Proc. European Conf. Speech Commun., Technol. (Aalborg, Denmark), pp. 1993-1996, Sept. 2001

Narrowband Perceptual Audio Coding: Enhancements for Speech

Hossein Najaf-Zadeh and Peter Kabal

Electrical & Computer Engineering
McGill University, Montreal, Canada

Abstract

This paper presents a bi-modal coding paradigm to compress narrowband audio signals at 8 kbit/s. In the general mode, the Enhanced Narrowband Audio Coder (ENPAC) exploits the characteristics of the human hearing system to adaptively code the perceptually important spectral components of the input audio. The other mode is employed to handle audio inputs with a strong harmonic structure. In that mode, the input block is represented by its audible harmonics. The spectral magnitude is modeled by the linear prediction analysis in the time domain. The phase of each harmonic is predicted and the phase residues are quantized using an adaptive bit allocation algorithm. This paper introduces a perceptually-based upper bound for phase errors of spectral components. The ENPAC encoder delivers good quality for narrowband speech and non-speech inputs.

1 Introduction

Over the past decade, research in audio coding has been concentrated on high quality compression of wideband audio signals. However, new applications such as broadcasting over the Internet, consumer multimedia, narrowband digital AM broadcasting and wireless networks are emerging. In such applications either the number of users is large (e.g., the Internet) or the available bandwidth is limited (e.g., wireless communications). Therefore, moderate audio quality at bit rates below 16 kbit/s is appropriate [1, 2, 3].

Available audio coders operate at bit rates above 16 kbit/s. On the other hand speech-specific coders operating at bit rates lower than 16 kbit/s are not suitable for encoding audio signals. There is a gap between the operating bit rates of state-of-the-art narrowband speech coders (8 kbit/s and below) and low bit rate audio coders operating at 16 kbit/s and above.

We present the Enhanced Narrowband Perceptual Audio Coder (ENPAC) using different coding tools based on the characteristics of the human hearing system to fill the gap and accommodate a wide range of narrowband audio inputs at 8 kbit/s.

Speech-specific coders deliver good quality for speech inputs by preserving the harmonic structure of voiced speech. Since the hearing system is very sensitive to distortion in harmonic signals, the structured spectrum of the input audio must be well represented. Low rate transform coders generally perform well on non-speech inputs. However, they may not deliver as high quality speech as speech-specific coders. One solution to universal coding is to use a source-driven variable rate scheme to better quantize the spectrum of speech. However, this solution is not appropriate for fixed rate coding. An alternative is to use two coding schemes to handle speech/non-speech inputs differently. In those paradigms some discontinuities in the output might be perceived.

In our coding paradigm, the encoder employs a general mode to handle audio inputs. If a strong harmonic structure is detected the harmonic mode is used. That mode represents the input by its harmonics [4, 5]. This way the perceptually-important spectral components are well reproduced. Good quality especially for harmonic inputs such as voiced speech is achieved. Although we use two different coding paradigms, almost no discontinuity is perceived. The reason is that the successive input frames overlap by 50% and the output signal is produced by an overlap-add method independent of the operating mode.

2 Encoder Overview

The ENPAC encoder has been developed to compress narrowband audio data. The input signal is band-limited from 50 Hz to 3.6 kHz, sampled at 8 kHz, and represented with 16 bit linear PCM. The input samples are grouped into overlapping blocks of 240 samples and are input to the encoder.

In this section, we discuss the operating mode selection criterion followed by a description of the two operating modes, i.e., general and harmonic modes. A schematic diagram of the ENPAC encoder is shown in Fig. 1.

2.1 Mode Selection and Pitch Estimation

The operating mode of the ENPAC encoder is selected based on the structure of the input signal. In doing so, the input block of 240 samples is windowed and filtered using a 10-th order lowpass butterworth filter with a cutoff frequency of 2800 Hz. The pitch period is determined using the autocorrelation of the filtered input. Since the length of each frame is 240 samples, the pitch value is confined in the range of 20–120 samples. In order to find a more accurate pitch value, the autocorrelation function is up-sampled by 10. The pitch value is represented by an integer and a fractional part. Alternative pitch detection algorithms that work directly in the frequency domain are under investigation.

In selecting the operating mode, the energy concentration in the harmonics (corresponding to the pitch value)

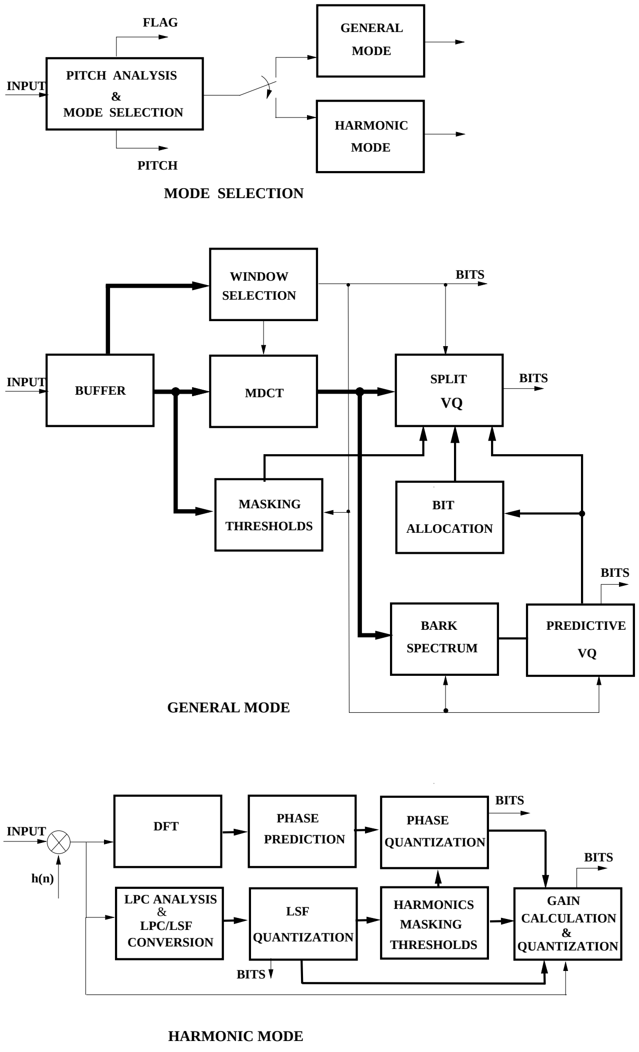

Fig. 1 Block diagram of the ENPAC encoder.

is compared to the average energy distribution in the DFT domain using the following criterion,

$$r = \frac{N_F \sum\limits_{k\in\mathcal{H}} |X_k|^2}{2N_H \sum\limits_{k=0}^{N_F/2-1} |X_k|^2}, \qquad (1)$$

where N_F is the length of the Fourier transform, N_H is the number of harmonics (including the fundamental frequency), $\mathcal{H}$ contains the indexes of the harmonics and X_k is the k-th Fourier coefficient. If r is greater than 3, the input frame is considered to have a strong harmonic structure and ENPAC operates in the harmonic mode. A one bit flag is employed to identify the operating mode.

2.2 Harmonic Mode

In this mode, the input frame has a strong harmonic structure. This situation often occurs for voiced speech and some single-instrument music. Our basic assumption in the harmonic mode is that a highly structured block of data can be well presented by its harmonic components as follows with almost no perceptual loss.

$$x(n) = \sum_{k=1}^{N_H} A_k \cos(2\pi k f_0 n + \phi_k), \qquad (2)$$

where A_k and ϕ_k are the amplitude and phase of the k-th harmonic, f_0 is the fundamental frequency.

2.2.1 Phase Modelling

In the harmonic mode, the phases of the audible components are coded. As is common in sinusoidal coding scheme, the phase for the m-th harmonic is linearly predicted from the corresponding value in the previous frame as follows.

$$\tilde{\phi}^{(j)} = \hat{\phi}^{(j-1)} + 2\pi m \bar{f}_0 \tau \qquad (3)$$

where $\tilde{\phi}^{(j)}$ is the predicted phase for the m-th harmonic in the j-th frame, $\hat{\phi}^{(j-1)}$ is the quantized phase of the m-th frequency component in the previous frame, τ is the time difference between successive frames, and $\bar{f}_0$ is the average fundamental frequency defined as

$$\bar{f}_0 = \frac{f_0^{(j)} + f_0^{(j-1)}}{2}. \qquad (4)$$

If the previous frame was labeled as non-harmonic, $\bar{f}_0 = f_0^{(j)}$. The phase residue is defined as

$$\triangle\phi^{(j)} = \phi^{(j)} - \tilde{\phi}^{(j)}. \qquad (5)$$

2.2.2 Quantization of Phase Residues

As is shown in Fig. 2, the linear prediction scheme performs more accurately at low frequencies, i.e., the histogram of the phase residues is narrower. We exploited this fact in quantizing the phase residues. The frequency band up to 3600 Hz is split into four subbands: 0–500, 500-1000, 1000–2000 and 2000–3600 Hz. For each subband seven scalar quantizers corresponding to different number of bits (1–7 bits) are designed. This quantizers are used along with a bit allocation scheme to quantize the phase residues.

2.2.3 Bit Allocation for Phase Quantization

Since there are different number of harmonics in each frame (9–54 for a pitch value of 20-120 samples in the frequency range up to 3600 Hz), we have developed a perception-based algorithm to allocate 63 bits to quantize the phase residues of audible harmonics. In doing so, the masking pattern due to the harmonics is computed based on the model proposed by Terhardt [6] to find the masking threshold for each harmonic. Note that the quantized harmonic magnitudes are used to calculate the masking pattern. Fig. 3 shows the masking pattern for a segment of voiced speech using the original harmonic magnitudes. As is seen, a number of harmonics are below the masking threshold. In the quantization scheme, 63 bits are equally distributed among the audible harmonics as follows

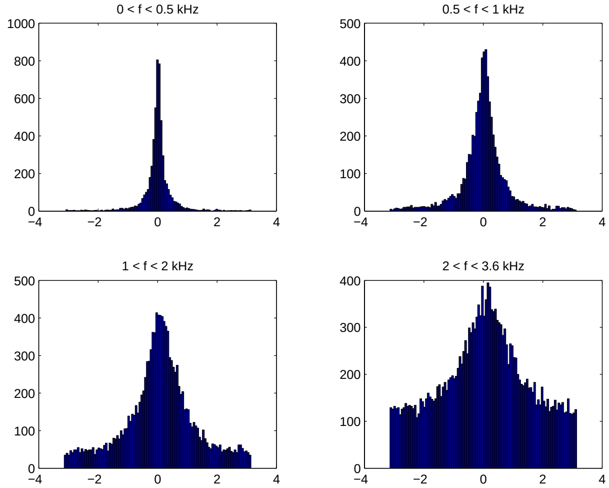

Fig. 2 Phase residues histograms.

$$b = \mathrm{int}(\frac{63}{N_{ah}}) \quad (6)$$

where b is number of bits assigned to each audible harmonics, int(.) gives the integer part and N_{ah} is the number of the audible harmonics. If the total number of bits assigned is less than 63, one more bit is allocated to the harmonics with the largest Signal-Mask-Ratio (SMR) until all the bits are assigned. The justification for the perception-based bit allocation is discussed in the following section. Note that, more accurate bit allocation schemes can easily be developed at the expense of higher computation.

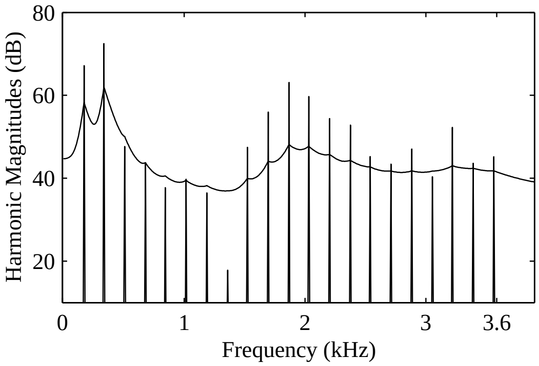

Fig. 3 Harmonics masking pattern for a block of voiced speech.

2.2.4 Upper Bound for Phase Errors

In this section, we present a perception-based upper bound for phase errors. For any audible spectral component, the error energy due to a phase error must be below the masking threshold as follows.

$$|(A^2\cos^2(\phi) - A^2\cos^2(\phi + \triangle\phi)| \leq m_{th} \quad (7)$$

where A and ϕ are the amplitude and phase of the component, $\triangle\phi$ is the phase error and m_th is the corresponding masking threshold. The worst case occurs when the cosine function has the highest rate of variation that is when $\phi = \frac{\pi}{2}$. Hence,

$$A^2\cos^2(\frac{\pi}{2} + \triangle\phi) \leq m_{th}. \quad (8)$$

Therefore, the upper bound for phase errors is given by

$$\triangle\phi \leq \arcsin\sqrt{\frac{m_{th}}{A^2}}. \quad (9)$$

Since the Signal-to-Mask-Ratio (SMR) is defined as A^2/m_{th}, Eq. (9) can be written as

$$\triangle\phi \leq \arcsin\frac{1}{\sqrt{\mathrm{SMR}}}. \quad (10)$$

This bound gives the audible level of phase errors for each component. As is seen, a large SMR requires a small phase error. That result is consistent with the perceptual importance of different spectral components.

2.2.5 Harmonic Magnitudes Modelling and Quantization

In the harmonic mode, the magnitudes of the harmonics are modeled using a 10-th order Linear Prediction (LP) analysis. The LP coefficients are converted to the LSF's and and a vector quantization scheme is employed to compress the LSF's. The LSF's are divided into four groups: 1–2, 3–4, 5–6, 7–10. A total number of 37 bits, i.e., (9+9+8+10) bits, are assigned to the LSF groups.

2.2.6 Pitch Information

The pitch period is confined to 20–120 samples and 11 bits are used to specify the integer and fractional parts of the pitch.

2.2.7 Gain Determination and Quantization

We assume that the major part of the energy of a harmonic frame is concentrated in the harmonics. However, in order to adjust the energy of the synthesized signal, we set the energy of the windowed original frame to the energy of the windowed synthesized frame of data. Therefore the gain is found as follows,

$$g = \frac{\sum_{n=0}^{N-1} x(n)h(n)}{\sum_{n=0}^{N-1} x_s(n)h(n)}. \quad (11)$$

where N is the length of the input frame, $x(n)$, $x_s(n)$ and $h(n)$ are the input signal, synthesized signal and window function respectively. A scalar quantizer with 256 levels, i.e., 8 bits, has been trained to quantize the gain.

2.3 General Mode

This operating mode has been previously reported in [7, 8]. In this mode, a frame of either 240 or 80 samples of the input signal is transformed into the frequency domain by means of a Modified Discrete Cosine Transform (MDCT) [9]. An energy-based window switching criterion is used to choose the length of the input frame. The MDCT coefficients are grouped into 17 critical bands. A Gain/Shape approach is taken to quantize the MDCT coefficients. The shape vectors are quantized using a split VQ scheme. The masking threshold is calculated to be used in the bit allocation and shape quantization. The SMR-based adaptive bit allocation scheme [10] and a perceptual error criterion is used in quantizing the shape vectors. An adaptive predictive VQ scheme is employed to quantize the gain factors.

This mode is suitable to compress inputs with either unstructured (e.g., noise like) or complex spectra (e.g., music).

2.4 Subjective Evaluation

It is difficult to make valid comparisons with existing coders as there are not many examples of coders designed for both music and speech. However, we have evaluated the performance of the ENPAC encoder on a large variety of narrowband speech and music signals. The ENPAC encoder has been compared to an audio coder in the same category, i.e., the Real Audio music encoder operating at 8 kb/s. According to all the listeners, ENPAC delivers better quality for all audio material. As for speech inputs, the bi-modal ENPAC achieves better performance compared to the general mode. Moreover, compared to other speech-specific coders operating at 8 kb/s, ENPAC delivers comparable quality for speech inputs.

3 Concluding Remarks

We have presented Enhanced Narrowband Perceptual Audio Coder (ENPAC) to compress narrowband inputs. The ENPAC encoder operates in two modes; general mode and harmonic mode. In the general mode, a block of input is converted to the frequency domain using an MDCT. The MDCT coefficients are quantized using perceptually-based VQ and adaptive bit allocation. In the harmonic mode, the input block is represented by the harmonics and the parameters of the harmonics, i.e., magnitudes, phases and the fundamental frequency, are encoded. A perceptually-based bit algorithm for phase quantization has been developed and used in ENPAC. An upper bound for phase errors as a function of the Signal-to-Mask-Ratio (SMR) has been obtained.

We have observed that ENPAC operates in the harmonic mode around 40% of the time for speech material. For music signals that figure depends on the input and varies from 5% to 35%. The use of the harmonic mode makes a considerable impact on the quality of the output signal especially for speech inputs.

The performance of ENPAC was improved by operating in two modes. Further improvement may be possible by employing another mode for perceptually-important transitional frames (e.g., frames containing an on-set or a rapid transition in the pitch contour). That new operating mode may be a combination of the two existing modes to differently treat spectral components based on their local strength.

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] M. Dietz, H. Popp, K. Brandenburg, and R. Friedrich, “ Audio Compression for Network Transmission ,” *J. Audio Eng. Soc.*, vol. 44, pp. 58–72, Jan. 1996.

[3] B. Feiten, R. Schwalbe, and F. Feige, “Dynamically scalable internet audio transmission,” *104th AES Convention* (Amsterdam), 1998. Preprint 4686.

[4] R. McAulay and T. Quatieri, “Sinusoidal Coding,” in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 4, pp. 427–428, Elsevier, 1995.

[5] R. McAulay and T. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[6] E. Terhardt, G. Stoll, and M. Seewann, “Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals,” *J. Acoust. Soc. Am.*, vol. 71, pp. 679–688, Mar. 1982.

[7] 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.

[8] H. Najafzadeh-Azghandi and P. Kabal, “Improving Perceptual Coding of Narrowband Audio Signals at Low Rates,” *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Phoenix, Arizona), pp. 913–916, 1999.

[9] 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.

[10] N. Najafzadeh and P. Kabal, “Perceptual Bit Allocation for Low Rate Coding of Narrowband Audio ,” *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing* (Istanbul), pp. 893–896, 2000.