and S. Hayashi, "An 8 kbit/s Speech Coder Based on Conjugate Structure CELP," Proc. ICASSP'93, pp. II-592–595. 1993.

[10] T. Moriya, "Two-channel Conjugate Vector Quantizer for Noisy Channel Speech Coding," IEEE JSAC, vol. 10 pp. 866–874. 1992.

[11] K. Ikeda, T. Moriya N. Iwakami and S. Miki: "Error Protected TwinVQ Audio Coding at less than 64 kbit/s," *Proc. IEEE Speech Coding Workshop*, pp. 33–34, 1995.

[12] T. Moriya and M. Honda, "Transform Coding of Speech Using a Weighted Vector Quantizer," IEEE JSAC, vol. 6, pp. 425–431, 1988,