Autoregressive Models of Amplitude Modulations in Audio Compression

Sriram Ganapathy*, *Student Member, IEEE,* Petr Motlicek, *Member, IEEE,*
Hynek Hermansky *Fellow, IEEE*

Abstract

We present a scalable medium bit-rate wide-band audio coding technique based on frequency domain linear prediction (FDLP). FDLP is an efficient method for representing the long-term amplitude modulations of speech/audio signals using autoregressive models. For the proposed audio codec, relatively long temporal segments (1000 ms) of the input audio signal are decomposed into a set of critically sampled sub-bands using a quadrature mirror filter (QMF) bank. The technique of FDLP is applied on each sub-band to model the sub-band temporal envelopes. The residual of the linear prediction, which represents the frequency modulations in the sub-band signal [1], are encoded and transmitted along with the envelope parameters. These steps are reversed at the decoder to reconstruct the signal. The proposed codec utilizes a simple signal independent non-adaptive compression mechanism for a wide class of speech and audio signals. The subjective and objective quality evaluations show that the reconstruction signal quality for the proposed FDLP codec compares well with the state-of-the-art audio codecs in the 32-64 kbps range.

Index Terms

Speech and Audio coding, Modulation Spectrum, Frequency Domain Linear Prediction (FDLP), Objective and subjective evaluation of audio quality.

EDICS: SPE-ANLS, AUD-ANSY, AUD-ACOD.

S. Ganapathy and H. Hermansky are with the ECE Dept., Johns Hopkins University, Baltimore, USA, (phone: +1-410-516-7031; fax +1-410-516-5566; email: {ganapathy, hynek}@jhu.edu).

P. Motlicek is with the Idiap Research Institute, Martigny, Switzerland, (phone: +41-277-217-749; fax +41-277-217-712; email: motlicek@idiap.ch.)

I. Introduction

Demanded by new audio services, there has been new initiatives in standardization organizations like 3GPP, ITU-T, and MPEG (for example [2]) that aim for the development of a unified codec which can efficiently compress all kinds of speech and audio signals and which may require new audio compression techniques. Conventional approaches to speech coding are developed around a linear source-filter model of the speech production using the linear prediction (LP) [3]. The residual of this modelling 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 multimedia coding applications (for example [4], [5]) are not as efficient for speech content.

In traditional applications of speech coding (i.e., for conversational services), the algorithmic delay of the codec is one of the most critical variables. However, there are many services, such as audio file downloads, voice messaging etc., where the issue of the codec delay is much less critical. This allows for a whole set of different analysis and compression techniques that could be more effective than the conventional short-term frame based coding techniques.

In this paper, we describe a technique, which employs 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 [6] for predicting speech intelligibility and characterizing room acoustics are now widely used in the industry [7]. Recently, there has been many applications of such concepts for robust speech recognition [8], [9], [10], [11], audio coding [12] and noise suppression [13].

In this paper, the approach to audio compression 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 linear prediction in spectral domain [1], [14], [15], which is an efficient technique for autoregressive modelling of the temporal envelopes of a signal. For audio coding applications, frequency domain linear prediction (FDLP) is performed on real spectral representations using symmetric extensions [11], [16], [17]. This idea was first applied for audio coding in the MPEG2-AAC (advanced audio coding) [18], where it was primarily used for removing pre-echo artifacts. The proposed codec employs FDLP for an entirely different purpose. We use FDLP to model relatively long (1000 milliseconds) segments of AM envelopes in sub-bands. A non-uniform quadrature mirror filter (QMF) bank is used to derive 32 critically sampled frequency sub-bands. This non-uniform QMF analysis emulates the critical band decomposition observed in the human auditory system. FDLP is applied on these sub-band signals to estimate the sub-band Hilbert envelopes. The remaining residual signal (Hilbert carrier) is further processed using the modified discrete cosine transform (MDCT) and the transform components are quantized, entropy coded and transmitted. At the decoder, the sub-band signal is reconstructed by modulating the inverse quantized Hilbert carrier with the AM envelope. This is followed by a QMF synthesis to obtain the audio signal back.

The main goal of this paper is to illustrate the use of FDLP based signal analysis technique for purpose of wide-band audio coding using a simple compression scheme. In this regard, the proposed codec does not use any psycho-acoustic models or signal dependent windowing techniques and employs relatively unsophisticated quantization methodologies. 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. The proposed codec is evaluated using the speech/audio samples provided by MPEG for the development of unified speech and audio codec [2], [19]. In the objective and subjective quality evaluations, the proposed FDLP codec provides competitive results compared to the state-of-the-art codecs at similar bit-rates.

The rest of the paper is organized as follows. Section II provides a mathematical description for the autoregressive modelling of AM envelopes using the FDLP technique. The various blocks in the proposed codec are described in Section III. The results of the objective and subjective evaluations are reported in Section IV. This followed by a summary in Section V.

II. Autoregressive modelling of the 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) [20]. 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 [1], [15]. 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.

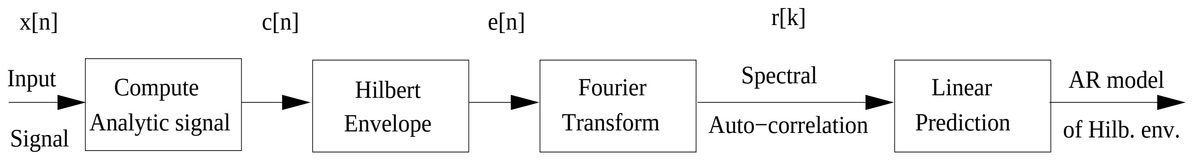

Fig. 1. Steps involved in deriving the autoregressive model of AM envelope.

A. 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 Hilbert envelope of the signal and the auto-correlation in the spectral domain form Fourier transform pairs [18]. 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 [20]).

The block schematic showing the steps involved in deriving the AR model of Hilbert envelope is shown in figure 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 [21]. 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 $c[n]$ [21].

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

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

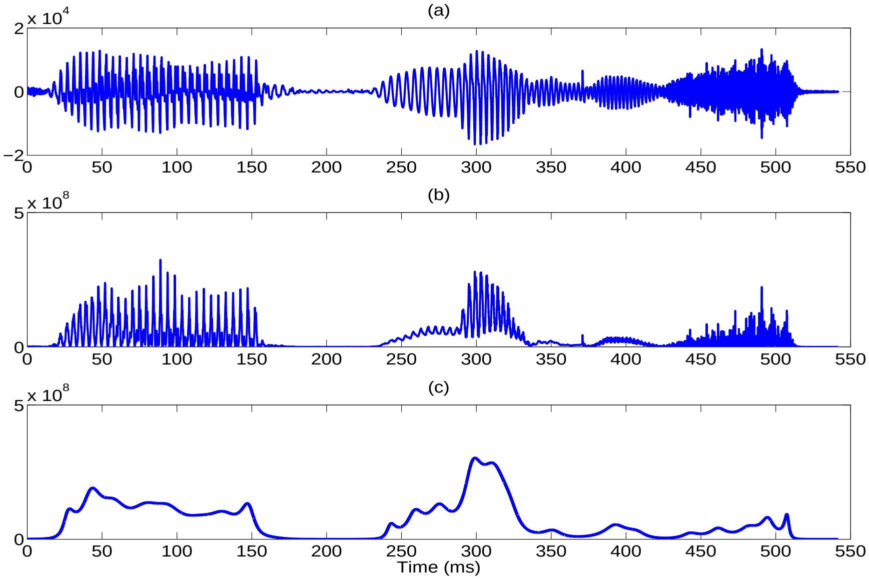

Fig. 2. Illustration of the AR modelling property of FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) all pole model obtained using FDLP.

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. 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 [1]. The remaining steps given in figure 1 follow the mathematical relations described previously.

B. 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 [1], [14], [15]. 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 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

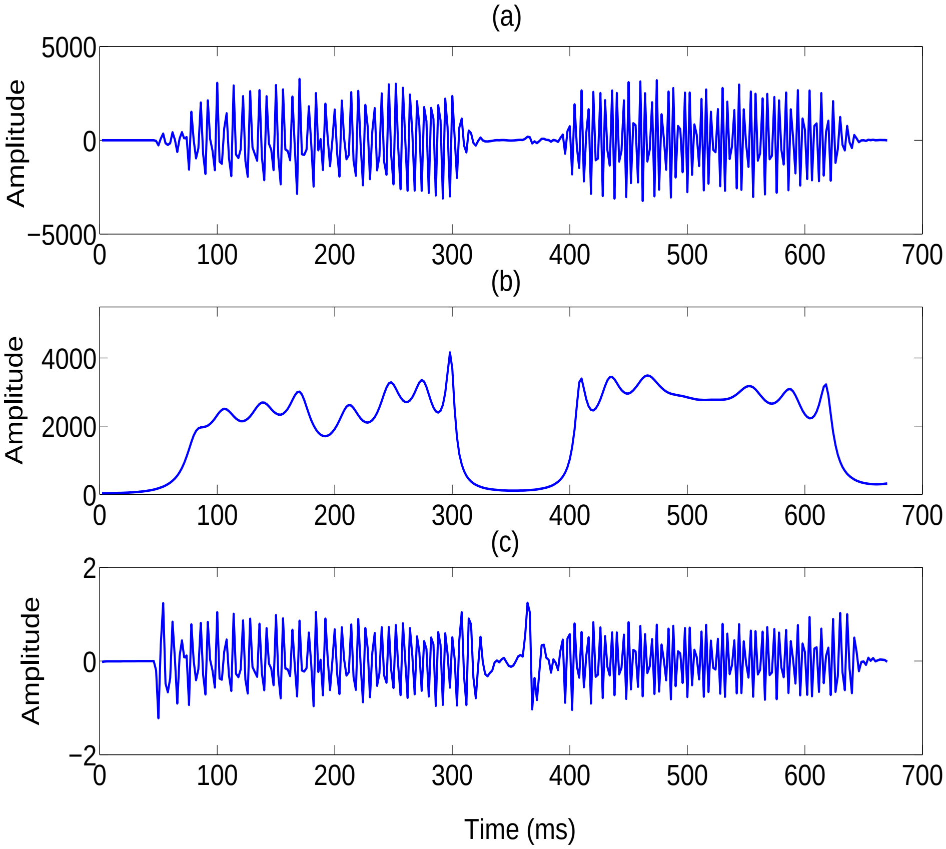

Fig. 3. Illustration of AM-FM decomposition using FDLP. (a) a portion of band pass filtered speech signal, (b) its AM envelope estimated using FDLP and (c) the FDLP residual containing the FM component.

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 figure 2, where we plot a portion of speech signal, its Hilbert envelope computed from the analytic signal [21] 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 [22]. 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 [1]. 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 [1]. This forms a parametric approach to AM-FM decomposition of

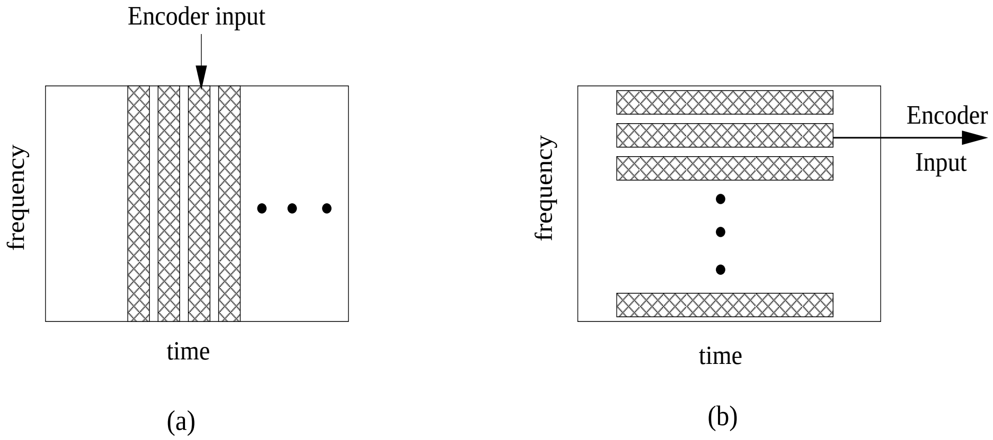

Fig. 4. Overview of time-frequency energy representation for (a) conventional codecs and (b) proposed FDLP codec.

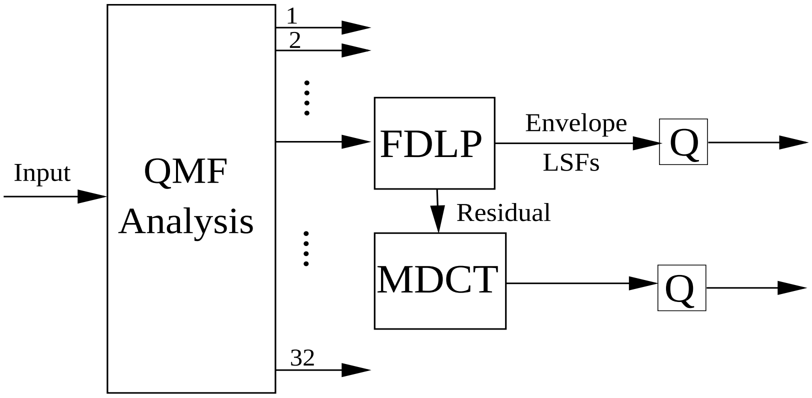

Fig. 5. Scheme of the FDLP encoder.

a signal. There are other non-parametric approaches for AM-FM decomposition using these results [23]. 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 [24] 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 figure 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.

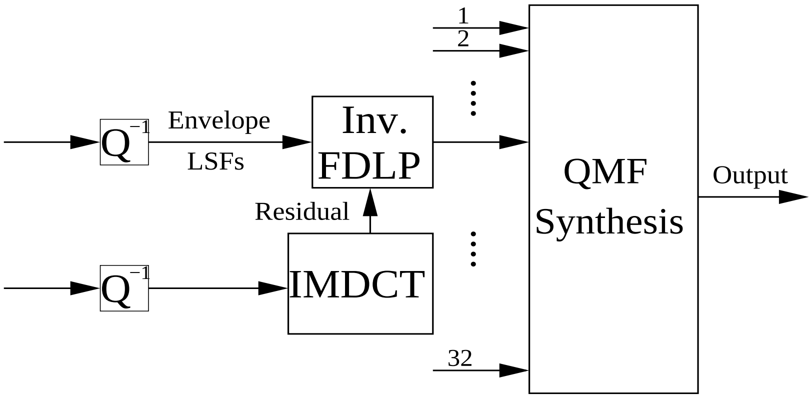

Fig. 6. Scheme of the FDLP decoder.

C. 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 (figure 4). The state-of-the-art audio codecs (for example AAC [18]) 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.

III. 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 a set of prediction coefficients is obtained using the Levinson-Durbin recursion [25]. These prediction coefficients are converted to envelope line spectral frequencies (LSFs) (in a manner similar to the conversion of TDLP coefficients to LSF parameters). The envelope LSFs represent the location of the poles on the temporal domain. Specifically, the envelope LSFs take values in the range of $(0, 2\pi)$ radians corresponding to temporal locations in the range of $(0, 1000$ ms) of the sub-band signal. Thus, the angles of poles of the FDLP model indicate the timing of the peaks of the signal [16].

In each sub-band, these LSFs approximating the sub-band temporal envelopes are quantized using vector quantization (VQ). The residual signals (sub-band Hilbert carrier signals) are processed in transform

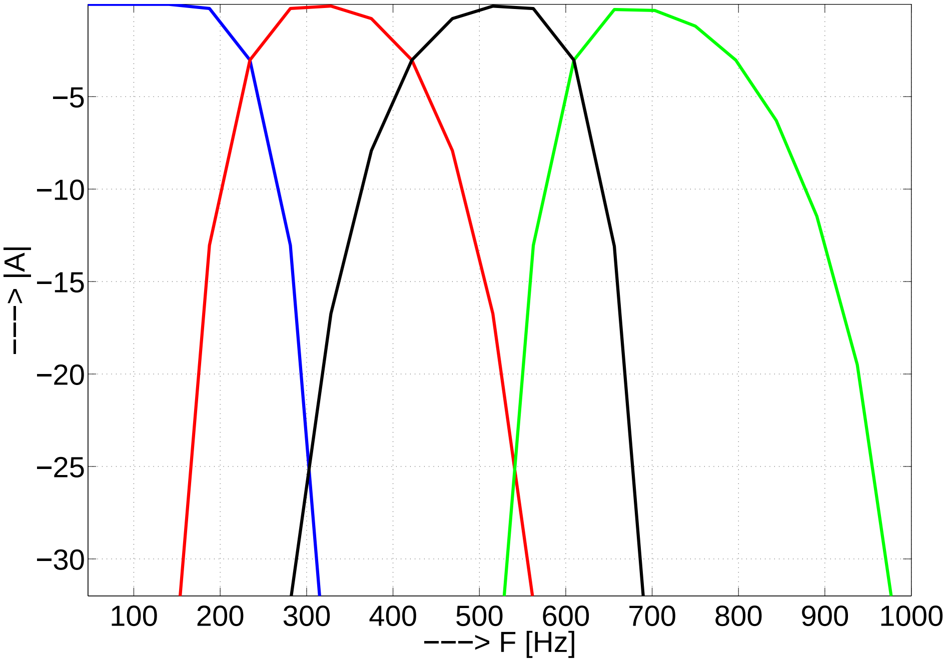

Fig. 7. Magnitude frequency response of first four QMF bank filters.

domain using the modified discrete cosine transform (MDCT). The MDCT coefficients are also quantized using VQ. Graphical scheme of the FDLP encoder is given in figure 5.

In the decoder, shown in figure 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 (obtained 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.

The important blocks are:

A. Non-uniform sub-band decomposition

A non-uniform quadrature mirror filter (QMF) bank is used for the sub-band decomposition of the input audio signal. QMF provides sub-band sequences which form a critically sampled and maximally decimated signal representation (i.e., the total number of sub-band samples is equal to the number of input samples). In the proposed non-uniform QMF analysis, the input audio signal (sampled at 48 kHz) is split into 1000 ms long frames. Each frame is decomposed using a 6 stage tree-structured uniform QMF analysis to provide 64 uniformly spaced sub-bands. A non-uniform QMF decomposition into 32 frequency sub-bands is obtained by merging these 64 uniform QMF sub-bands [26]. This sub-band decomposition is motivated by critical band decomposition in the human auditory system. Many uniformly spaced sub-bands at the higher auditory frequencies are merged together while maintaining perfect reconstruction. The non-uniform QMF decomposition provides a good compromise between fine spectral resolution for low frequency sub-bands and a smaller number of FDLP parameters for higher bands.

In order to reduce the leakage of quantization noise from one sub-band to another, the QMF analysis and synthesis filters are desired to have a sharp transition band. This would result in a significant delay for the QMF filter bank. Since we use an initial decomposition using a tree structured QMF filter bank, the overall filter bank delay can be considerably reduced by reducing the length of filters in the successive stages of the tree. Although the width of the transition band in the sub-sampled domain increases due to the reduction in filter length, the transition bandwidth at the original sampling rate remains the same [27]. The overall delay for the proposed QMF filter bank is about 30 ms. Magnitude frequency responses of first four QMF filters are given in figure 7.

B. Encoding FDLP residual signals using MDCT

In the previous version of the FDLP codec [28], the sub-band FDLP residual signals were transformed using DFT and the magnitude and phase components were quantized separately. Although this baseline FDLP codec provides good reconstruction signal quality at high bit-rates (66 kbps), there is strong requirement for scaling to lower bit-rates while meeting the reconstruction quality constraints similar to those provided by the state-of-the-art codecs. The simple encoding set-up of using a DFT based processing for the FDLP residual signal ([28]) offers little freedom in reducing the bit-rates. This is mainly due to the fact that small quantization errors in the DFT phase components of the sub-band FDLP residual signals (which consume 60 % of the final 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 MDCT. The MDCT, originally proposed in [29], 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) [30]).

For the proposed FDLP codec, 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

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

TABLE I

PEAQ SCORES AND THEIR MEANINGS.

objective quality evaluations). 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 quantized levels are Huffman encoded for further reduction of bit-rates. This entropy coding scheme results in a bit-rate reduction of about 10%. The MDCT coefficients for the lower frequency sub-bands are quantized using higher number of VQ levels as compared to those from the higher bands. VQ of the MDCT coefficients consumes about 80% of the final bit-rate.

For the purpose of scaling the bit-rates, all sub-bands are treated uniformly and the number of VQ levels are suitably modified so as to meet the specified bit-rate. The current version of the codec follows a simple signal independent bit assignment mechanism for the MDCT coefficients and provides bit-rate scalability in the range of 32-64 kbps.

IV. 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 [2], [19]. 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 mono and stereo versions of these audio samples were used for the recent low bit-rate evaluations of unified speech and audio codec [31].

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.

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 II

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

2) The previous version of the FDLP codec using DFT based residual signal processing [28], at 48 and 66 kbps, denoted as FDLP-DFT.
3) LAME MP3 (MPEG 1, layer 3) [33], at 32, 48 and 64, kbps denoted as LAME.
4) MPEG-4 HE-AAC, v1, at 32, 48 and 64 kbps [30], denoted as AAC. The HE-AAC coder is the combination of spectral band replication (SBR) [34] and advanced audio coding (AAC) [35].
5) AMR-WB plus standard [36], at 32 kbps, denoted as AMR.

A. Objective Evaluations

The objective measure employed is the perceptual evaluation of audio quality (PEAQ) distortion measure [32]. 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 I. The mean PEAQ score for the 28 speech/audio files from [19] is used as the objective quality measure.

The first two set of results given in table II compare the objective quality scores for the proposed FDLP codec at with the FDLP-DFT codec. These results show the advantage of using the MDCT for encoding the FDLP residuals instead of using the DFT. The results in table II also show the average PEAQ scores for the proposed FDLP codec, AAC and LAME codecs at 48 kbps and the scores for these

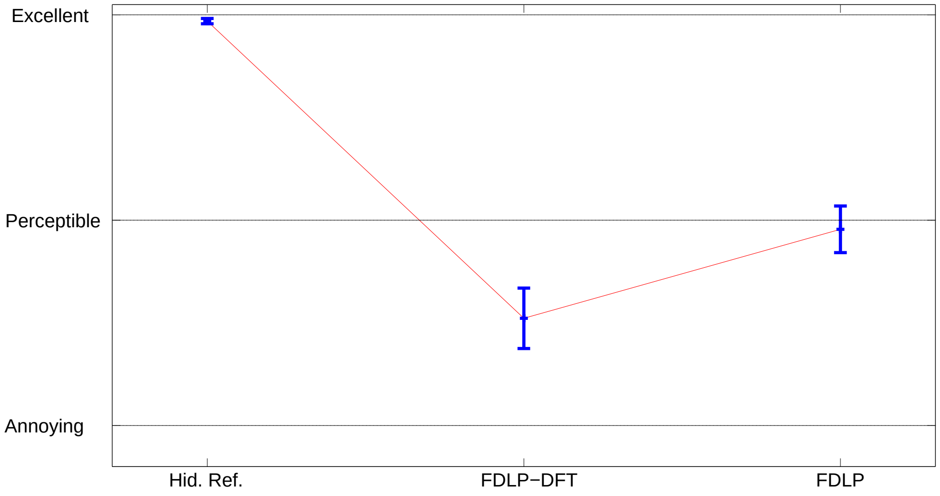

Fig. 8. BS.1116 results for 6 speech/audio samples using two coded versions at 48 kbps (FDLP-MDCT (FDLP), FDLP-DFT) and hidden reference (Hid. Ref.) with 9 listeners.

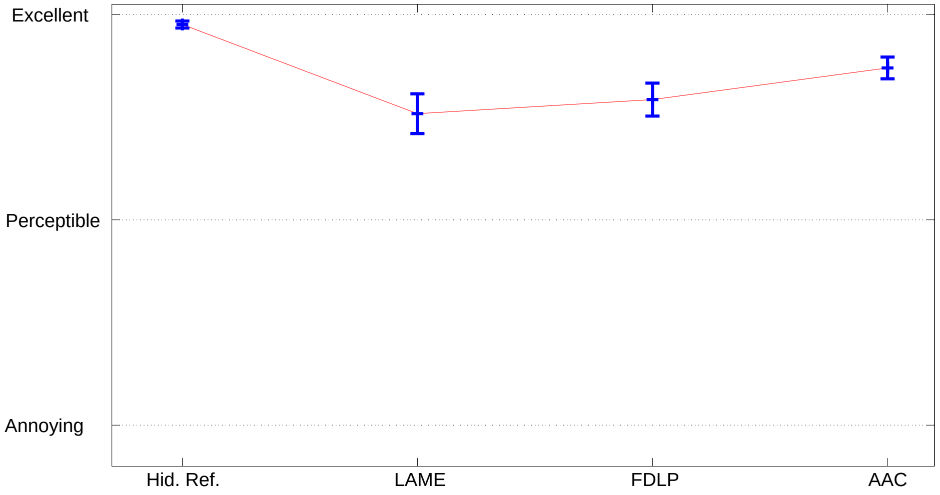

Fig. 9. BS.1116 results for 5 speech/audio samples using three coded versions at 64 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME MP3 (LAME)), hidden reference (Hid. Ref.) with 7 listeners.

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 compared to that of the state-of-the-art codecs.

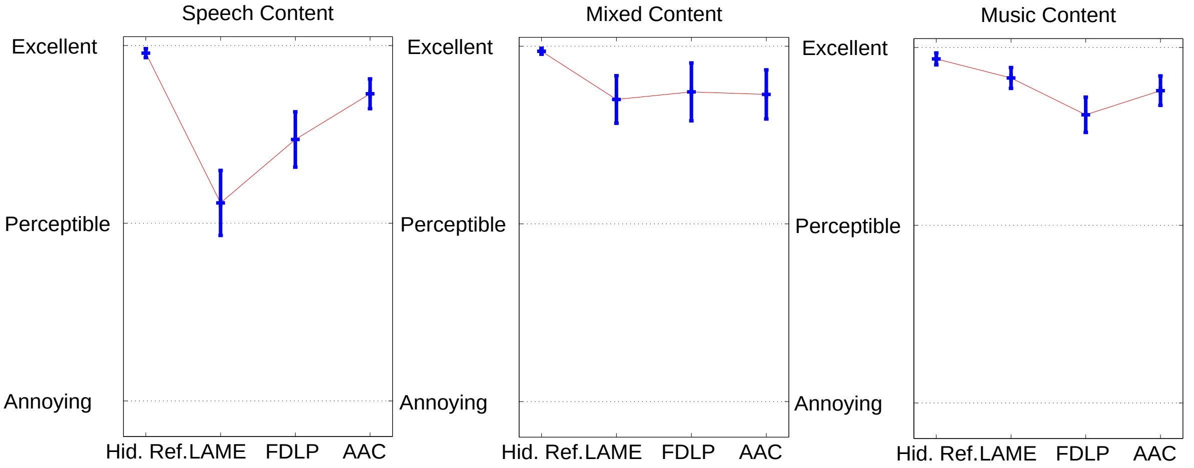

Fig. 10. BS.1116 results for each audio sample type namely speech, mixed and music content using three coded versions at 64 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME MP3 (LAME)), hidden reference (Hid. Ref.) with 7 listeners. Average results for all these audio samples are present in figure 9.

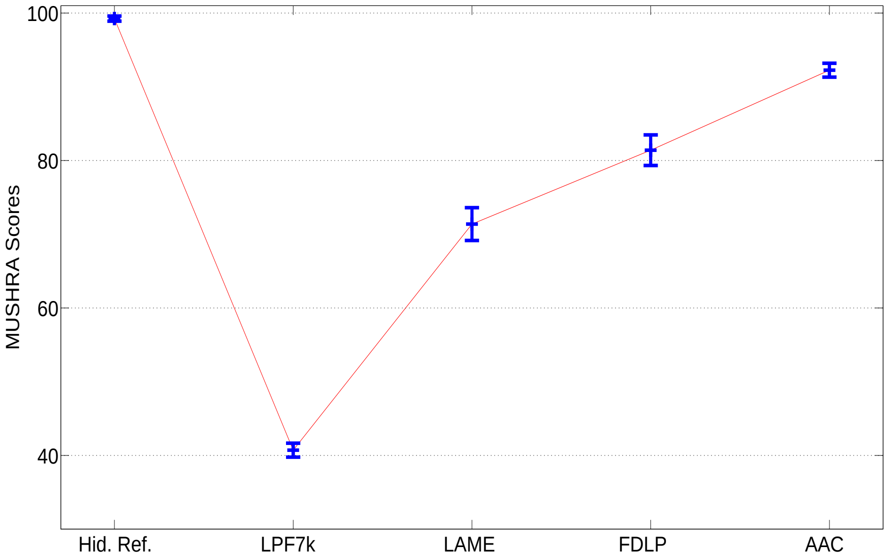

Fig. 11. MUSHRA results for 6 speech/audio samples and 8 listeners using three coded versions at 48 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k).

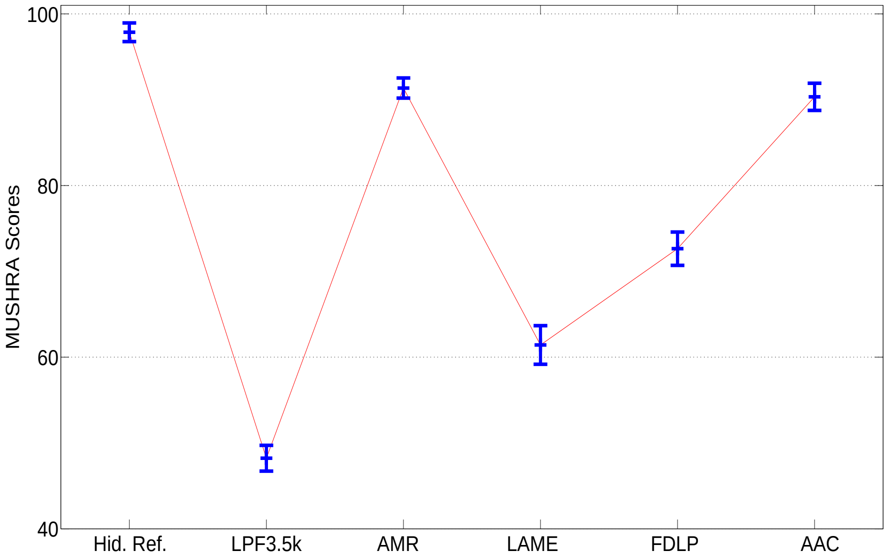

Fig. 12. MUSHRA results for 6 speech/audio samples and 6 listeners using four coded versions at 32 kbps (AMR-WB+ (AMR), FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 3.5 kHz low-pass filtered anchor (LPF3.5k).

B. Subjective Evaluations

The audio files chosen for the subjective evaluation consist of a subset of speech, music as well as mixed signals from the set of 28 audio samples given in [19]. The first set of experiments compare the proposed FDLP codec with the previous version of the codec which utilizes DFT based carrier processing [28]. We perform the BS.1116 methodology of subjective evaluation [37]. The results of the subjective evaluation with 6 speech/audio samples is shown in figure 8. These results show that the MDCT based residual processing is considerably better than the previous version of the FDLP codec. Furthermore, the MDCT processing simplifies the quantization and encoding step.

Since the MDCT processing of the FDLP carrier signal is found to be efficient, the rest of the subjective evaluations use the FDLP-MDCT configuration. We perform the BS.1116 methodology of subjective evaluation [37] for the comparisons of three coded versions (LAME, FDLP and AAC) at 64 kbps along with the hidden reference. The subjective evaluation results with 7 listeners using 5 speech/audio samples from the database is shown in figure 9. Here, the mean scores are plotted with 95% confidence interval. At 64 kbps, the proposed FDLP codec as well as the LAME and AAC codec, are subjectively judged to

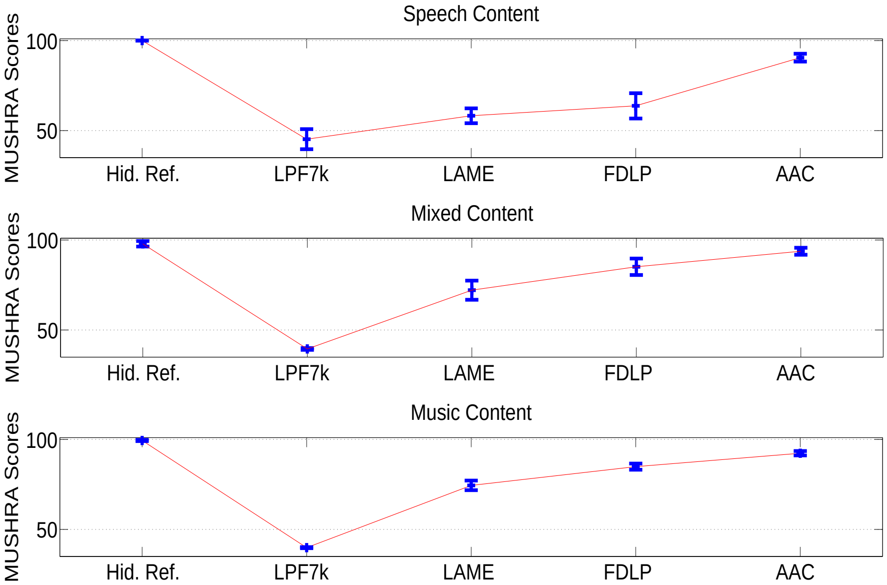

Fig. 13. MUSHRA results for each audio sample type namely speech, mixed and music content obtained using three coded versions at 48 kbps (FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 7 kHz low-pass filtered anchor (LPF7k) with 8 listeners. Average results for all these audio samples are present in figure 11.

have imperceptible noise content.

The subjective results for individual sample types (namely speech, mixed and music content) are shown in figure 10. These results show that the performance of the FDLP codec was better than the LAME codec for speech content and mixed content, whereas it was slightly worse for music content at 64 kbps. At 64 kbps, the proposed FDLP codec gives the best performance for mixed content and the performance for the speech content is the least among these audio sample types.

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 [38]. We perform the MUSHRA tests on 6 speech/audio samples from the database. The mean MUSHRA scores (with 95% confidence interval), for the subjective listening tests at 48 kbps and 32 kbps (given in figure 11 and figure 12, respectively), show that the subjective quality of the proposed codec is slightly poorer compared to the AAC codec but better than the LAME codec.

The subjective results for individual sample types (namely speech, mixed and music content) at 48 kbps and 32 kbps are shown in figure 13 and figure 14. For all the individual sample types, the performance of

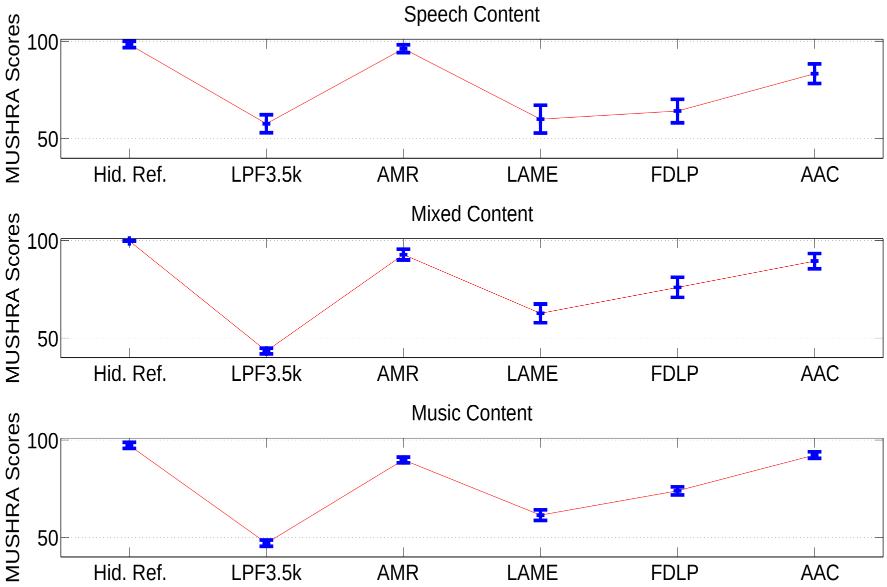

Fig. 14. MUSHRA results for each audio sample type namely speech, mixed and music content obtained using four coded versions at 32 kbps (AMR-WB+ (AMR), FDLP-MDCT (FDLP), MPEG-4 HE-AAC (AAC) and LAME-MP3 (LAME)), hidden reference (Hid. Ref.) and 3.5 kHz low-pass filtered anchor (LPF3.5k) with 6 listeners. Average results for all these audio samples are present in figure 12.

the FDLP codec is worser than the AAC codec but better than the LAME codec. The subjective scores are higher for the audio samples with music and mixed content compared to those with speech content.

V. Discussions and conclusions

A technique for autoregressive modelling of the AM envelopes is presented, which is employed for developing a wide-band audio codec operating in medium bit-rates. 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 adaptively captures fine temporal nuances with high temporal resolution while at the same time summarizes the spectrum in time scales of hundreds of milliseconds. The proposed compression scheme is relatively simple and suitable for coding speech, music and mixed signals.

Although the application of linear prediction in the transform domain is used in temporal noise shaping (TNS) [18], the proposed technique is fundamentally different from this approach. While the TNS tries to remove coding artifacts in transient signals in a conventional short-term transform codec like AAC [5], the proposed FDLP technique is able to model relatively long (hundreds of milliseconds) segments of AM envelopes in sub-bands. Specifically, the proposed codec exploits the AM-FM decomposition property of FDLP in the sub-bands of speech and audio signals.

The performance of the proposed codec is objectively evaluated using PEAQ distortion measure, standardized in ITU-R (BS.1387). Final performances of the FDLP codec, in comparison with other state-of-the-art codecs, at variety of bit-rates in $32 - 64$ kbps range, are also evaluated using subjective quality evaluation methodologies like MUSHRA and BS.1116, standardized in ITU-R (BS.1534 and BS.1116, respectively). The subjective evaluations suggest that the proposed wide-band FDLP codec provides perceptually better audio quality than LAME - MP3 codec and produces slightly worse results compared to MPEG-4 HE-AAC standard. Although the improvements are modest, the potential of the proposed analysis technique for encoding speech and audio signals is clearly illustrated by the quality evaluations.

The performance of the proposed codec is dependent on the efficient processing of the FDLP carrier signal. The MDCT based processing simplifies the codec design. The quantizer can be designed effectively for fixed length MDCT coefficients of the carrier signal. Furthermore, the objective and subjective quality evaluations show that the MDCT processing provides good improvements compared to the FDLP-DFT codec.

Furthermore, the proposed codec yields reconstruction signal quality comparable to that of the state-of-the-art codecs without using many additional techniques that are becoming standard in the conventional codecs. Specifically, neither SNRs in the individual sub-bands are evaluated nor signal dependent non-uniform quantization in different frequency sub-bands (e.g. module of simultaneous masking) or at different time instants (e.g. bit-reservoir) are employed. There are no signal dependent windowing techniques and the quantization scheme is relatively simple. Inclusion of some of these sophisticated bit-rate reduction techniques should further reduce the target bit-rates and enhance the bit-rate scalability. These form part of our future work.

ACKNOWLEDGMENT

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 (IM2)”; managed by the IDIAP Research Institute on behalf of the Swiss Federal Authorities. The authors would like to thank H. Garudadri for his active involvement during the development of the codec.

REFERENCES

[1] 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.

[2] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on Unified Speech and Audio Coding," Shenzhen, China, Oct. 2007, MPEG2007/N9519.

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

[4] K. Brandenburg, G. Stoll, Y. F. Dehery, J. D. Johnston, L. V. D. Kerkhof, and E. F. Schroeder, "The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio," *Audio Engg. Soc.*, 92nd Convention, Vienna, Austria, May 1992.

[5] J. Herre and J. M. Dietz, "MPEG-4 high-efficiency AAC coding", *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137 - 142, May 2008.

[6] 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.

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

[8] V. Tyagi, C. Wellekens, "Fepstrum representation of speech signal," *Proc. of the IEEE Workshop in Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, Dec. 2005.

[9] V. Tyagi, C. Wellekens, "Fepstrum: An Improved Modulation Spectrum for ASR Vivek Tyagi," *Proc. of Interspeech*, Belgium, Aug. 2007.

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

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

[12] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 5, pp. 3277-3280, Salt Lake City, USA, Apr. 2001.

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

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

[15] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal", *IEEE Sig. Proc. Letters*, Vol.5, No.10, Oct. 1998.

[16] M. Athineos and D. P. W. Ellis, "Autoregressive modelling of temporal envelopes," *IEEE Trans. Speech and Audio Processing*, Vol. 55, pp. 5237-5245, Nov. 2007.

[17] M. Athineos, and D. P. W. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pp. 648-651, Hong Kong, Apr. 2003.

[18] J. Herre, and J. H. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," *Audio Engg. Soc.*, 101st Convention, Los Angeles, USA, Nov. 1996.

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

[20] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. of IEEE*, Vol. 63, No. 4, Apr. 1975.

[21] L. S. Marple, “Computing the Discrete-Time Analytic Signal via FFT”, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 47 (9), pp. 2600-2603, Sep. 1999.

[22] 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.

[23] V. Tyagi, C. Wellekens, “Fepstrum and Carrier Signal Decomposition of Speech Signals Through Homomorphic Filtering”, *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 5, pp. 1041-1044, France, May 3006.

[24] 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.

[25] S. M. Kay,“ Modern Spectral Estimation: Theory and Application”, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[26] P. Motlicek, S. Ganapathy, H. Hermansky, H. Garudadri and M. Athineos, “Perceptually motivated Sub-band Decomposition for FDLP Audio Coding,” *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, DE, pp. 435-442, Sep. 2008.

[27] X.M. Xie, S. C. Chan, and T. I. Yuk, “M-band perfect-reconstruction linear-phase filter banks,” *Proc. of the IEEE Signal Processing Workshop on Statistical Signal Processing*, pp. 583-586, Singapore, Aug. 2001.

[28] S. Ganapathy, P. Motlicek, H. Hermansky and H. Garudadri, “Autoregressive modelling of Hilbert Envelopes for Wide-band Audio Coding,” *Audio Engg. Soc.*, 124th Convention, Amsterdam, Netherlands, May 2008.

[29] J. Princen, A. Johnson and A. Bradley, “Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation,” *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 87, pp 2161-2164, Dallas, USA, May 1987.

[30] 3GPP TS 26.401: Enhanced aacPlus general audio codec; General Description.

[31] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre and B. Grill,“Unified speech and audio coding scheme for high quality at low bitrates,” *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

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

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

[34] M. Dietz, L. Liljeryd, K. Kjorling and O. Kunz, “Spectral Band Replication, a novel approach in audio coding,” *Audio Engg. Soc.*, 112th Convention, Munich, May 2002.

[35] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, “ISO/IEC MPEG-2 Advanced Audio Coding,” *J. Audio Engg. Soc.*, Vol. 45, no. 10, pp. 789-814, Oct. 1997.

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

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

[38] ITU-R Recommendation BS.1534: “Method for the subjective assessment of intermediate audio quality,” Jun. 2001.