Proc. IEEE Workshop Speech Coding (Pocono Manor, PA), pp. 109-110, Sept. 1997

PERCEPTUAL CODING OF NARROWBAND AUDIO SIGNALS AT 8 KBIT/S*

Hossein Najafzadeh-Azghandi and Peter Kabal

Department of Electrical Engineering
McGill University
Montreal, Quebec H3A 2A7
E-mail: kabal@ee.mcgill.ca

ABSTRACT

This paper proposes a VQ-based transform coding scheme for audio signals (sampled at 8 kHz) at very low bit rates. This coder uses a new perceptually based distortion measure, which takes into account the energy of audible noise, in both training the codebooks and selecting the best codewords. An adaptive bit allocation strategy based on the distribution of the energy of the transform coefficients above the masking threshold is employed to assign more bits to perceptually important critical bands. This coder delivers good quality for most audio signals at 1 bit/sample.

1. INTRODUCTION

Users of personal communication systems require high quality reproduction of all signals that can be presented to a common carrier. Conventional speech coders achieve compression by utilizing models of speech production. These models are not necessarily appropriate for other signals such as noisy speech, multiple speakers, music and background noise. For instance for many speech coders, music-on-hold is coded with annoying artifacts. On the other hand, a perceptually based coder can accommodate different signals. By employing perceptual coding criteria, based on the human hearing system, such coders can handle a wide variety of input signals. Although, models of the human hearing system have been successfully used in high quality broadband audio coding, in low rate coding where the transmission rate should be kept low, masking effects are not fully exploited.

In this paper, we propose a low rate coder that will reproduce most of the audio signals of 8 kHz sampling rate with good quality using 1 bit/sample. The efficacy of this scheme comes from the definition of the *distortion measure* which is used in training the codebooks and selecting the best codewords as well as a perceptually based bit allocation to different critical bands. This perceptually based VQ scheme improves the coder subjective performance by taking into account perceptual redundancies in the transform coefficients.

*This work was supported by the Canadian Institute for Telecommunication Research.

2. BASIC CODER AND DECODER STRUCTURE

Fig. 1 illustrates the major processing blocks for the proposed coder and decoder. In this scheme, the input signal is partitioned into overlapping frames which are transformed into the frequency domain by means of a Modified Discrete Cosine Transform (MDCT) [1]. There is a 50% overlap between successive frames which implies a higher frequency resolution while maintaining critical sampling. This overlap also has the advantage of reducing block edge effects which exist in traditional transform coding systems. For blocks with no sharp transients, a frame of 240 samples is transformed into 120 coefficients each representing a bandwidth of 33.33 Hz at an 8 kHz sampling rate; it provides the good spectral resolution that is required for the calculation of masking thresholds. However, this good frequency resolution is achieved at the expense of temporal resolution which is needed for reproducing transients. Since quantization error spreading over a window length can produce pre-echo artifacts, a shorter window with a length of 10 msec is used whenever a strong transient is detected. The window switching is done based on the criterion proposed in [2] . Because the switch-over is not instantaneous, a start window is used to switch from long to short windows, and a stop window switches back [3].

2.1. Calculation of the Masking Threshold

Transform coefficients are used to calculate the energy in each critical band. Then the model proposed in [4], which accounts for perceptually important characteristics of the input signal, is used to determine the masking threshold for the critical bands.

2.2. Perceptually Trained VQ

The masking thresholds for different critical bands are used to vector quantize the transform coefficients. There are 18 critical bands for the bandwidth of 4 kHz. For each critical band, a structured codebook with 2048 codewords is designed. To form the training sets, the vectors in each training set are normalized by the corresponding energy above the masking threshold. (If the energy above the masking threshold for any band is zero, that subvector will not be included in the training set.) In the process of training the codebooks, we employ the following distortion measure

based on the audible noise energy. For a vector of N normalized transform coefficients X and the jth codeword $C^{(j)}$

$$D(i) \triangleq |X(i) - C^{(j)}(i)|^2 - M(i) \quad (1)$$

where M is the vector of normalized masking thresholds corresponding to X and $i = 1, \ldots, N$.

The energy of the audible noise is calculated by

$$E(X, C^{(j)}) = \sum_{i=1}^{N} (|D(i)| + D(i))/2 \quad (2)$$

The centroid of each Voronoi region is determined by minimizing the energy of the audible noise as follows

$$C^{(j)}_{\text{opt}} = \arg\min_{C^{(j)}} \sum_{k=1}^{L} E(X^{(k)}, C^{(j)}) \quad (3)$$

where L is the number of the vectors in region j.

2.3. Bit Allocation Scheme

The number of bits assigned to each critical band is determined in an iterative manner using the following formula

$$b_i = \bar{b} + \alpha_i \log_2(\hat{E}_i / \hat{E}_{gm}) \quad (4)$$

where $\bar{b}$ is chosen to be 5 bits, $\hat{E}_i$ is the quantized energy above the masking threshold for critical band i , $\hat{E}_{gm}$ is the geometric mean of $\hat{E}_i$'s and α_i is a constant between 0.3 to 0.5. The maximum number of bits assigned to each band is set to 11 bits.

2.4. Predictive VQ of E_i's

Since the transform coefficients in each critical band are normalized by the corresponding energy above the masking threshold E_i, the vector containing E_i's must be transmitted to the receiver as side information. Due to the 50% overlap between successive frames, the spectra and therefore the masking pattern for those blocks are highly correlated. We exploit this fact to vector quantize E_i's by using a predictive scheme.

3. SUBJECTIVE TEST RESULTS

Test audio files including a piece of female speech, symphony orchestra, noisy speech and multiple speakers were coded by the proposed coder and by two other low rate speech coders, i.e., ITU–T G.729 and EIA/TIA IS–96 at 8 kbit/sec. The subjective quality of the coded signal by the proposed coder was clearly better for almost all audio files. The exceptions were for files containing a signgle speaker. Even for these case, the quality is not far below that of the coders specifically designed for speech input.

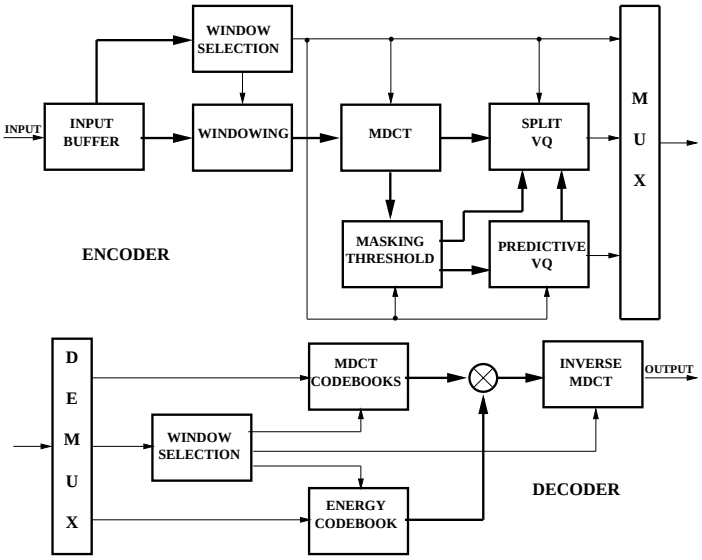

Figure 1: Block diagram of the coder and decoder.

4. CONCLUSION

In this paper, we have presented a transform coder suited for a wide range of natural sounds such as speech, multiple speaker, noisy speech and music at 8 kbit/s. A perceptually based split–VQ scheme with subbands matching the critical bands of the auditory system has been employed to remove both statistical and perceptual redundancies in the input signal to a large extent. We believe that, this work has revealed the suitability of VQ–based perceptual coding systems at a rate of as low as 8 kbit/s. This coder delivers acceptable quality for most audio signals while other state-of-the-art speech coders operating at the same rate have uneven results for non-speech signals.

5. REFERENCES

[1] J. P. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time-domain aliasing cancellation," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, pp. 1153–1161, Oct. 1986.

[2] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.

[3] K. Brandenburg, G. Stoll, Y. Dehery, J. D. Johnston, L. V. Kerkhof, and E. F. Schroeder, "The ISO/MPEG audio codec: A generic standard for coding of high quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–791, Oct. 1994.

[4] J. D. Johnston, "Transform coding of audio signals using the perceptual noise criteria," *IEEE-SAC*, vol. 6, pp. 314–323, Feb. 1988.