IMPROVING PERCEPTUAL CODING OF NARROWBAND AUDIO SIGNALS AT LOW RATES

Hossein Najafzadeh-Azghandi and Peter Kabal

Electrical & Computer Engineering
McGill University, Montreal, Canada

Abstract

This paper discusses perceptual coding of narrowband audio signals at low rates. In particular, it proposes a new error measure which shapes the noise inside the critical bands, a window switching criterion based on the temporal masking effect of the hearing system, a more accurate model of the simultaneous masking effect of the hearing system, perceptually-based bit allocation algorithms based on two different approaches towards quantization noise shaping and a predictive vector quantization scheme to code the scale factors. The resulting coding scheme outperforms existing low rate speech coders for non-speech signals.

1 Introduction

The goal is to represent digital audio signals at low bit rates with minimum perceived loss of signal quality. At low bit rates, it is necessary that we design the coding algorithm to minimize a perceptually based measure of signal distortion, rather than more traditional criterion such as the mean square difference between the waveform at the input and output of the coding system. By doing so, the distortion (or noise) introduced in the coding process is distributed in such a way that it will be masked by the input signal. In low rate coding, much of the present emphasis has been on coding speech signals. However, such coders perform badly for non-speech signals such as music. Moreover, a new trend of low rate coding of narrowband audio signals is emerging [1]. Such coders would be needed in multimedia communications over wireless links (personal communications systems) and over wired links (such as the Internet). Following this new trend, we discuss different aspects of low rate coders and propose criteria and algorithms suitable for these coders in general and more specifically to enhance the low rate coder proposed in [2]. The proposed coder has been designed based on a model of the human hearing system and codes narrowband audio signals (sampled at 8 kHz) at 1 bit/sample.

2 Coder Description

In the proposed coder shown in Fig. 1, a frame of 240 samples of the input signal is transformed into the frequency domain by means of a Modified Discrete Cosine Transform (MDCT). It is desired to localize short burst of quantization noise to prevent it spreading over a long period of time.

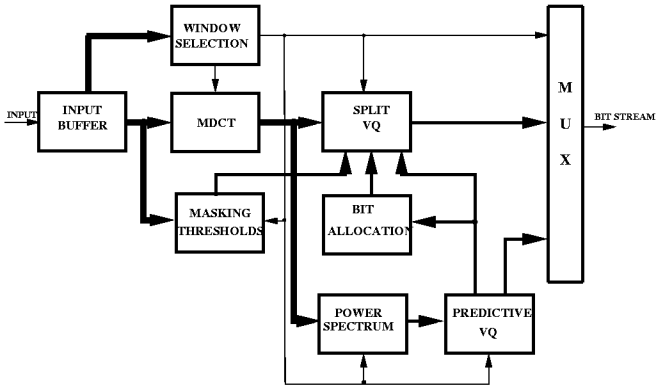

Fig. 1 Block diagram of the coder.

One way to handle the problem is to switch to a shorter window when facing a strong jump in the energy. However, short windows reduce the coding gain and should be avoided when they do not improve the coded signal quality. Since backward temporal masking lasts for about 4 msec while forward temporal masking lasts for about 200 msec, we have to make a distinction between rises and falls in the energy of the signal. A simple criterion which works based on the relative positive change in the energy of the input signal is used. In the time domain, a local estimate is made of the change in signal energy. This is done by splitting the input frame into 80 intervals and calculating the energy of the coefficients in each interval. The maximum of positive changes will be found as follows

$$r = \max\left(\frac{e_{j+1} - e_j}{e_j}\right), \quad j = 0, ..., 79 \tag{1}$$

where e_j is the energy of interval j. If r exceeds a threshold value, then the switching to a shorter widow will be done through a start window. Note that in order to maintain perfect reconstruction of the combined analysis and synthesis stages, two transitional windows, i.e. start and stop windows are needed.

2.1 Calculation of the Masking Threshold

In calculating the masking threshold, we use a modified version of the model proposed in [3] . The masking calculation consists the following steps:

- Calculation of the Bark energy spectrum which is the total energy in each Bark (in the DFT domain)
- Convolution of the Bark energy spectrum with the spreading function to get the spread Bark spectrum
- In contrast to [3] in which the spectral flatness measure is used to identify the nature of the whole frame as being tone-like or noise-like, we take another approach based on the predictability of the transform coefficients in each critical band. Note that, most audio signals have a noise-like structure at high frequencies despite the fact that they may have a strong harmonic structure at low frequencies. Considering this fact, it would be more accurate to identify the nature of the spectrum locally at different Barks. The tonality factor will be calculated for each Bark using

$$\tilde{X}^{(j)} = 2X^{(j-1)} - X^{(j-2)} \tag{2}$$

where $\tilde{X}^{(j)}$ is a linear prediction of the current subvector based on the observation of previous subvectors $X^{(j-1)}$ and $X^{(j-2)}$. The relative prediction error is calculated

$$e = \frac{\| X^j - \tilde{X}^{(j)} \|}{\| X^j \| + \| \tilde{X}^j \|} \tag{3}$$

The relative prediction error will be converted to the tonality factor according to [4]

$$a = \min(1, \max(-0.3 - 0.43\log(e), 0)) \tag{4}$$

- Subtraction of an offset from the log spread Bark spectrum to get the masking threshold

$$\text{offset} = a(14.5 + i) + 5.5(1 - a) \tag{5}$$

where i is the index of the critical band.

- Comparison of the masking threshold with the absolute threshold of hearing
- Since the masking threshold is calculated based on the DFT of the input frame, it is not accurate to use this masking threshold for the MDCT coefficients. Instead, we consider the following relationship between the DFT and MDCT to find a more accurate masking threshold for MDCT coefficients (see Appendix A).

$$C(k) = \sqrt{2/M}|S(k)|\cos[\frac{2\pi n_0(k+0.5)}{N} - \angle S(k)] \tag{6}$$

where $S(k)$ is the Fourier transform of the modulated windowed input signal, $C(k)$ is the MDCT, $n_0 = (M+1)/2$, M and N are the number of samples in the frequency and time domain respectively. If m_{DFT} is the masking threshold corresponding to the kth DFT coefficient, then in order to have the same Signal-to-Mask Ratio (SMR) at any coefficient in the DFT and MDCT domain, the following relation should hold:

$$C^2(k)/m_{\text{MDCT}} = |S(k)|^2/m_{\text{DFT}} \tag{7}$$

Considering the relation between the MDCT and DFT, we find the masking threshold for the kth MDCT coefficient,

$$m_{\text{MDCT}} = (2/M)m_{\text{DFT}}\cos^2(\frac{2\pi n_0(k+0.5)}{N} - \angle S(k)) \tag{8}$$

2.2 Quantization of the Transform Coefficients

One way to accomplish good quantization is to consider the characteristics of the hearing system such as masking phenomena and limited temporal and frequency resolution. Due to dearth of bits for coding the transform coefficients, vector quantization is used rather than scaler quantization. In [2], a perceptually-based vector quantization scheme has been proposed. In that scheme, each frame of transform coefficients is divided into the critical bands of the hearing system. In order to reduce the complexity of vector quantization, the subvectors are decomposed into gains and shapes which are coded separately. The distortion measure used in [2] counts only the audible part of the quantization noise, meaning that as long as the quantization noise is below the masking threshold, the coded signal will be perceptually identical to the original one.
With a limited number of bits, it is not possible to have transparent coding, but combining the masking phenomenon with an adaptive VQ scheme is an appropriate way to keep the quantization noise as close to the masking threshold as possible. Since the quantization noise level usually goes above the masking threshold, it will be appropriate to shape the quantization noise inside each band too. Therefore, we modify the error criterion proposed in [2] as follows

$$e(i) = \max(\frac{|X(i) - C^{(j)}(i)|^2 - M(i)}{X^2(i) + M(i)}, 0) \tag{9}$$

$$D(X, C^{(j)}) = \sum_{i=1}^{L} e(i) \tag{10}$$

where D is the total quantization noise above the masking threshold, $X(i)$ is the transform coefficient, $C^{(j)}$ is the j-th codevector, $M(i)$ is the corresponding masking threshold and L is the dimension of X. By making this modification, we allow the audible quantization noise to get shaped according to the distribution of energy inside a critical band.

2.3 Predictive VQ of the Scale Factors

In the process of the vector quantization of the transform coefficients, in order to reduce the dynamic range of the input vectors, they are normalized to unit energy vectors. In the proposed coder, a predictive/non-predictive VQ scheme is used in the log domain to quantize the scale factors. Since the level of similarity between successive vectors containing the scale factors are varying according to the nature of the signal, we use the predictive scheme whenever the

root mean squared difference of the current vector and its prediction is less then 4 dB, otherwise the vector of scale factors will be quantized directly. This coding strategy is compatible with the mechanism of the hearing system; in steady parts of the input signal such as voiced speech we need finer quantization of both spectral shapes and gains, whereas for 'unstructured' or noise-like parts more coarse quantization is adequate. This also can be justified through the masking property of the hearing system. As is well known, the masking threshold in the case of tone-masking-noise is lower than that of noise-masking-noise. For that reason, we need finer quantization for pseudo-periodic parts of the input signal. In our scheme, we quantize the vectors containing the scale factors through the following steps:

- Take the logarithm of the vector of scale factors
- Subtract the average and quantize the average using 5 bits
- Predict the current normalized vector from the previous normalized vector using the best prediction matrix out of 64 possible matrices
- If the root mean squared difference of the current vector and its prediction is less than 4 dB, the difference vector will be quantized by two stage VQ otherwise the input vector will be directly quantized.

This approach leads to finer quantization of the scale factors in steady state parts of the input signal which is highly desired for high quality of the coded signal.

2.4 Bit Allocation

In low rate coding of audio signals, due to scarcity of bits, the existence of unmasked quantization noise (audible noise) is inevitable. The final goal in low rate coding is to deliver acceptable quality with no annoying artifacts. This contrasts with transparent coding which is required in high rate wideband audio coding. Two different strategies can be considered to shape the audible noise spectrum[5]. In one approach, the quantization noise will be shaped in parallel with the masking threshold. This way, the audible noise will become equally audible in different frequency bands. An alternative approach is to follow the above noise-shaping scheme only for noise levels up to the masking threshold and have a flat noise spectrum above the masking threshold. The audible noise would then first fill the spectral valleys of the masking threshold and therefore would have a higher level difference to the masking threshold. Based on our experiments, it is hard to draw a solid conclusion on which of the two approaches is better.

Energy-based Bit Allocation

In this scheme, the energy above masking threshold for each critical band is calculated and transmitted as side information. Since in our coding scheme, the transform coefficients in each critical band are vector quantized, we have found the rate-distortion curve for each codebook. Note that the distortion is considered as only the audible part of the quantization noise, i.e. the noise above the masking threshold. Based on the rate-distortion curves for each band, we have found the following experimental formula relating the audible noise to the energy above the masking threshold as a function of the number of bits assigned to that specific band:

$$D_i = c_i \, \hat{E}_i \, 2^{-b_i/\beta_i} \tag{11}$$

where D_i is the quantization noise above the masking threshold, $\hat{E}_i$ is the quantized energy above the masking threshold, c_i and β_i are constant found from the corresponding rate-distortion curve. Having the above relation for each band, we go through the following minimization procedure to find the number of bits assigned to each band;

$$\arg\min_{b_i} \sum_{i=1}^{17} D_i \quad \text{s.t.} \quad \sum_{i=1}^{17} b_i = b_T \tag{12}$$

b_T is the total number of bits available for each frame.

$$b_i = \max\Big(\frac{\beta_i b_T}{\sum_{j=1}^{17} \beta_j} + \log_2\big(\frac{\hat{E}_i c_i}{\hat{E}_{gm}}\big), 0\Big) \tag{13}$$

where

$$E_{gm} = \left(\prod_{i=1}^{17} (c_i \hat{E}_i)^{\beta_i}\right)^{\left(1/\sum_{i=1}^{17} \beta_i\right)} \tag{14}$$

Then decimal parts of the b_i's will be discarded to leave the integer parts. The remaining bits will be distributed based on a greedy algorithm, meaning that one bit at a time to the band with highest distortion. In this approach, the level of audible noise will be relatively higher in spectrum valleys due to the fact that there is less energy above masking threshold compared to unmasked spectral peaks.

Signal-to-Mask Ratio (SMR)-based Bit Allocation

In this approach bit allocation is performed based on the Signal-to-Mask Ratio (SMR). This way, the resulting noise spectrum will be parallel to the masking threshold. Each critical band is considered as a single entity with its corresponding SMR. In doing so, for each band the total amount of energy is calculated and transmitted. Since in our coding scheme, the subbands are matched to the critical bands, we can calculate the masking threshold at the receiver should the tonality factor (which indicates how the spectrum looks like, i.e., either tone-like or noise-like) be transmitted as well. In order to save bits, we can fix the value of the tonality factor at 0.5 which assumes the signal is neither pure tone nor noise. As mentioned before, we use a shorter widow when a strong transient is detected. In that case, since the signal does not have any periodic structure, we simply set the masking threshold 5 dB below the spread Bark spectrum. SMR is indicative of the ratio of signal energy to quantization noise in the case of transparent coding,

i.e., noise level below the masking threshold. SMR for each band is calculated in the following manner:

$$\mathrm{SMR}_i = \hat{E}_i - T_i \tag{15}$$

where $\hat{E}_i$ is the quantized log energy in band i, and T_i is the log masking threshold in that band. Using the values of SMR_i's, the first allocation of the number of bits to each band is performed according to the following formula,

$$b_i = \max(\mathrm{SMR}_i L_i b_T / \sum_{i=1}^{17} \mathrm{SMR}_i L_i, 0) \tag{16}$$

where b_T is the total number of bits available to quantize the shape of the frequency spectrum within the critical bands. L_i is the number of coefficients in band i. After the first round of bit allocation, the decimal parts of b_i's will be discarded to leave the integer parts. As a result of using the floor function the total number of bits allocated in the first step will be less than b_T. To allocate the remaining bits, the Noise-to-Mask Ratio (NMR) is approximated for each band taking into account the bits already allocated in the first step,

$$\mathrm{NMR}_i = \hat{E}_i - T_i - \alpha_i b_i \tag{17}$$

Note that the assignment of few remaining bits is based on the assumption that NMR_i is equal to SMR_i for $b_i = 0$ and also a reduction of α_i dB in NMR_i for one bit assigned to band i. Note that α_i is obtained from the rate-distortion curve for each codebook. The value of NMR_i is calculated for all bands and an additional bit is then allocated to the band with the largest value of NMR. This process will continue until all remaining bits are allocated.

2.5 Concluding Remarks

We have developed a transform coder appropriate for a wide range of input signals. In order to achieve a high coding gain, we have used different means such as perceptually-based VQ and adaptive bit allocation. This coder outperforms other low rate speech coders for non-speech signals. Based on our observations, as long as there is no specific structure in the input signal, i.e. strong harmonic structure, this coder works well even at rates considerably lower than 8 kb/s. In the case of pseudo-periodic parts of the input signal, due to high sensitivity of the human ear to small variations of the harmonic structure, distortion is perceivable. Possible solutions are finer quantization of the harmonics and making the coder operate at two different rates based on the harmonic structure of the input frame.

Appendix A

Relation between the DFT and MDCT

The MDCT of a frame of input signal $x(n)$ is defined as [6]

$$C(k) = \sqrt{2/M} \sum_{i=0}^{N-1} x(n)h(n) \cos[\frac{\pi}{M}(n+n_0)(k+0.5)] \tag{18}$$

where $h(n)$ is the window function, N is the length of the input frame, $M = N/2$ is the number of transform coefficients in each frame and n_0 is a constant equal to $(M+1)/2$. Write the above formula as

$$C(k) = \sqrt{2/M} \sum_{i=0}^{N-1} \Re\{x(n)h(n) \exp[\frac{-j\pi(n+n_0)(k+0.5)}{M}]\} \tag{19}$$

$$= \sqrt{2/M} \Re\{\exp[j\phi(k)] \mathcal{F}(s(n))\} \tag{20}$$

where $\Re$ denotes the real part and $\mathcal{F}$ denotes the Fourier transform,

$$\phi(k) = \frac{-\pi(N+2)(k+0.5)}{2N} \tag{21}$$

$$s(n) = \exp(\frac{-j\pi n}{N}) x(n)h(n) \tag{22}$$

Finally we get

$$C(k) = \sqrt{2/M} |S(k)| \cos[\frac{2\pi n_0(k+0.5)}{N} - \angle S(k)] \tag{23}$$

References

[1] M. Dietz, J. Herre, B. Teichmann, and K. Brandenburg, "Bridging the Gap: Extending MPEG Audio down to 8 kbit/s," *102nd AES Convention* (Munich), 1997. Preprint 4508.

[2] H. Najafzadeh-Azghandi and P. Kabal, "Perceptual Coding of Narrowband Audio Signals at 8 kb/s," *Proc. IEEE Workshop on Speech Coding* (Pocono Manor, Penn.), pp. 109–110, 1997.

[3] J. D. Johnston, "Transform coding of audio signals using the perceptual noise criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, pp. 314–323, Feb. 1988.

[4] 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.

[5] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis.* Elsevier, 1995. pp. 427–428.

[6] H. Malvar, *Signal Processing with Lapped Transforms.* Artech House, 1992.