Audio Engineering Society

Convention Paper

Presented at the 121st Convention
2006 October 5–8 San Francisco, CA, USA

New Results in Low Bit Rate Speech Coding and Bandwidth Extension

Raghuram Annadana[1], Harinarayanan. E.V[1], Anibal Ferreira[1,2], and Deepen Sinha[1]

[1] ATC Labs, New Jersey, USA

[2] University of Porto, Portugal

Correspondence should be addressed to Raghuram Annadana (raghu@atc-labs.com)

ABSTRACT

Emerging digital audio applications for broadcast radio and multimedia systems are presenting new challenges such as the need to code mixed audio content, error robustness, higher audio bandwidth and quality at low bit rates; demanding a paradigm shift in the existing speech coding techniques. This paper describes continuation of our research in low bit rate audio/speech coding and in the recently introduced Audio Bandwidth Extension Toolkit (ABET). Several new modes of operation have been introduced in the codec, in particular making innovative use of perceptual coding tools. Additionally, a new mode in ABET is added to improve the efficiency of the temporal shaping tool, Multi Band Temporal Amplitude Coding (MBTAC), by exploiting time and frequency correlation. The structure of the codec and its performance in these modes of operation are detailed. Further information is available at http://www.atc-labs.com/lbr/ and http://www.atc-labs.com/abet/.

1. INTRODUCTION

Speech and audio coding at low bit rates (e.g., 2.4-12 kbps) has numerous well established applications. These include telephony and other interpersonal communication frameworks, Voice over IP (VOIP), teleconferencing, broadcasting, etc. Several well known speech coding schemes exist and have been standardized by organizations such as ITU. However, emerging application scenarios require an improved performance with regard to the metrics viz. high robustness to mixed content, error robustness, higher audio bandwidth and scalability, not directly addressed by existing coding schemes.

We describe enhancements to our recently introduced low bit rate coding scheme [1] and the Audio Bandwidth Extension Toolkit (ABET) [2] that achieve significantly improved performance with respect to several of the above metrics. Several new algorithm components have been incorporated:

- New techniques for coding the excitation based on perceptual coding tools. The Psychoacoustic model requires certain adaptations for the coding of the residual signal, which is described.

- The ABET Toolkit incorporates a technique for accurate coding and application of the temporal envelope using a secondary utility filter bank, which compensates for the cases where the time resolution of primary coding filterbank is low. We have added new enhancements to MBTAC to take advantage of the special characteristics of the primary ABET bandwidth extension filterbank (MDCT/ODFT) and resulting frequency correlation as well as time correlation. The improved temporal envelope coding technique is also utilized in a new multichannel coding scheme [3].

The proposed enhancements in combination of previous results [4] yield an integrated coding scheme that supports coding modes from 3-96 kbps and has several of the attributes described above. Algorithmic components, and test results including audio samples are presented.

The organization of the rest of the paper is as follows. In section 2 we introduce the architecture and operation of the encoder operating in the 8 – 16 kbps range. Section 3 describes the corresponding architecture for the decoder. Section 4 discusses the additional enhancements that have been incorporated in the MBTAC tool. Section 5 presents some of the preliminary results and comparisons and is followed by conclusions in section 6

2. ENCODER ARCHITECTURE

Figure 1 illustrates the architecture of the codec and is detailed in this section. The enhancements have been made in the codec to incorporate new modes of operation in the 8–16 kbps range of bit rates. Linear Prediction (LP) filtering is used to eliminate the short term correlation in the signal and generate an excitation signal. LP parameters are estimated using the Leroux and Gueguen algorithm [5] from the autocorrelation of windowed speech frame. The residual is now analyzed in the MDCT domain. Optionally the input audio signal is also analyzed in the UFB and ODFT domains depending on the mode of operation and the compression efficiency required.

In the simplest mode of operation, none of the algorithmic blocks in ‘dots’ are in use and the output of the MDCT analysis filter bank is quantized with the aid of an elaborate psychoacoustic model. The goal in our case is to quantize the excitation signal in such a way that the quantization noise is either fully masked or rendered less annoying after signal reconstruction due to masking.

The core of perceptual modeling is the concept of auditory masking [6]. Building the perception model in an audio codec typically involves the utilization of the following four key concepts: *simultaneous masking*, *temporal masking*, *frequency spread of masking*, and, *tone vs. noise like nature* of the masker. Simultaneous masking is a phenomenon whereby a *masker* is found to mask the perception of a *maskee* occurring at the same time. Temporal masking refers to a phenomenon in which a *masker* masks a *maskee* occurring either prior to or after its occurrence. Frequency spread of masking refers to the phenomenon that a masker at a certain frequency has a masking potential not only at that frequency but also at neighboring frequencies. Finally, the masking potential of a narrow band masker is strongly dependent on the tone vs. noise like nature of the masker. These factors are utilized to estimate desired quantization accuracy, or Signal to Mask Ratio (*SMR*) for each band of frequency. In order to tune the psychoacoustic model to quantize the residual, the following strategy has been deployed. The scale factors have been reduced empirically so as to allow for a greater Signal to Mask Ratio (SMR). In addition, at the decoder, some of the factors lost due to the quantization process are filled based on previous values so as to improve the naturalness of the output audio.

At bit rates such as 12, 14 and 16 Kbps bandwidth extension can be optionally applied to the high resolution MDCT spectral representation of the input audio signal using the Audio Bandwidth Extension Toolkit (ABET). The toolkit consists of two main algorithms for bandwidth extension, viz. Fractal Self Similarity Model (FSSM) [7] and the Accurate Spectral Replacement [8]. The FSSM model works across a wide class of natural audio and is capable of providing detailed and natural sounding audio reconstruction. ASR is capable of an extremely accurate reconstruction

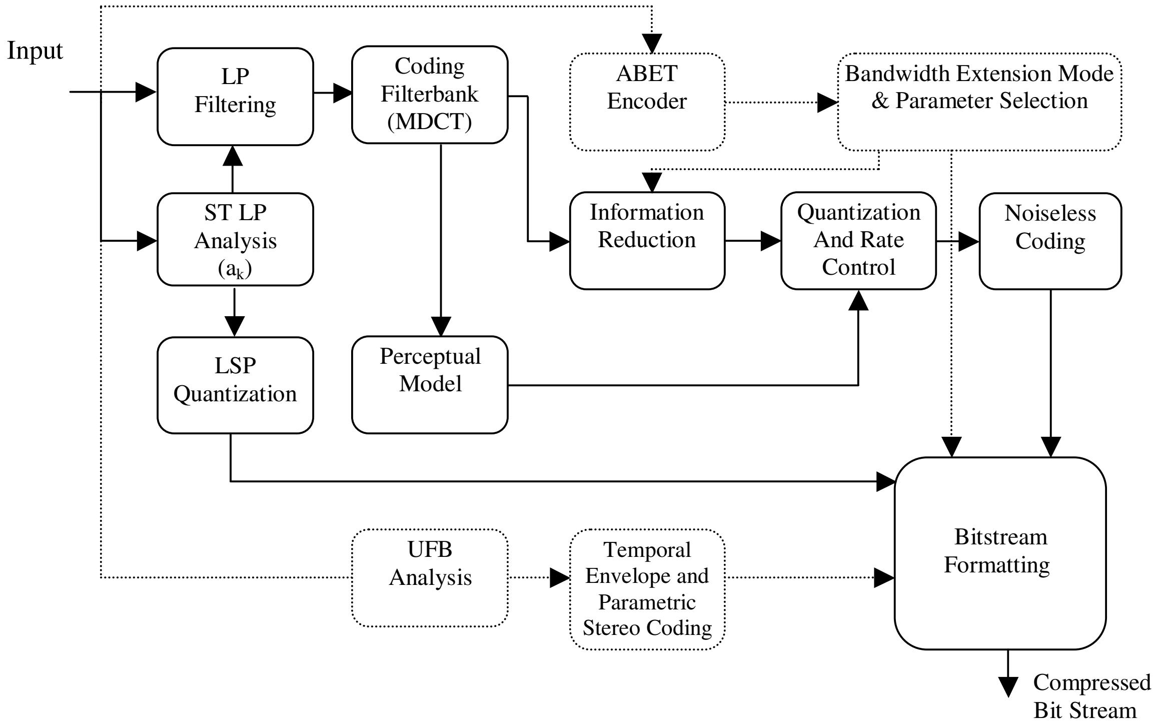

Figure 1: Architecture of Encoder

of the tonal components and harmonic structures in the synthesized high frequency spectrum of the signal. ABET also incorporates a third essential tool “Multi Band Temporal Amplitude Coding” (MBTAC). MBTAC may (optionally) be employed when the time resolution of the primary MDCT/ODFT filterbank is too low to allow for suitable temporal shaping of the reconstructed high frequency components. For the computation and application of MBTAC, the audio signal is analyzed using a secondary Utility Filter Bank (UFB) that has a significantly better time resolution.

The linear prediction coefficients are transformed to Line Spectral Pair (LSP) domain and are quantized using the scheme presented in [1]. Statistical redundancies in the quantized MDCT spectral coefficients are reduced by Huffman coding followed by bit stream formatting.

3. DECODER ARCHITECTURE

Figure 2 shows the corresponding architecture for the decoder. The dotted blocks indicate they are optional and are not used in some modes of operation. The major processing steps at the decoder consist of Huffman decoding, inverse quantization and signal reconstruction. Huffman decoding and inverse quantization yields MDCT coefficients. The excitation is obtained by an inverse MDCT transform of the obtained coefficients. The signal reconstruction involves inverse LP filtering the obtained residual to obtain the output audio. The ABET decoder performs optional signal synthesis using the *FSSM* and *ASR* model in the MDCT domain. In the cases where the time resolution of the MDCT/ODFT filterbank is too high to allow for adequate temporal shaping, the MBTAC information is applied in the UFB domain.

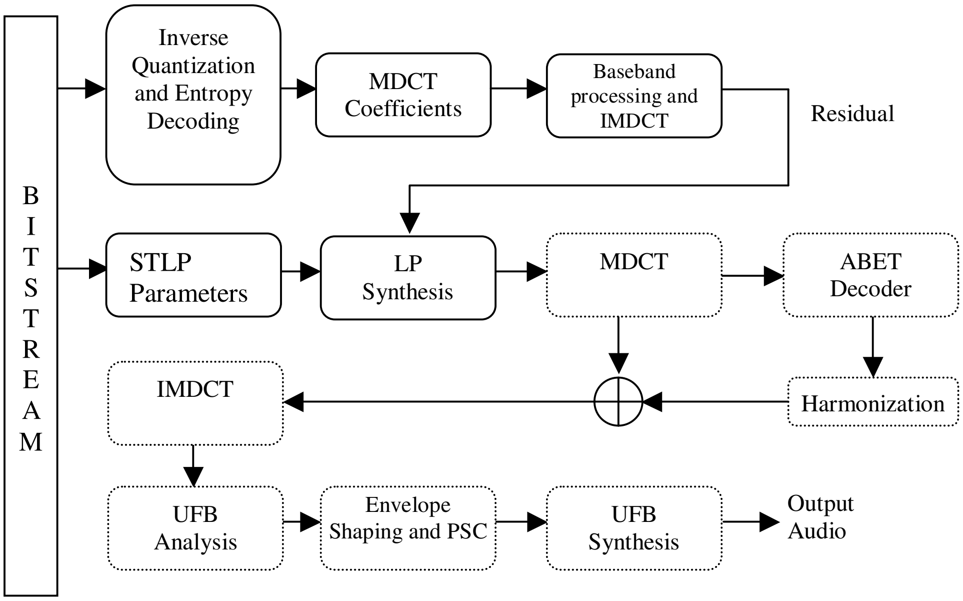

Figure 2: Architecture of Decoder

4. TEMPORAL ENVELOPE CODING

The Multi Band Temporal Amplitude Coding (MBTAC) tool is utilized to extract and code the temporal envelope of the signal. The input audio signal is analyzed using a Utility Filter Bank (UFB) to obtain a high resolution secondary filter bank compensating for the lower resolution MDCT analysis filter bank employed in bandwidth extension.

The MBTAC is designed as a cascade of time-frequency grouping techniques on the time-frequency plane followed by an innovative difference coding scheme. It utilizes a two level time-frequency envelope grouping scheme. Grouping is essentially performed on energies on the time-frequency grid.

An initial first level of grouping on the time-frequency plane of the UFB is carried out separately in time and frequency. Frequency grouping is performed on the critical band scale. The time grouping is designed to have a higher resolution above the transition frequency and a lower resolution below the transition frequency. A second level of grouping is performed to exploit the energy stationarity in the time and frequency planes.

4.1. Correlation Based Coding

In order to reduce the bit demand further, a novel correlation based scheme is incorporated to exploit the high correlation that exists between the grouped time-frequency envelopes. A block diagram illustrating this scheme is presented in Figure 3. This technique identifies and replaces correlated envelopes by a single parametric value thereby, reducing the overall bit demand of the system. The method involves finding a parametric correlation value ‘a’ between a pair of envelopes which we will call as a primary and a secondary envelope represented by Equations 1 and 2.

$$X = [x_1 \quad x_2 \quad . \quad . \quad . \quad x_n]^T \tag{1}$$

$$Y = [y_1 \quad y_2 \quad . \quad . \quad . \quad y_n]^T \tag{2}$$

The parameter ‘a’ is computed so as to minimize distance between any two envelopes with a scalar product and the expression is as shown in Equation 3.

$$a = \frac{x_1 . y_1 + x_2 . y_2 + \ldots + x_n . y_n}{x_1^2 + x_2^2 + \ldots + x_n^2} \tag{3}$$

With the knowledge of the correlation parameter, the secondary envelope is synthesized at the encoder. The original secondary and synthesized envelopes are compared by computing the Euclidian distance D between them, given by the expression.

$$D = | Y - a.X | \tag{4}$$

Where, |.| represent the modulus.

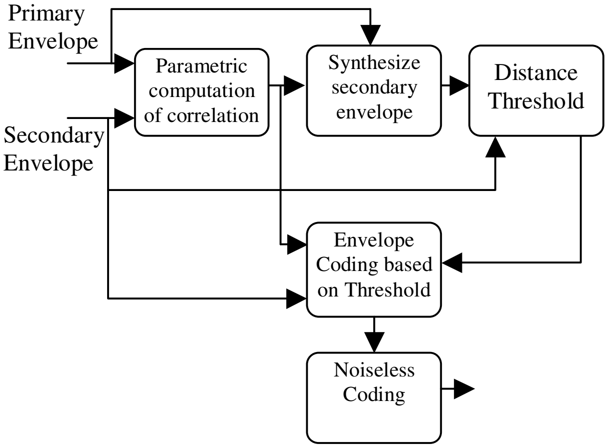
Primary Envelope
Secondary Envelope
Parametric computation of correlation
Synthesize secondary envelope
Distance Threshold
Envelope Coding based on Threshold
Noiseless Coding

Figure 3: Illustration of Correlation based coding

A final decision on whether to code the parameter 'a' or the secondary envelope as a whole is made based on a threshold.

5. RESULTS

The proposed coding scheme is capable of operating in the 8 -16 Kbps range. Various audio samples demonstrating the performance of the scheme are available at the web site: www.atc-labs.com/lbr.

Figure 4 and Figure 5 displays the bit-rate gained over different threshold values with the encoder operating at 12 Kbps and 48 Kbps respectively in the correlation based coding scheme. When the distance threshold fails between the primary and the waveform of comparison, the primary waveform is updated to this new waveform making the algorithm more robust.

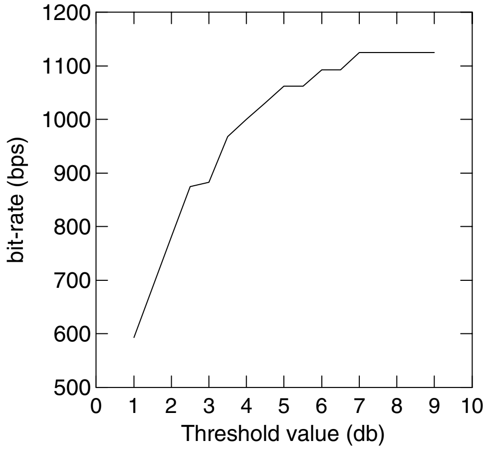

Figure 4: A Plot of bit-rate gained encoder operational at 12 Kbps.

Informal subjective tests have been conducted and more formal ones are in progress.

- 8 Kbps: No bandwidth extension and temporal shaping tools are used at this bit rate. At this bit rate, the audio quality is superlative to that obtained from G729 codec and is comparable to that of AMR-WB+ codec. In particular, sections of music and mixed content have features of the original audio and are less "flat".

- 10 Kbps: No bandwidth extension tools are used in this mode of operation and only temporal shaping tools are used. A 4 kHz mono audio bandwidth is maintained at this bit rate. This bit rate exhibits improved audio qualities readily noticeable in speech segments. The audio clarity improves in comparison in the 8 Kbps mode.

- 12 Kbps: At this bit rate, the ABET bandwidth extension algorithm is used to extend frequencies above 4 kHz. Natural wideband audio is obtained using this mode of operation.

- 14 Kbps: 12 kHz mono audio bandwidth is maintained at this bit rate. The ABET bandwidth extension algorithm is used to reconstruct frequencies above 4 kHz. MBTAC is used to accurately reproduce the temporal envelope.

- 16 Kbps: This mode of operation is similar to the 12 kbps mode with the exception of more accurate baseband coding.

In particular, these coding modes have been found to be increasingly robust to mixed and non-speech content.

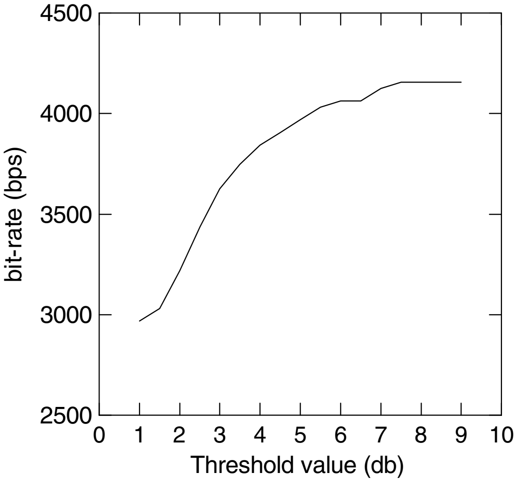

Figure 5: A Plot of bit-rate gained with encoder operational at 48 Kbps.

6. CONCLUSIONS

We have presented the architectures and the operation of our codec in the 8 – 16 kbps range. Techniques for transform coding the excitation using perceptual coding tools have been presented. In addition, new techniques incorporated in the ABET toolkit for accurate coding and application of temporal envelope have been presented. The coding modes have been found to be robust to non speech content making these modes of operation attractive for application scenarios such as VOIP. These enhancements in combination with our previous results yield an integrated coding scheme supporting coding modes from 3 – 96 kbps. Audio demos are available at http://www.atc-labs.com/lbr/.

7. REFERENCES

[1] Raghuram. A, A. J. S Ferreira, and D. Sinha, “A New Low Bit Rate Speech Codec for Mixed Content”, *in the preprints of 120th Convention of the Audio Engineering Society, May 2006.*

[2] D. Sinha, A. J. S Ferreira and Harinarayanan E. V., “A Novel Integrated Audio Bandwidth Extension Toolkit (ABET)”, *in the preprints of 120th Convention of the Audio Engineering Society, May 2006.*

[3] Chandresh Dubey, Richa Gupta, Deepen Sinha and Anibal Ferreira, “A Novel Very Low Bit Rate Multi-Channel Audio Coding Scheme using Accurate Temporal Envelope Coding and Signal Synthesis Tools”, *In the preprints of AES 121st Convention.*

[4] D. Sinha and A. J. S Ferreira, “A New Broadcast Quality LBR Audio Coding Scheme Utilizing Novel Bandwidth Extension Tools”, *in the preprints of 119th Convention of the Audio Engineering Society, Oct 2005.*

[5] J. Le Roux and C. Gueguen, "A fixed point computation of Partial Correlation Coefficients", *IEEE Trans on Acoustics, Speech and Signal Proc.*, vol. 27, no. 3, pp. 257 - 259, June 1977.

[6] B.C.J. Moore, *An Introduction to the Psychology of Hearing, 5th Ed.*, Academic Press, San Diego (2003).

[7] D. Sinha, A. J. S Ferreira and D. Sen, “A Fractal Self-Similarity Model for the Spectral Representation of Audio Signals”, *In the Preprint of 118th AES Convention*, Barcelona, Spain.

[8] A. J. S. Ferreira and D. Sinha, “Accurate Spectral Replacement”*, In the Preprint of 118th AES Convention*, Barcelona, Spain. Convention Paper 6383.