Audio Engineering Society

Convention Paper

Presented at the 127th Convention
2009 October 9–12 New York NY, USA

MDCT for Encoding Residual Signals in Frequency Domain Linear Prediction

Sriram Ganapathy[1], Petr Motlicek[2], and Hynek Hermansky[1,3*]

[1] *Department of Electrical and Computer Engineering, Johns Hopkins University, USA*

[2] *Idiap Research Institute, Martigny, Switzerland*

[3] *Human Language Technology Center of Excellence, Johns Hopkins University, USA*

Correspondence should be addressed to Sriram Ganapathy (ganapathy@jhu.edu)

ABSTRACT

Frequency domain linear prediction (FDLP) uses autoregressive models to represent Hilbert envelopes of relatively long segments of speech/audio signals. Although the basic FDLP audio codec achieves good quality of the reconstructed signal at high bit-rates, there is a need for scaling to lower bit-rates without degrading the reconstruction quality. Here, we present a method for improving the compression efficiency of the FDLP codec by the application of the modified discrete cosine transform (MDCT) for encoding the FDLP residual signals. In the subjective and objective quality evaluations, the proposed FDLP codec provides competent quality of reconstructed signal compared to the state-of-the-art audio codecs for the $32 - 64$ kbps range.

1. INTRODUCTION

Conventional approaches to speech coding achieve signal 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.

A new speech/audio coding technique based on modeling the temporal evolution of the spectral dy-

*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.

namics was proposed in [2, 3]. This technique is based on representing amplitude modulating (AM) signal using the Hilbert envelope estimate and frequency modulating (FM) signal using the Hilbert carrier. The technique exploits the predictability of slowly varying amplitude modulations for encoding speech/audio signals. Input signals are analyzed using a non-uniform quadrature mirror filter (QMF) bank to decompose the signal into frequency sub-bands. For each sub-band signal, Hilbert envelopes are estimated using frequency domain linear prediction (FDLP), which is an efficient technique for autoregressive (AR) modeling of the temporal envelopes of a signal [4]. The parameters of the AR model are transmitted to the decoder. The FDLP residual signals are transformed using discrete Fourier transform (DFT) and the magnitude and phase components are quantized separately [3]. At the decoder, these steps are inverted to reconstruct the signal back.

The base-line FDLP codec provides good reconstruction signal quality at high bit-rates ~ 66 kbps. However, there is strong requirement for scaling to lower bit-rates while meeting the reconstruction quality constraints similar to those provided by the state-of-art codecs. The simple encoding set-up of using a DFT based processing for the FDLP residual signal ([3]) offers little freedom in reducing the bit-rates. This is mainly due to the fact that small quantization errors in DFT phase components of the sub-band FDLP residual signals (which consume 60 % of the bit-rate) give rise to significant coding artifacts in the reconstructed signal.

In this paper, we propose an encoding scheme for the FDLP residual signals using modified discrete cosine transform (MDCT). The MDCT, proposed in [5], outputs a set of critically sampled transform domain coefficients. Perfect reconstruction is provided by time domain alias cancellation and the overlapped nature of the transform. All these properties make the MDCT a potential candidate for application in many popular audio coding systems (for example Advanced Audio Coding (AAC) [6]).

In the proposed FDLP codec, MDCT is applied on short segments (50 ms) of the FDLP residual signals in each sub-band. These coefficients are vector quantized (VQ) and transmitted to the receiver along with the parameters of AR model. At the decoder, the MDCT coefficients of the residual are inverse transformed and are used to modulate the FDLP envelope for reconstructing the sub-band signal. Bit-rate scalability is provided by altering the number of VQ levels. The current version of the codec provides high-fidelity audio compression for speech/audio content operating in the bit-rate range of $32 - 64$ kbps. In the objective and subjective quality evaluations, the proposed FDLP codec provides competitive results compared to the state-of-art codecs at similar bit-rates.

The rest of the paper is organized as follows. Sec. 2 describes the FDLP technique for AR modelling of AM Envelopes. The basic structure of the proposed FDLP codec is described in Sec. 3. The objective and subjective evaluations are reported in Sec. 4.

2. AUTOREGRESSIVE MODELLING OF AM ENVELOPES

Autoregressive (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) [7]. The duality between the time and frequency domains means that AR modelling can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples [8]. 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 described below. These relations form the basis for the autoregressive modelling of AM 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

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

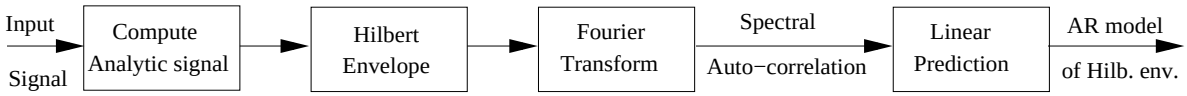

Fig. 1: Steps involved in deriving the autoregressive model of AM envelope.

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 Hilbert envelope of the signal and the auto-correlation in the spectral domain form Fourier transform pairs [9]. In a manner similar to the computation of the time domain 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 similar to those in [7]).

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 [10]. The input signal is transformed using DFT and the DFT sequence is made causal. The application of inverse DFT to the causal spectral representation gives the analytic signal $c[n]$ [10].

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

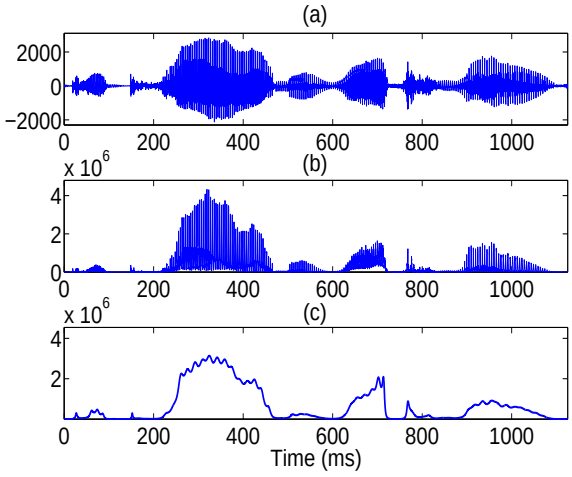

Fig. 2: Illustration of the AR modelling property of FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) all-pole model obtained using FDLP.

FDLP envelope is made causal. This step of generating a real valued spectral auto-correlation function is done for simplicity in the computation, although, the linear prediction can be done equally well for complex valued signals [8]. The remaining steps given in Fig. 1 follow the mathematical relations described previously.

2.2. FDLP Based AM-FM Decomposition

As the conventional AR models are used effectively on signals with spectral peaks, the AR models of the temporal envelope are appropriate for signals with peaky temporal envelopes [8, 11]. The individual poles in the resulting polynomial are directly associated with specific energy maxima in the time domain waveform. For signals that are expected to consist of a fixed number of distinct energy peaks in a given time interval, the AR model could well approximate these perceptually dominant peaks and

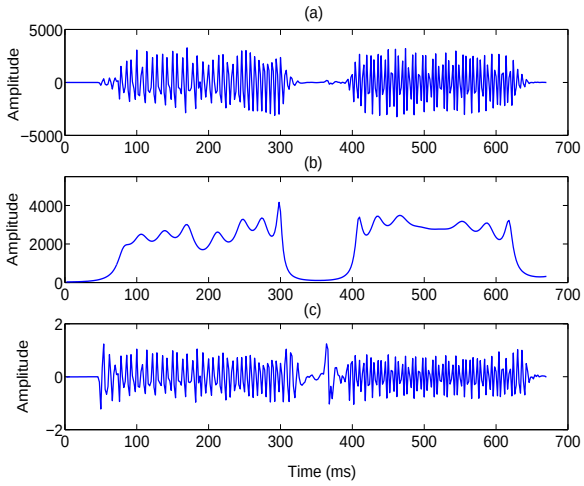

Fig. 3: Illustration of AM-FM decomposition using FDLP. (a) a portion of band pass filtered speech signal, (b) its AM envelope estimated using FDLP and (c) the FDLP residual containing the FM component.

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 computed from the analytic signal [10] and the AR model fit to the Hilbert envelope using FDLP.

For many modulated signals in the real world, the quadrature version of a real input signal and its Hilbert transform are identical [12]. This means that the Hilbert envelope is the squared AM envelope of the signal. The operation of FDLP estimates the AM envelope of the signal and the FDLP residual contains the FM component of the signal [8]. The FDLP technique consists of two steps. In the first step, the envelope of the signal is approximated with an AR model by using the linear prediction in the spectral domain. The resulting residual signal is obtained using the original signal and the AR model of the envelope obtained in the first step [8].

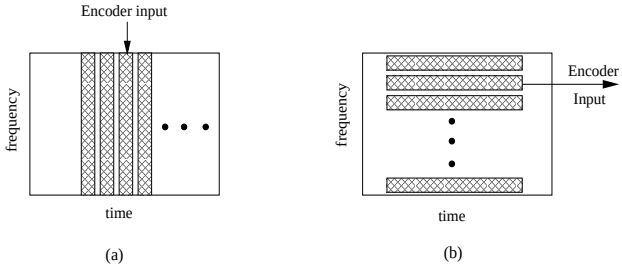

Fig. 4: Overview of time-frequency energy representation for (a) conventional codecs and (b) proposed FDLP codec.

This forms a parametric approach to AM-FM decomposition of a signal. In this paper, we extend the parametric AM-FM decomposition for the task of wide-band audio coding.

Speech signals in sub-bands are modulated signals [13] and hence, FDLP technique can be used for AM-FM decomposition of sub-band signals. An illustration of the AM-FM decomposition using FDLP is shown in Fig. 3, where we plot a portion of band pass filtered speech signal, its AM envelope estimate obtained as the square root of FDLP envelope and the FDLP residual signal representing the FM component of the band limited speech signal.

2.3. Time Frequency Signal Representation

For the proposed codec, the representation of signal information in the time-frequency domain is dual to that in the conventional codecs (Fig. 4). The state-of-the-art audio codecs (for example AAC [9]) encode 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 recreating the individual time frames. In the proposed FDLP codec, relatively long temporal segments of the signal (typically of the order hundreds of ms) are processed in narrow sub-bands (which emulate the critical band decomposition in human auditory system). At the decoder, the signal reconstruction is achieved by recreating the individual sub-bands signals which is followed by a sub-band synthesis.

3. SPEECH/AUDIO CODEC BASED ON FDLP

The block schematic of the FDLP based encoder is

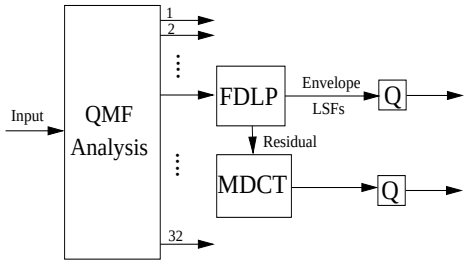

Fig. 5: Scheme of the FDLP encoder.

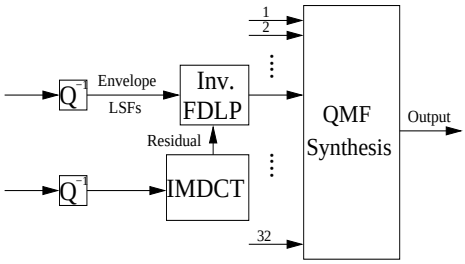

Fig. 6: Scheme of the FDLP decoder.

given in Fig. 5.

3.1. FDLP Analysis

Long temporal segments (1000 ms) of the input speech/audio signals are decomposed into 32 non-uniform QMF sub-bands which approximate the critical band decomposition in the auditory system. In each sub-band, the FDLP analysis is applied to obtain a set of AR model parameters and the FDLP residual signal. The FDLP envelope coefficients are converted to line spectral frequencies (LSF) which approximate the sub-band temporal envelopes. These LSF parameters are quantized using vector quantization (VQ).

3.2. Encoding FDLP Residual Signals Using MDCT

The sub-band FDLP residual signals are split into relatively short frames (50 ms) and transformed using the MDCT. We use the sine window with 50 % overlap for the MDCT analysis as this was experimentally found to provide the best reconstruction quality (based on objective quality scores). Since a full-search VQ in the MDCT domain with good resolution would be computationally infeasible, the split VQ approach is employed. Although the split VQ approach is suboptimal, it reduces the computational complexity and memory requirements to manageable limits without severely degrading the VQ performance. The VQ codebooks are trained on a large audio database using the LBG algorithm. The quantized levels are Huffman encoded for further reduction of bit-rates (bit-rate reduction of about 10 %). Quantization of the MDCT coefficients using the split VQ consumes around 80% of the bit-rate. The MDCT coefficients for the lower frequency sub-bands are quantized using higher number of VQ levels compared to those from the higher bands. For the purpose of scaling the bit-rates, all the sub-bands are treated uniformly. The current version of the codec follows a simple signal independent bit assignment for the MDCT coefficients.

3.3. Decoder

In the decoder, shown in Fig. 6, quantized MDCT coefficients of the FDLP residual signals are reconstructed and transformed back to the time-domain using inverse MDCT (IMDCT). The reconstructed FDLP envelopes (from LSF parameters) are used to modulate the corresponding sub-band residual signals. Finally, sub-band synthesis is applied to reconstruct the full-band signal.

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 [14]. This database 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 the following codecs are considered:

1. The proposed FDLP codec with MDCT based residual signal processing, at 32, 48 and 64 kbps, denoted as FDLP.

2. The previous version of the FDLP codec [3], at 66 kbps, denoted as FDLP-DFT.

bit-rate [kbps]	64	64	66	64
Codec	LAME	AAC	FDLP-DFT	FDLP
PEAQ	-1.6	-0.8	-1.2	-0.7
bit-rate [kbps]	48	48	48	48
Codec	LAME	AAC	FDLP-DFT	FDLP
PEAQ	-2.5	-1.1	-2.5	-1.2
bit-rate [kbps]	32	32	32	32
Codec	LAME	AAC	AMR	FDLP
PEAQ	-3.0	-2.4	-2.2	-2.4

Table 1: Average PEAQ scores for 28 speech/audio files at 64, 48 and 32 kbps.

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

Table 2: PEAQ scores and their meanings.

3. LAME MP3 (MPEG 1, layer 3) [15], at 32, 48 and 64, kbps denoted as LAME.

4. MPEG-4 HE-AAC, v1, at 32, 48 and 64 kbps [6], denoted as AAC. The HE-AAC coder is the combination of spectral band replication (SBR) [16] and advanced audio coding (AAC) [17].

5. AMR-WB plus standard [18], at 32 kbps, denoted as AMR.

4.1. Objective Evaluations

The objective measure employed is the perceptual evaluation of audio quality (PEAQ) distortion measure [19]. 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 2. The mean PEAQ score for the 28 speech/audio files in [14] is used as the objective quality measure.

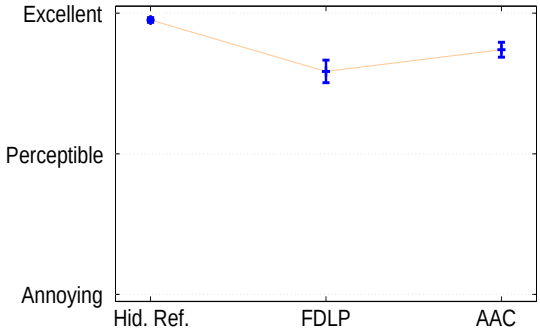

Fig. 7: BS.1116 results for 5 speech/audio samples using two coded versions at 64 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC)), hidden reference (Hid. Ref.).

The first set of results given in Table 1 compare the objective quality scores of the proposed FDLP codec at 64 kbps with the FDLP-DFT codec at 66 kbps. The objective quality scores for AAC and LAME codecs at 64 kbps are also shown. This table shows the advantage of using the MDCT for encoding the FDLP residuals instead of using the DFT.

The next set of results in Table 1 show the average PEAQ scores for the proposed FDLP codec with AAC and LAME codecs at 48 kbps and the scores for these codecs along with the AMR codec at 32 kbps. The objective scores for the proposed FDLP codec at these bit-rates follow a similar trend to that of the state-of-the-codecs.

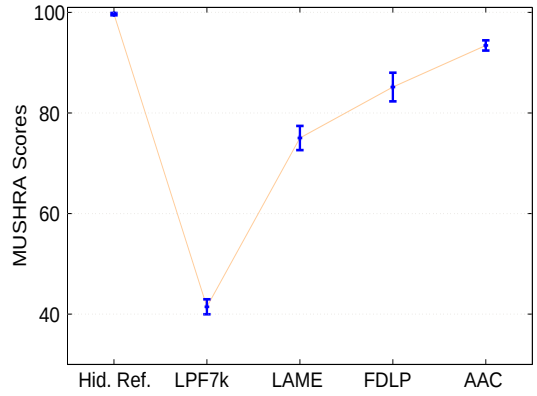

Fig. 8: MUSHRA results for 6 speech/audio samples using three coded versions at 48 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k).

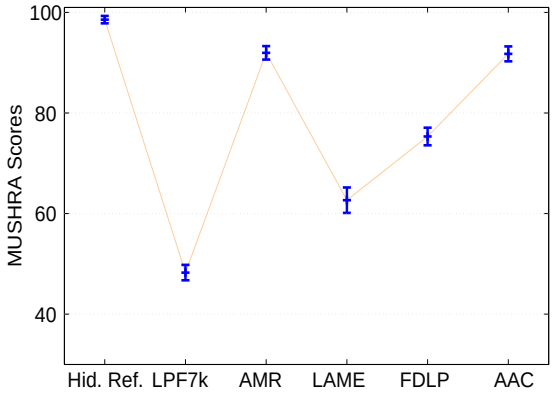

Fig. 9: MUSHRA results for 6 speech/audio samples using four coded versions at 32 kbps (AMR-WB+ (AMR), FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k).

4.2. Subjective Evaluations

Since the encoded audio signals at 64 kbps have small impairments compared to the original, we perform the BS.1116 methodology of subjective evaluation [20]. As this subjective evaluation is time consuming, only two coded versions (FDLP and AAC) are compared at 64 kbps along with the hidden reference. The subjective results with 7 listeners using 5 speech/audio samples from the database is shown in Fig. 7. Here, the mean scores are plotted with 95% confidence interval. The proposed FDLP codec at 64 kbps is judged to be similar to the AAC codec at the same bit-rate.

For the audio signals encoded at 48 kbps and 32 kbps, the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) methodology for subjective evaluation is employed. It is defined by ITU-R recommendation BS.1534 [21]. We perform the MUSHRA tests on 6 speech/audio samples from the database with 5 listeners. The mean MUSHRA scores (with 95% confidence interval) for the subjective listening tests at 48 kbps and 32 kbps (given in Fig. 8 and Fig. 9 respectively) show that the subjective quality of the proposed codec is slightly poorer than AAC codec but better than LAME codec.

5. CONCLUSIONS

In order to improve the compression efficiency of audio codecs based on spectral dynamics, we propose a new method of encoding the FDLP residual signals by the application of MDCT. This new technique offers the advantage of bit-rate scalability similar to the state-of-the-art codecs. Objective evaluations justify the improvement provided by the use of MDCT as compared to the use of DFT in the previous versions of the FDLP codec. The current version of the codec provides subjective results which are competitive to the state of the art codecs in the bit-rate range of 32-64 kbps. Furthermore, this performance is achieved without utilizing standard modules like psycho-acoustic modelling and signal adaptive windowing. The inclusion of these techniques form part of the future work.

ACKNOWLEDGEMENTS

The authors would like to thank Harinath Garudadri for his active involvement during the development of the codec and Marios Athineos for MDCT code fragments.

6. REFERENCES

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

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

[3] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri, "Autoregressive Modelling of Hilbert Envelopes for Wide-band Audio Coding", *Audio Engg. Soc.*, 124th Convention, May 2008.

[4] M. Athineos, and D. Ellis, "Autoregressive Modeling of Temporal Envelopes", *IEEE Trans. on Signal Proc.*, Vol. 55, pp. 5237 - 5245, Nov. 2007.

[5] J. Princen, A. Johnson and A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", *Proc. of ICASSP*, Vol. 87, pp 2161-2164, May 1987.

[6] 3GPP TS 26.401: Enhanced aacPlus general audio codec; General Description.

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

[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] J. Herre and J.D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)", *Audio Engg. Soc.*, 101st Convention, pp. 1-24, 1996.

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

[11] A. Rao and R. Kumaresan, "A parametric modeling approach to Hilbert transformation," *IEEE Sig. Proc. Letters*, Vol.5, No.1, Jan. 1998.

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

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

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

[15] LAME MP3 codec: *<http://lame.sourceforge.net>*

[16] M. Dietz, L. Liljeryd, K. Kjorling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding", *Audio Eng. Soc.* , 112th Convention, May 2002.

[17] 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, pp. 789-814, Oct. 1997.

[18] "Extended AMR Wideband codec", *<http://www.3gpp.org/ftp/Specs/html-info/26290.htm>*

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

[20] ITU-R Recommendation BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", Oct. 1997.

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