IDIAP RESEARCH REPORT

Autoregressive Modelling of Hilbert Envelopes for Wide-band Audio Coding

Sriram Ganapathy [a b] Petr Motlicek [a]
Hynek Hermansky [a b] Harinath Garudadri [c]

IDIAP–RR 08-40

June 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[c] Qualcomm Inc., San Diego, CA, USA

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-40

Autoregressive Modelling of Hilbert Envelopes for Wide-band Audio Coding

Sriram Ganapathy Petr Motlicek Hynek Hermansky
Harinath Garudadri

June 2008

Abstract. Frequency Domain Linear Prediction (FDLP) represents the technique for approximating temporal envelopes of a signal using autoregressive models. In this paper, we propose a wide-band audio coding system exploiting FDLP. Specifically, FDLP is applied on critically sampled sub-bands to model the Hilbert envelopes. The residual of the linear prediction forms the Hilbert carrier, which is transmitted along with the envelope parameters. This process is reversed at the decoder to reconstruct the signal. In the objective and subjective quality evaluations, the FDLP based audio codec at 66 kbps provides competitive results compared to the state-of-art codecs at similar bit-rates.

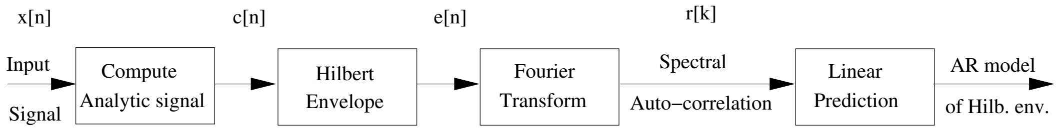


Figure 1: Steps involved in deriving the AR model of Hilbert envelope.

1 Introduction

Conventional approaches to speech coding achieves compression with a linear source-filter model of speech production using the linear prediction (LP) [1]. The residual of this modeling process represents the source signal. While such approaches are commercially successful for toll quality conversational services, they do not perform well for mixed signals in many emerging multimedia services. On the other hand, perceptual codecs typically used for multi-media coding applications are not as efficient for speech content.

In this paper, we propose to exploit the predictability of slowly varying amplitude modulations for encoding speech/audio signals. Spectral representation of amplitude modulation in sub-bands, also called "Modulation Spectra", have been used in many engineering applications. Early work done in [2] for predicting speech intelligibility and characterizing room acoustics are now widely used in the industry [3]. Recently, there has been many applications of such concepts for robust speech recognition [4, 5], audio coding [6], noise suppression [7], etc.

Our approach is based on the assumption that speech/audio signals in critical bands can be represented as a modulated signal with the Amplitude Modulating (AM) component obtained using Hilbert envelope estimate and Frequency Modulating (FM) component obtained from the Hilbert carrier. The Hilbert envelopes are estimated using Frequency Domain Linear Prediction (FDLP), which is an efficient technique for autoregressive modelling of the temporal envelopes of the signal [8, 9]. This idea was first applied for audio coding in the MPEG2-AAC (Advanced Audio Coding) [10], where it was primarily used for Temporal Noise Shaping (TNS). Specifically, the technique was used to eliminate pre-echo artifacts associated with transients by removing the parameterized time envelope prior to quantization. At the decoder, the temporal envelope is used to modulate the reconstructed residual signal to obtain the original signal back.

In the proposed audio codec, we utilize the concept of linear prediction in spectral domain on sub-band signals. We use a non-uniform Quadrature Mirror Filter (QMF) bank to derive 32 critically sampled frequency sub-bands. The QMF sub-bands simulate the critical band decomposition observed in the human auditory system. FDLP is applied over relatively long segments ($\sim$ 1000 ms) to estimate Hilbert envelopes in each sub-band. The remaining residual signal (Hilbert carrier) is further processed and its frequency components are selectively quantized. At the decoder, the sub-band signal is reconstructed by modulating the inverse quantized residual with the Hilbert envelope. This is followed by a QMF synthesis to obtain the audio signal back.

The current version of the codec provides high-fidelity audio compression at $\sim$ 66 kbps for speech/audio content. In the subjective listening tests, the FDLP codec was judged to be better than MPEG-1 Layer 3 (MP3) [11] and similar to MPEG-4 HE-AAC codec [12].

The rest of the paper is organized as follows. Section 2 explains the FDLP technique for the autoregressive modelling of Hilbert envelopes, where a mathematical description is provided. Section 3 describes the different blocks of the proposed codec in detail. The results of the objective and subjective evaluations are reported in Section 4 followed a summary in Section 5.

2 Autoregressive modelling of the Hilbert Envelopes

Autoregressive (AR) modelling is a technique that removes the simple form of redundancy in signal sequences by representing the current sample by an optimal linear combination of a fixed length history. AR models describe the original sequence as the output of filtering a temporally-uncorrelated (white) excitation sequence through a fixed length all-pole digital filter. Typically, AR models have been used in speech/audio applications for representing the envelope of the power spectrum of the signal by performing the operation of Time Domain Linear Prediction (TDLP) [13].

This paper utilizes AR models for obtaining smoothed, minimum phase, parametric models of temporal rather than spectral envelopes. The duality between the time and frequency domains means that AR modeling can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples. Since we apply the LP technique to exploit the redundancies in the frequency domain, we call this approach Frequency Domain Linear Prediction (FDLP) [14]. For the FDLP technique, the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal (in a manner similar to the approximation of the power spectrum of the signal by the TDLP).

The relation between the Hilbert envelope of a signal and the auto-correlation of the spectral components is shown in the following subsection. These relations form the basis for the FDLP technique in autoregressive modelling of the Hilbert envelopes.

2.1 A Simple Mathematical Description

Let $x[n]$ denote a discrete-time real valued signal of finite duration N. Let $c[n]$ denote the complex analytic signal of $x[n]$ given by

$$c[n] = x[n] + j\ \mathcal{H}\big[x[n]\big], \tag{1}$$

where $\mathcal{H}[.]$ denotes the Hilbert Transform operation. Let $e[n]$ denote the Hilbert envelope (squared magnitude of the analytic signal), i.e.,

$$e[n] = |c[n]|^2 = c[n]c^*[n], \tag{2}$$

where $c^*[n]$ denotes the complex conjugate of $c[n]$. The Discrete Fourier Transform (DFT) of the Hilbert envelope can be written as

$$E[k] = C[k] * C^*[-k] = \sum_{p=1}^{N} C[p]C^*[p-k] = r[k], \tag{3}$$

where $C[k]$ denotes the DFT of the analytic signal and $r[k]$ denotes the spectral auto-correlation function. The above expression can be re-written as

$$r[k] = \mathcal{F}\left\{e[n]\right\}, \tag{4}$$

where $\mathcal{F}$ denotes the DFT operation. Eq. 4 shows that the Hilbert envelope of the signal and the auto-correlation in the spectral domain form Fourier Transform pairs. In a manner similar to the computation of the auto-correlation of the signal using the inverse Fourier Transform of the power spectrum, the spectral auto-correlation function can be obtained as the Fourier transform of the Hilbert envelope of the signal. These spectral auto-correlations are used for AR modelling of the Hilbert envelopes (by solving a linear system of equations [13]).

The block schematic showing the steps involved in deriving the AR model of Hilbert envelope is shown in Fig. 1. The first step is to compute the analytic signal for the input signal. For a discrete time signal, the analytic signal can be obtained using the DFT [15]. Specifically, the procedure for creating a complex-valued N-point discrete-time analytic signal $c[n]$ from a real-valued N-point discrete time signal $x[n]$, is given as follows:

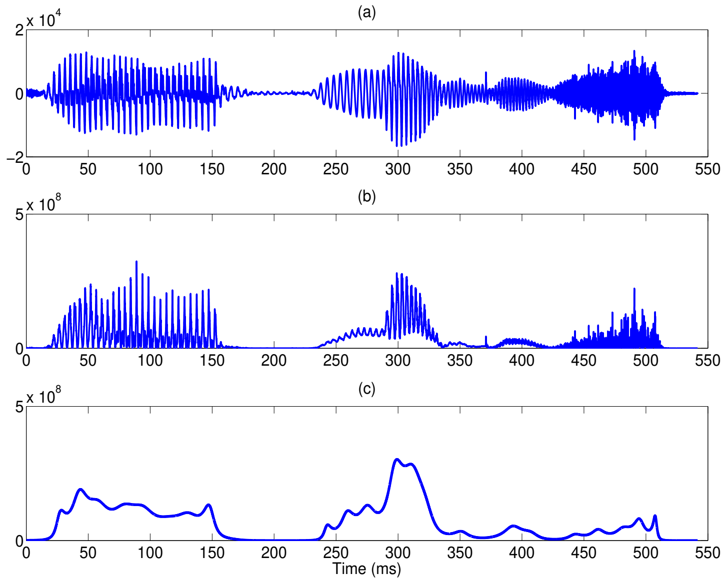


Figure 2: Illustration of the AR modelling property of FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) all pole model obtained using FDLP.

- Find the N-point DFT of the signal $X[k]$.
- Form the N-point one-sided discrete-time analytic signal spectrum by making the signal DFT $X[k]$ causal (assuming N to be even)

$$C[k] = \begin{cases} X[0], & \text{for } k = 0 \\ 2X[k], & \text{for } 1 \leq k \leq \frac{N}{2} - 1 \\ X[\frac{N}{2}], & \text{for } k = \frac{N}{2} \\ 0. & \text{for } \frac{N}{2} + 1 \leq k \leq N \end{cases}$$

- Find the N-point inverse DFT (IDFT) of $C[k]$ to obtain $c[n]$.

In general, the spectral auto-correlation function will be complex since the Hilbert envelope is not even-symmetric. In order to obtain a real auto-correlation function (in the spectral domain), we symmetrize the input signal in the following manner

$$x_e[n] = \frac{x[n] + x[-n]}{2},$$

where $x_e[n]$ denotes the even-symmetric part of $x[n]$. The Hilbert envelope of $x_e[n]$ will also be even-symmetric and hence, this will result in a real valued auto-correlation function in the spectral domain. Once the AR modelling is performed, the resulting FDLP envelope is made causal to correspond to the original causal sequence $x[n]$. This step of generating a real valued spectral auto-correlation is done for simplicity in the computation, although, the linear prediction can be done equally well for complex valued signals. The remaining steps given in Fig. 1 follow the mathematical relation given in Eq. 4.

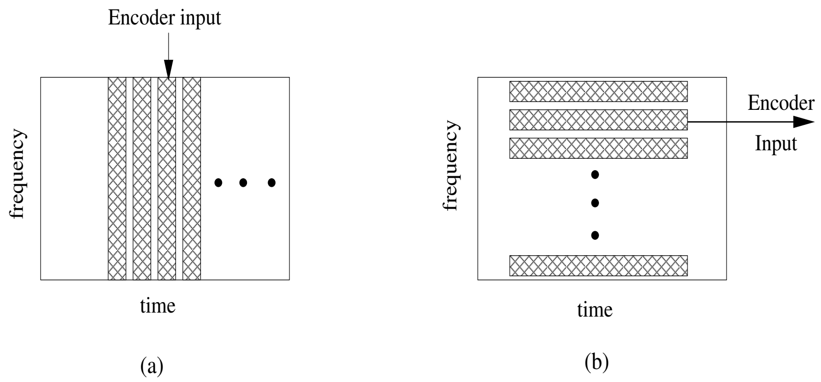


Figure 3: Overview of time-frequency energy representation for (a) conventional codecs and (b) proposed FDLP codec.

2.2 Modelling Temporal Envelopes with FDLP

Just as the conventional AR models are used effectively on signals with spectral peaks, the AR models of the temporal envelope are appropriate for peaky temporal envelopes. The individual poles in the resulting polynomial are directly associated with specific energy maxima in the time waveform. For signals that are expected to consist of a fixed number of distinct transients, fitting an AR model constrains the modelled envelope to be a sequence of maxima, and the AR fitting procedure removes the finer-scale detail. This suppression of detail is particularly useful in audio coding applications, where the goal is to extract the general form of the signal by means of a parametric model and to characterize the residual with a small number of bits. An illustration of the all-pole modelling property of the FDLP technique is shown in Fig. 2, where we plot a portion of speech signal, its Hilbert envelope and the AR model fit to the Hilbert envelope.

With a small approximation error, the Hilbert envelope can be shown to be the squared AM envelope [16]. Hence, the operation of splitting a signal into FDLP envelope and FDLP residual can be considered as a AM-FM decomposition. We apply the FDLP technique to speech/audio signals in critical bands where these signals are assumed to be modulated signals. In this manner, the proposed FDLP operation is similar to AM-FM decomposition proposed in [17].

The representation of signal information in the time-frequency domain is dual to that in the conventional codecs (Fig. 3). The state-of-art audio codecs (for example MP3, AAC, etc.) represent the time-frequency energy distribution of the signal by quantizing the short-term spectral or transform domain coefficients. The signal at the decoder is reconstructed by decoding the individual time frames. In the proposed FDLP codec, long temporal segments of the signal (typically of the order hundreds of ms) are processed in narrow sub-bands (which simulate the critical band decomposition). The reconstruction is achieved by recreating the individual sub-bands signals which is followed by a sub-band synthesis.

3 FDLP based Audio Codec

Long temporal segments (typically 1000 ms) of the full-band input signal are decomposed into frequency sub-bands. In each sub-band, FDLP is applied and Line Spectral Frequencies (LSFs) approximating the sub-band temporal envelopes are quantized using Vector Quantization (VQ). The residuals (sub-band carriers) are processed in spectral domain with the magnitude spectral parameters quantized using VQ. Since a full-search VQ in this high dimensional space would be computationally infeasible, the split VQ approach is employed. Although this forms a suboptimal approach to VQ, it

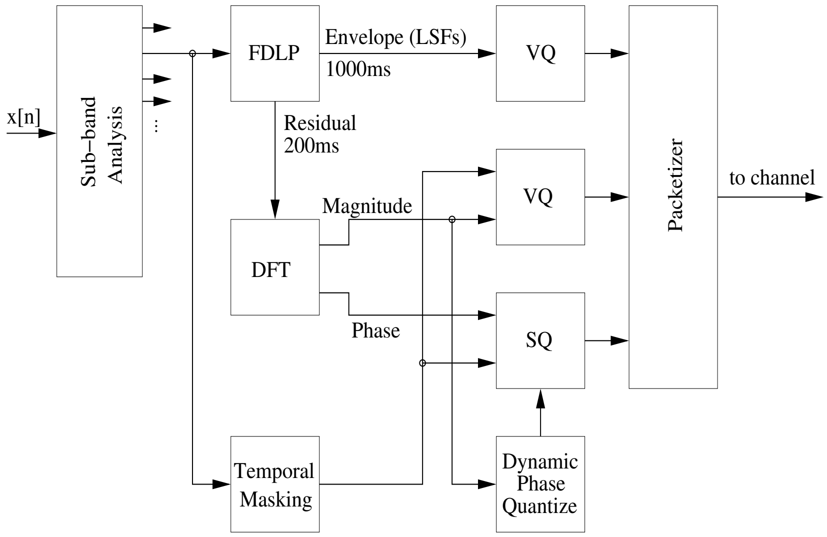


Figure 4: Scheme of the FDLP encoder.

reduces computational complexity and memory requirements without severely affecting the VQ performance. Phase spectral components of the sub-band residuals are Scalar Quantized (SQ). Graphical scheme of the FDLP encoder is given in Fig. 4.

In the decoder, shown in Fig. 5, quantized spectral components of the sub-band carriers are reconstructed and transformed into time-domain using inverse DFT. The reconstructed FDLP envelopes (from LSF parameters) are used to modulate the corresponding sub-band carriers. Finally, sub-band synthesis is applied to reconstruct the full-band signal. The final version of the FDLP codec operates at ~ 66 kbps. The important blocks are:

Non-Uniform QMF decomposition - The FDLP codec utilizes a perfect reconstruction non-uniform QMF bank to decompose the input signal to 32 sub-bands. An initial decomposition with a 6 stage tree-structured uniform QMF analysis gives 64 uniformly spaced sub-bands. A non-uniform QMF decomposition into 32 frequency sub-bands is obtained by merging these 64 uniform QMF bands. The merging operation is performed in such a way that bandwidths of the resulting critically sampled sub-bands emulate the characteristics of the critical band filters in the human auditory system [18].

Temporal Masking - Temporal masking refers to the hearing phenomenon, where the exposure to a sound reduces response to following sounds for a certain period of time (up to 200 ms) [19]. In the proposed version of the codec, a first order forward masking model of the human ear is implemented and informal listening experiments were performed to obtain the exact noise masking thresholds. Subsequently, this masking model is employed in encoding the sub-band FDLP carrier signal [20]. Application of the temporal masking in the FDLP codec results in a bit-rate reduction of about 10% without degrading the quality.

Dynamic phase quantization - It is found that the phase spectral components of the sub-band residual signal are uncorrelated. These phase components have a distribution close to uniform, and therefore, have high entropy. To prevent excessive consumption of bits, those corresponding to relatively low magnitude spectral components are transmitted using lower resolution SQ, i.e., the magnitude codebook vector is processed at the encoder with adaptive thresholding [21]. The spectral phase components whose magnitudes lie above a threshold are transmitted using high

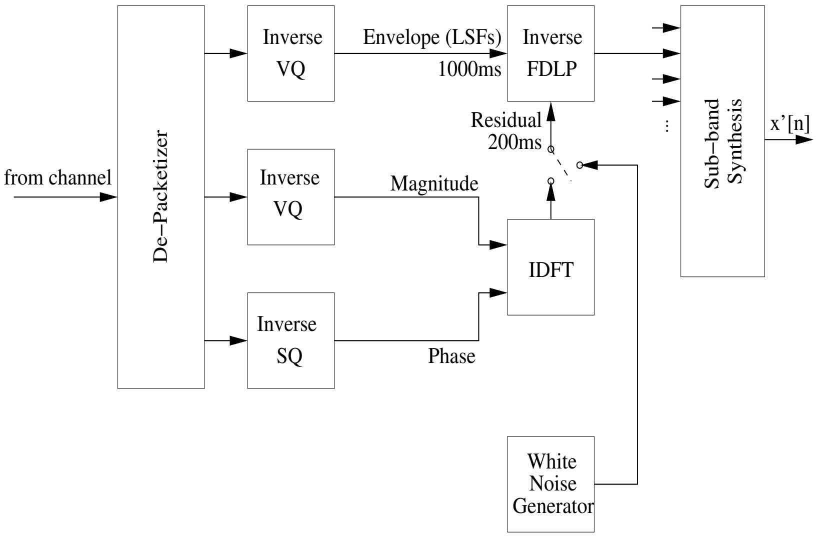


Figure 5: Scheme of the FDLP decoder.

ODG Scores	Quality
0	imperceptible
−1	perceptible but not annoying
−2	slightly annoying
−3	annoying
−4	very annoying

Table 1: PEAQ scores and their meanings.

resolution SQ and those lying below the threshold are transmitted with fewer bits. The threshold is adapted dynamically to meet a specified bit-rate. As the dynamic phase quantization follows an analysis-by-synthesis scheme, no side information needs to be transmitted.

Noise substitution - FDLP residuals in frequency sub-bands above 12 kHz are not transmitted, but they are substituted by white noise in the decoder. Subsequently, white noise residuals are modulated by corresponding sub-band FDLP envelopes. Such an operation has a minimum impact on the quality of reconstructed audio (even for tonal signals) and provides a bit-rate reduction of 15%.

4 Quality Evaluations

The subjective and objective evaluations of the proposed audio codec are performed using audio signals (sampled at 48 kHz) present in the framework for exploration of speech and audio coding [22]. It is comprised of speech, music and speech over music recordings. The music samples contain a wide variety of challenging audio samples ranging from tonal signals to highly transient signals.

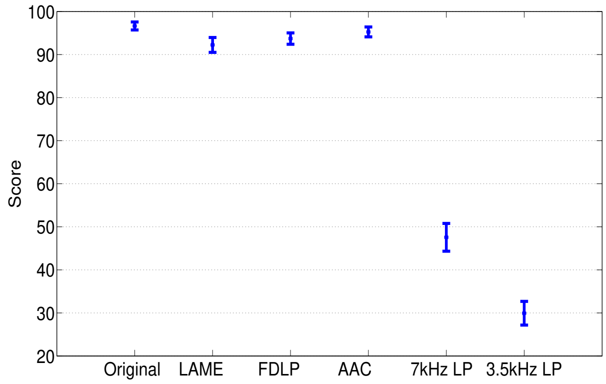
The objective and subjective quality evaluations of following 3 codecs are compared:

1. The proposed FDLP codec at ~ 66 kbps.

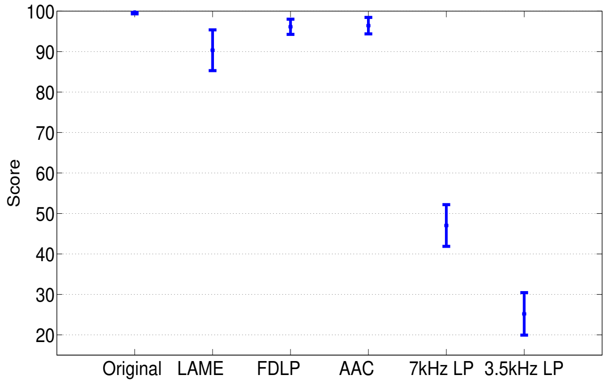
2. LAME MP3 (MPEG 1, layer 3) [11] at 64 kbps.

bit-rate [kbps]	66	64	64
system	FDLP	LAME	HE-AAC
ODG Scores	-1.11	-1.61	-0.77

Table 2: Mean objective quality test results provided by PEAQ [25] for 27 files with mixed signal content from MPEG database for explorations in Speech and Audio Coding [22].



(a) 22 listeners.



(b) 4 expert listeners.

Figure 6: MUSHRA results for 8 audio samples using three coded versions (FDLP, MPEG-4 HE-AAC (AAC+) and LAME MP3), hidden reference (original) and two anchors (7 kHz and 3.5 kHz low-pass filtered).

3. MPEG-4 HE-AAC, v1 at ∼64 kbps [12]. The HE-AAC coder is the combination of Spectral Band Replication (SBR) [23] and Advanced Audio Coding (AAC) [24].

4.1 Objective Evaluations

The objective measure employed is the Perceptual Evaluation of Audio Quality (PEAQ) distortion measure [25]. In general, the perceptual degradation of the test signal with respect to the reference signal is measured, based on the ITU-R BS.1387 (PEAQ) standard. The output combines a number of model output variables (MOV's) into a single measure, the Objective Difference Grade (ODG) score, which is an impairment scale with meanings shown in Table 1. For the objective quality evaluations, the PEAQ scores were computed for 27 audio samples from the database. The mean objective scores, for the three codecs in consideration, are shown in Table 2.

4.2 Subjective Evaluations

The qualitative performance of the complete codec is evaluated using MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening tests [26] performed on 8 audio samples from MPEG audio exploration database. The cumulative MUSHRA scores (mean values with 95% confidence) are shown in Fig. 6(a) and (b). MUSHRA tests were performed independently in two different labs (with the same setup). Fig. 6(a) shows mean scores for the results from both labs (combined scores for 18 non-expert listeners and 4 expert listeners), while Fig. 6(b) shows mean scores for 4 expert listeners in one lab. The subjective evaluations show that the FDLP codec performs better than LAME-MP3 and achieves subjective results close to MPEG-4 HE-AAC standard.

5 Conclusions

A technique for autoregressive modelling of the Hilbert envelopes is presented, which is employed for developing a wide-band audio codec. Specifically, the technique of linear prediction in the spectral domain is applied on relatively long segments of speech/audio signals in QMF sub-bands (which follow the human auditory critical band decomposition). The FDLP technique is able to adaptively capture the fine temporal nuances with high temporal resolution while at the same time summarizes the spectrum in time scales of hundreds of milliseconds. The objective and subjective evaluations, performed with the current version of the FDLP codec, suggest that the FDLP codec operating at $\sim$ 66 kbps provides better audio quality than the LAME - MP3 codec at 64 kbps and gives competent results compared to MPEG-4 HE-AAC standard at $\sim$ 64 kbps. Furthermore, the current version of the FDLP codec does not utilize the standard modules for compression efficiency provided by entropy coding and simultaneous masking. These form part of the future work.

6 Acknowledgements

This work was partially supported by grants from ICSI Berkeley, USA; the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2"; managed by the IDIAP Research Institute on behalf of the Swiss Federal Authorities, and by the European Commission 6^{th} Framework DIRAC Integrated Project. The authors would also like to thank Vijay Ullal for his active involvement in the subjective listening tests.

References

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *Proc. of the ICASSP* , Vol. 10, Apr. 1985, pp. 937-940.

[2] T. Houtgast, H.J.M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics," *Acoustica 46*, pp. 60-72, 1980.

[3] IEC 60268-16: "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index", <*http://www.iec.ch/*>

[4] B. E. D. Kingsbury, N. Morgan, S. Greenberg,"Robust speech recognition using the modulation spectrogram",*Speech Communication*, Vol. 25 , Issue 1-3, Aug. 1998, pp. 117-132.

[5] M. Athineos, D. Ellis, "Frequency-domain linear prediction for temporal features", *IEEE Workshop on Automatic Speech Recognition and Understanding* , pp. 261-266, December 2003.

[6] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," *in Proc. ICASSP*, Vol.5, pp. 3277-3280, April 2001.

[7] T.H. Falk, S. Stadler, W.B. Kleijn, Wai-Yip Chan, "Noise Suppression Based on Extending a Speech-Dominated Modulation Band", *Interspeech 2007*, August 2007.

[8] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", *Journal of Acoustical Society of America*, vol 105, no 3, pp. 1912-1924, Mar. 1999.

[9] R. Kumerasan,"An inverse signal approach to computing the envelope of a real valued signal", *IEEE Signal Processing Letters*, Vol 5, Issue 10, pp 256-259, Oct. 1998.

[10] J. Herre and J.D Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)," in *Proc. of 101st AES Conv.*, Los Angeles, USA, pp. 1-24, 1996.

[11] LAME MP3 codec: http://lame.sourceforge.net

[12] 3GPP TS 26.401: Enhanced aacPlus general audio codec; General Description.

[13] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[14] P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Frequency Domain Linear Prediction for QMF Sub-bands and Applications to Audio coding", Proc. of MLMI 2007, LNCS Series, Springer-Verlag, Berlin, 2007.

[15] L.S. Marple, "Computing the Discrete-Time Analytic Signal via FFT", in *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. 47, pp. 2600-2603, 1999.

[16] A.H. Nuttal and E. Bedrosian, "On the Quadrature Approximation to the Hilbert Transform of modulated signals", *Proc. IEEE*, Vol. 54 (10), pp. 1458-1459, Oct. 1966.

[17] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis ", *IEEE Trans. on Signal Proc.*, Vol. 41, Issue 10, pp 3024-3051, Oct. 1993.

[18] P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Non-uniform QMF Decomposition for Wide-band Audio Coding based on Frequency Domain Linear Prediction ", *Tech. Rep., IDIAP*, RR 07-43, 2007.

[19] Walt Jesteadt, Sid P. Bacon, and James R. Lehman, "Forward Masking as a function of frequency, masker level, and signal delay", *Journal of Acoustical Society of America*, Vol. 71(4), pp. 950-962, April 1982.

[20] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri, "Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction", *to appear in Proc. ICASSP*, April 2008.

[21] P. Motlicek, S. Ganapathy, H. Hermansky, and H. Garudadri, "Scalable Wide-band Audio Codec based on Frequency Domain Linear Prediction", *Tech. Rep., IDIAP*, RR 07-16, version 2, September 2007.

[22] ISO/IEC JTC1/SC29/WG11, "Framework for Exploration of Speech and Audio Coding", *MPEG2007/N9254*, July 2007, Lausanne, CH.

[23] Martin Dietz, Lars Liljeryd, Kristofer Kjorling and Oliver Kunz, "Spectral Band Replication, a novel approach in audio coding", *Proc. of 112th AES Conv.*, May 2002.

[24] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding", J. Audio Eng. Soc., vol. 45, no. 10, pp. 789-814, October 1997.

[25] ITU-R Recommendation BS.1387, "Method for objective psychoacoustic model based on PEAQ to perceptual audio measurements of perceived audio quality", December 1998.

[26] ITU-R Recommendation BS.1534: "Method for the subjective assessment of intermediate audio quality", June 2001.