IDIAP RESEARCH REPORT

Scalable Wide-band Audio Codec based on Frequency Domain Linear Prediction (version 2)

Petr Motlicek [*] Sriram Ganapathy [*]
Hynek Hermansky [*] Harinath Garudadri [+]

IDIAP–RR 07-16

September 2007

premire rvision : Avril 2007, version 1
seconde révision : September 2007, version 2

[*] IDIAP Research Institute, Martigny, Switzerland
[+] Qualcomm Inc., San Diego, California, US

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

Rapport de recherche de l'IDIAP 07-16

Scalable Wide-band Audio Codec based on Frequency Domain Linear Prediction (version 2)

Petr Motlicek Sriram Ganapathy Hynek Hermansky
Harinath Garudadri

September 2007

première révision : Avril 2007, version 1
seconde révision : September 2007, version 2

Résumé. This paper proposes a technique for wide-band audio applications based on the predictability of the temporal evolution of Quadrature Mirror Filter (QMF) sub-band signals. An input audio signal is first decomposed into 64 sub-band signals using QMF decomposition. The temporal envelopes in critically sampled QMF sub-bands are approximated using frequency domain linear prediction applied over relatively long time segments (e.g. 1000 ms). Line Spectral Frequency parameters related to autoregressive models are computed and quantized in each frequency sub-band. The sub-band residuals are quantized in the frequency domain using a combination of split Vector Quantization (VQ) (for magnitudes) and uniform scalar quantization (for phases). In the decoder, the sub-band signal is reconstructed using the quantized residual and the corresponding quantized envelope. Finally, application of inverse QMF reconstructs the audio signal. Even with simple quantization techniques and without any sophisticated modules, the proposed audio coder provides encouraging results on objective quality tests. Also, the proposed coder is easily scalable across a wide variety of bit-rates.

1 Introduction

Digital audio representation brings many advantages including unprecedented high fidelity, dynamic range and robustness in mobile and media coding applications. Due to the success provided by first generation digital audio applications, such as CD and DAT (digital audio tape), end-users have come to expect CD-quality audio reproduction from any digital system.

Furthermore, emerging digital audio applications for network, wireless, and multimedia computing systems face a series of constraints such as reduced and variable channel bandwidth, limited storage capacity and low cost.

New services are created rapidly on the Internet. Compared to the present situation, audio and video consumed only 2% of Internet traffic in 2000. IP networks, as a new service platform, introduce new possibilities for the customer. As a consequence, other services such as live audio and video streaming applications are possible (e.g. radio and TV broadcast over IP, multicast of a lecture, etc.). There are many reasons for this strong increase of audio and video traffic, such as faster Internet access, increased popularity of peer-to-peer applications (70% of the current Internet traffic), digital radio broadcasting in the web, technological advancements in soundcards and speakers, and development of high-quality compression techniques.

Interactive applications such as videophone or interactive games have a real-time constraint. This imposes a maximum acceptable end-to-end latency of the transmitted information, where end-to-end is defined as : capture, encode, transmit, receive, decode and display. The maximum acceptable latency depends on the application, but often is of the order of 150 ms. However, non-interactive applications have looser latency constraints, for example even few seconds. The critical constraints are few errors in transmission, few breaks in continuity, and the overall signal quality.

This paper mainly focuses on audio signal coding for non-interactive applications. A novel speech coding system, proposed recently [1], exploits the predictability of the temporal evolution of spectral envelopes of a speech signal using Frequency-Domain Linear Prediction (FDLP) [2, 3]. Unlike [2], this technique applies FDLP to approximate relatively long (up to 1000 ms) segments of the Hilbert envelopes in individual frequency sub-bands.

The approach was extended for wide-band applications (from 8 kHz up to 48 kHz) by including higher frequency sub-bands [4]. However, many difficulties arise specifically due to the need to preprocess and encode 1000 ms of a full-sampled input signal in each frequency sub-band. Efficient transmission of sub-band residual signals is also a challenge. This paper attempts to address most of these issues.

In contrast to the previous approaches, this paper proposes the use of FDLP on the sub-band signal and not on the full-band signal. First, the input signal is decomposed into frequency subbands using the maximally decimated QMF bank. Then, FDLP technique is applied in each critically sampled frequency sub-band independently. The Line Spectral Frequencies (LSFs) as well as the spectral components of the residual sub-band signals are quantized. Good properties of this novel coding scheme are shown and the first version of a variable bit-rate audio encoder based on QMF - FDLP techniques is proposed.

The rest of the paper is organized as follows : Section 2 provides a brief overview of the FDLP principle. Section 3 describes the QMF - FDLP codec. Simulation results and audio quality evaluations are given in Section 4, followed by conclusions and discussions in Section 5.

2 FDLP

The novelty of the proposed audio coding approach is the employment of FDLP method to parameterize Hilbert envelope (squared magnitude of an analytic signal) of the input signal [2, 3]. The FDLP can be seen as a method similar to Temporal Domain Linear Prediction (TDLP). In the case of TDLP, the AR model approximates the power spectrum of the input signal. The FDLP fits an AR model to the squared Hilbert envelope of the input signal. Using FDLP, we can adaptively capture

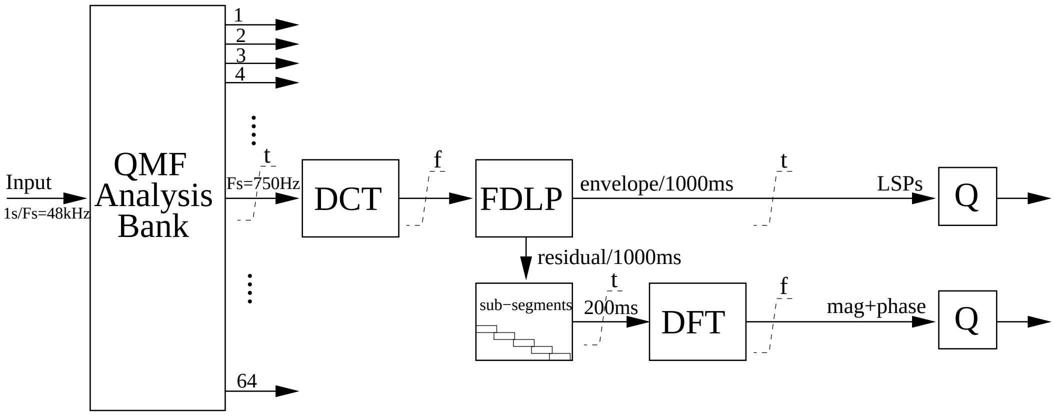

FIG. 1 – QMF-FDLP encoder structure (f - frequency domain, t - time domain).

fine temporal nuances with high temporal resolution while at the same time summarize the signal's gross temporal evolution in time scales of hundreds of milliseconds. In our system, we employ the FDLP technique to approximate the temporal envelope of sub-band signal in QMF sub-bands.

3 Structure of the codec

The first version of the coder based on FDLP for very low bit-rate narrow-band applications was proposed in [1]. The input speech signal is split into non-overlapping segments (hundreds of ms long). Then, each segment is DCT transformed and partitioned into unequal segments to obtain critical band-sized sub-bands. Finally, the DCT components which correspond to a given critical sub-band are used for calculating the FDLP model for that band. Since the FDLP model does not approximate the squared Hilbert envelope perfectly, the remaining residual signal (the carrier signal for the FDLP-encoded Hilbert envelope) is further processed and its frequency representatives are selectively quantized and transmitted.

However, the post-processing of sub-band residuals remains a complex operation with such setting. The sub-band residuals do not have good coding properties and henceforth, large bit-rate is required for high quality coding. In addition, this system setting is computationally expensive.

The next experiments, performed with audio signals sampled at 48 kHz, were motivated by the MPEG-1 architecture [5, 6]. In the MPEG-1 encoder, the input signal is first decomposed into critically sub-sampled frequency sub-bands using QMF bank (a polyphase realization), whose channels are uniformly spaced. In our system, the same operation is performed on the input signal. The band-pass outputs are decimated by a factor M (the number of sub-bands), yielding the sub-band sequences which form a critically sampled and maximally decimated signal representation (i.e. the number of sub-band samples is equal to the number of input samples). The QMF filters have flat pass-band response, which is advantageous for exploiting the predictability of frequency components of sub-band signals.

In the coder, we employ a 64 band decomposition compared to MPEG-1 standard of 32 frequency sub-bands. The parameters representing the FDLP model in each sub-band are not expensive from the final bit-rate point of view. The use of a higher number of sub-bands provides the following advantages :

1. Sub-band residuals are more frequency limited, and are easier to quantize using split VQ.
2. It is more advantageous when a psychoacoustic model (in the future) is employed to attenuate perceptually irrelevant frequency sub-bands.
3. Slightly better objective results.

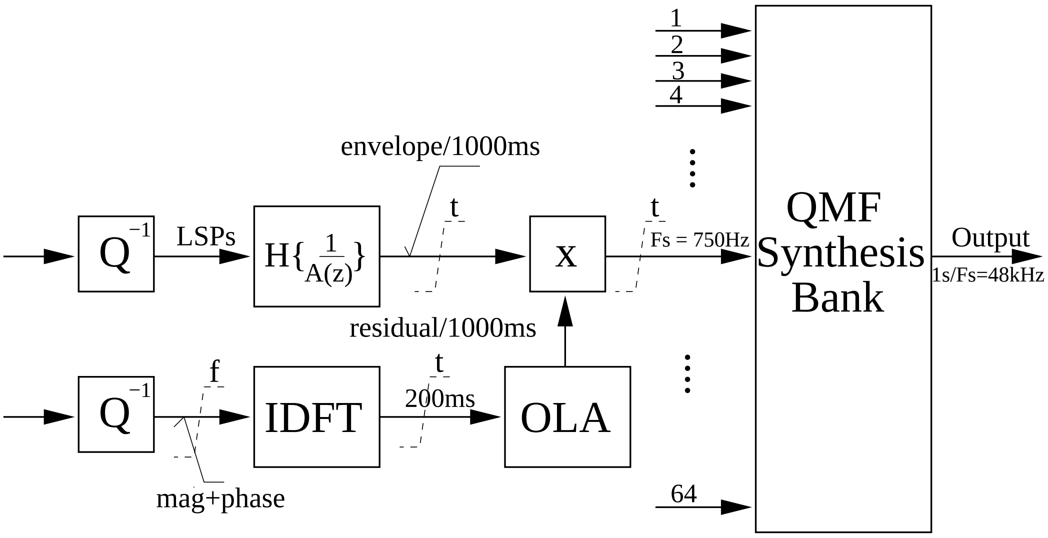

Fig. 2 – QMF-FDLP decoder structure (f - frequency domain, t - time domain).

The structure of the encoder and the decoder is depicted in figure 1 and 2, respectively.

3.1 Time-frequency analysis

The input signal is split into 1000 ms long frames. Each frame is decomposed into 64 sub-bands by QMF. A 99-th-order prototype filter with the structure as direct-form FIR polyphase decimator is used for the frequency decomposition. The prototype filter was designed for high sidelobe attenuation in the stop-band of each analysis channel (around 78 dB), which ensures that intraband aliasing due to quantization noise remains negligible. The magnitude and phase frequency response of the FIR filter is depicted in figure 3. In order to perform decomposition into 64 sub-bands, a cascade implementation of 2 band QMF decomposition provided by the prototype filter is utilized. The algorithmic delay of the implementation is around 130 ms.

3.2 Critically sampled sub-band processing

Each sub-band is DCT transformed which yields the input to the FDLP module. The magnitude frequency response of AR model, computed through the autocorrelation LPC analysis on the DCT transformed sub-band signal, approximates the squared Hilbert envelope of the 1000 ms sub-band signal. Spectral Transform Linear Prediction (STLP) technique [7] is used to control the fit of AR model to the Hilbert envelope of the input. The associated FDLP-LSF parameters are quantized and then transmitted.

The 1000 ms residual signal in each critically sampled sub-band, which represents the Hilbert carrier signal for the FDLP-encoded Hilbert envelope, is split into five 210 ms long overlapping sub-segments with 10 ms overlap. This is to take into account the non-stationarity of the Hilbert carrier over the 1000 ms frame. An overlap length of 10 ms ensures smooth transitions when the sub-segments of the residual signal are concatenated in the decoder. Finally, each sub-segment is DFT transformed which results in 79 complex spectral components distributed over 0 Hz to $F_s/2$ (= 375 Hz) with a frequency resolution of 4.75 Hz. The magnitude and phase components of the complex spectral representations are then quantized and transmitted to the decoder.

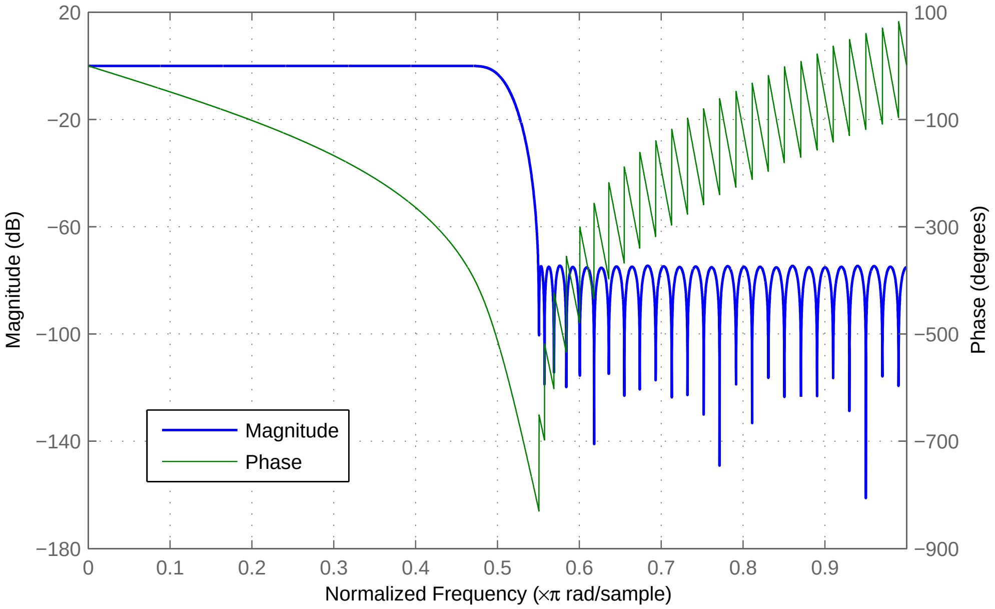

Fig. 3 – Magnitude and phase frequency response of theQMF prototype filter.

3.3 Quantization of parameters

Quantization of LSFs : The LSFs corresponding to the AR model in a given frequency sub-band over the 1000 ms input signal are vector quantized. In the experiments, a 20th order all-pole model is used. The total contribution of the FDLP models to the bit-rate for all the sub-bands is around 5 kbps.

Quantization of the magnitude components of the DFT transformed sub-segment residual : The magnitude spectral components are vector quantized. Since a full-search VQ in this high dimensional space would be computationally infeasible, the split VQ approach is employed. Although the split VQ approach is suboptimal, it reduces computational complexity and memory requirements to manageable limits without severely affecting the VQ performance. We divide the input vector of spectral magnitudes into separate partitions of a lower dimension. The VQ codebooks are trained (on a large audio database) for each partition using the LBG algorithm. Quantization of the magnitude components using the split VQ takes around 30 kbps for all the sub-bands.

Quantization of the phase components of the DFT transformed sub-segment residual : It is found that the phase components are uncorrelated across time. The phase components have a distribution close to uniform, and therefore, have a high entropy. Hence, we apply a 4 bit uniform scalar quantization for the phase components. To prevent excessive consumption of bits to represent phase coefficients, those corresponding to relatively low magnitude spectral components are not transmitted, i.e., the codebook vector selected from the codebook is processed by adaptive thresholding in the encoder as well as in the decoder. Only the spectral phase components whose magnitudes are above the threshold are transmitted. The threshold is adapted dynamically to meet a required number of spectral phase components (bit-rate). The options for reconstructing the signal at the decoder are :

1. Fill the incomplete phase components with uniformly distributed white noise.
2. Fill the incomplete phase components with zeros.
3. Make the magnitude components corresponding to incomplete phase components to zero.

In objective quality tests, it was found that the third option performed the best. Figure 4 shows

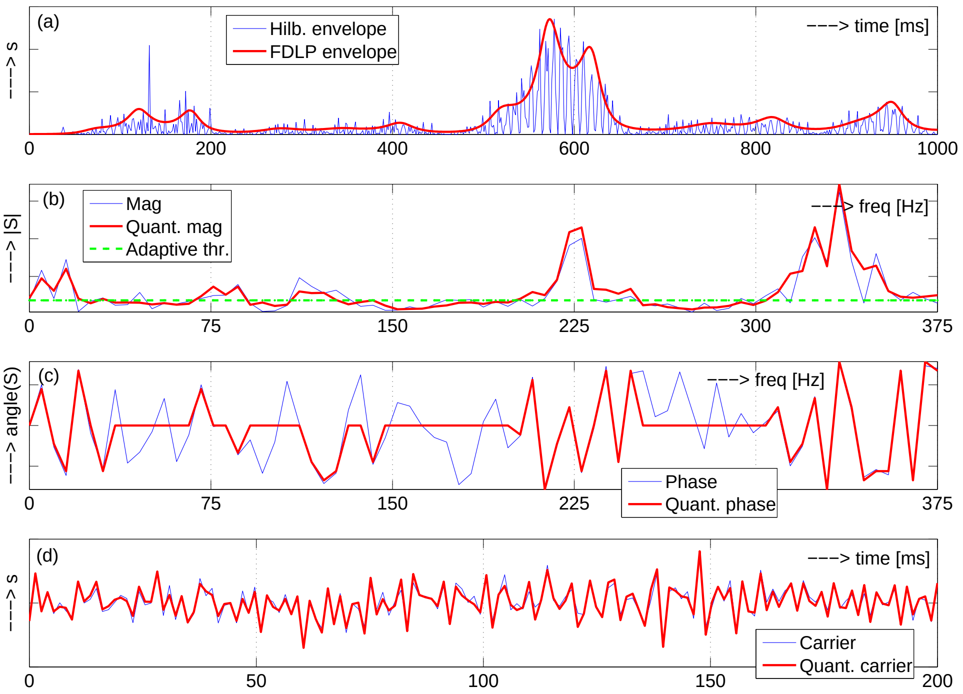

Fig. 4 – Time-frequency characteristics generated from a randomly selected audio example : (a) 1000 ms segment of the squared Hilbert envelope (estimated from the squared magnitude of an analytic signal) computed for the 3rd QMF sub-band, and its FDLP approximation. (b) Magnitude Fourier spectral components of the 200 ms residual sub-segment, and its reconstructed version, the adaptive threshold is also inserted. (c) Phase Fourier spectral components of the 200 ms residual sub-segment, and its reconstructed version. (d) Original 200 ms sub-segment of the residual signal (carrier) of the FDLP envelope, and its reconstructed version.

time-frequency characteristics of the original signal and the reconstructed signal for the proposed codec.

3.4 Decoding

In order to reconstruct the input signal, the carrier in each sub-band needs to be reproduced and then modulated by temporal envelope given by FDLP model.

The transmitted VQ codebook indices are used to select appropriate codebook vectors for the magnitude spectral components. Then, the adaptive threshold is applied on the reconstructed magnitudes and the transmitted scalar quantized phase spectral components are assigned to the corresponding magnitudes lying above the adaptive threshold. The sub-band carrier is created in the time domain from its spectral magnitude and phase information. The Overlap-add (OLA) technique is applied to obtain 1000 ms residual signal, which is then modulated by the FDLP envelope to obtain the reconstructed sub-band signal. Finally, a QMF synthesis bank is applied on the reconstructed sub-band signals to produce the output signal.

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

Tab. 1 – PEAQ scores and their meanings.

4 Experiments and Results

The qualitative performance of the proposed algorithm was evaluated using Perceptual Evaluation of Audio Quality (PEAQ) distortion measure [8]. 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.

The test was performed on 18 audio signals sampled at 48 kHz. These audio samples are part of the framework for exploration of speech and audio coding defined in [9]. They are comprised of speech, music and speech over music recordings. The ODG scores are presented in table 2.

First, the narrow-band speech coder [1] was extended for audio coding on 48 kHz sampled signal, where we employ Gaussian sub-band decomposition in 27 bands uniformly spaced in the Bark scale. The FDLP model was applied on these sub-band signals and residual signals were processed by adaptive threshold (half of the spectral components were preserved). This adaptive thresholding is performed to simulate the quantization process. The ODG scores for this technique are denoted as G-27h.

The results for QMF decomposition into 32 and 64 sub-bands are denoted as Q-32h and Q-64h, respectively. As in the previous case, the adaptive threshold is fixed to select half of the spectral components of the sub-band residual signals. Without quantization, Q-32h performs better than G-27h with approximately the same number of parameters to be encoded. In a similar manner, Q-64h slightly outperforms Q-32h.

ODG scores for Q-64 with quantization of spectral components of the sub-band residuals are presented for bit-rates $\sim$ 124 and $\sim$ 100 kbps. In both these experiments, magnitude spectral components are quantized using split VQ (10 codebooks each of dimension 8). Phase spectral components are scalarly quantized using 4 bits. To reduce the final bit-rates, the number of phase components to be transmitted is reduced using adaptive threshold (90% and 60% of phase spectral components resulting in 124 and 100 kbps respectively).

Finally, the ODG scores for standard audio coders : MPEG-1 Layer-3 LAME [6, 10] and MPEG-4 HE AACplus-v1 [11, 12], at bit-rates 64 and 48 kbps respectively, are also presented. The AACplus-v1 coder is the combination of Spectral Band Replication (SBR) [13] and Advanced Audio Coding (AAC) [14] and was standardized as High-Efficiency AAC (HE-AAC) in Extension 1 of MPEG-4 Audio [15].

5 Conclusions and Discussions

With reference to the ODG scores in table 2, the proposed codec needs to operate at 100 kbps in order to achieve the similar average quality as the MPEG-1 Layer-3 LAME standard at 64 kbps. In a similar manner, the proposed codec takes 124 kbps to perform as good as AACplus-v1 codec at 48kbps.

Even without any sophisticated modules like psychacoustic models, spectral band replication module and entropy coding, the proposed method (at the expense of higher bit-rate) is able to give

bit-rate	-	-	-	124	100	64	48
File	G-27h	Q-32h	Q-64h	Q-64	Q-64	MP3-LAME	AAC+ v1
SPEECH							
es01_s	-3.579	-1.380	-1.075	-0.811	-1.334	-2.054	-1.054
es02_s	-3.345	-1.826	-1.360	-0.820	-1.396	-1.432	-1.175
louis_raquin_1	-3.710	-2.785	-2.626	-1.670	-2.484	-1.931	-1.793
te19	-3.093	-1.829	-1.514	-1.016	-1.518	-1.140	-1.152
SPEECH OVER MUSIC							
Arirang_ms	-2.709	-2.056	-2.149	-1.741	-2.703	-1.750	-1.215
Green_sm	-2.656	-2.588	-2.332	-2.096	-2.765	-1.703	-1.147
noodleking	-2.312	-0.777	-0.677	-0.485	-0.705	-1.388	-0.912
te16_fe49	-2.316	-1.028	-1.106	-1.099	-1.678	-1.346	-1.558
te1_mg54	-2.668	-1.343	-1.340	-0.949	-1.487	-1.341	-1.319
twinkle_ff51	-2.174	-2.070	-1.705	-1.557	-2.162	-1.519	-0.822
MUSIC							
brahms	-2.635	-1.157	-1.038	-1.495	-1.788	-1.204	-1.777
dongwoo	-1.684	-0.675	-0.583	-0.471	-0.658	-1.369	-0.753
phi2	-2.194	-0.973	-0.696	-0.445	-0.834	-1.735	-0.748
phi3	-2.598	-1.263	-1.058	-0.774	-1.215	-1.432	-0.937
phi7	-3.762	-1.668	-1.635	-3.356	-3.624	-2.914	-1.551
te09	-2.997	-1.353	-1.239	-0.841	-1.490	-1.734	-1.579
te15	-2.006	-0.670	-0.644	-0.545	-0.845	-1.726	-0.995
trilogy	-2.002	-0.439	-0.461	-0.509	-0.694	-2.064	-0.951
AVERAGE	-2.691	-1.438	-1.291	-1.149	-1.632	-1.655	-1.191

Tab. 2 – PEAQ results on audio samples in terms of ODG scores : Music, Speech, Speech between music, Speech over music.

objective scores comparable to the state of the art codecs for the bit-rates presented. Furthemoe, by modifying the adaptive threshold, the proposed technique simply allows to scale the bit rates, while for example in MPEG-1 layer-3, such an operation is computationally intensive.

The succeeding versions of the proposed codec having perceptual bit allocation algorithms (for the carrier spectral components) is expected to reduce the bit rate considerably yet maintaining the quality.

Eventhough, the coder does not employ block switching scheme (thus avoiding structural complications), the FDLP technique is able to address temporal masking problems (e.g. pre-echo effect) in an efficient way [4]. Another important advantage of the proposed coder is its resiliency to packet loss. This results from reduced sensitivity of the human auditory system to drop-outs in a frequency band as compared to loss of a short-time frame.

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 6th Framework DIRAC Integrated Project.

Références

[1] Motlicek P., Hermansky H., Garudadri H., Srinivasamurthy N., "Speech Coding Based on Spectral Dynamics", *in Lecture Notes in Computer Science*, Vol 4188/2006, Springer Berlin/Heidelberg, DE, September 2006.

[2] Herre J., Johnston J. H., "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)", *in 101st Conv. Aud. Eng. Soc.*, 1996.

[3] Athineos M., Hermansky H., Ellis D. P. W., "LP-TRAP : Linear predictive temporal patterns", *in Proc. of ICSLP*, pp. 1154-1157, Jeju, S. Korea, October 2004.

[4] Motlicek P., Ullal V., Hermansky H., "Wide-Band Perceptual Audio Coding based on Frequency-domain Linear Prediction", *in Proc. of ICASSP*, Honolulu, USA, April 2007.

[5] Pan D., "A Tutorial on MPEG/Audio Compression", *IEEE Multimedia Journal*, pp. 60-74, Summer 1995.

[6] Brandenburg K. et all., "ISO-MPEG-1 Audio : A Generic Standard for Coding of High-Quality Digital Audio", J. Audio Eng. Soc., vol. 42, pp. 780-792, 1994.

[7] Hermansky H., Fujisaki H., Sato Y., "Analysis and Synthesis of Speech based on Spectral Transform Linear Predictive Method", *in Proc. of ICASSP*, Vol. 8, pp. 777-780, Boston, USA, April 1983.

[8] Thiede T., Treurniet W. C., Bitto R., Schmidmer C., Sporer T., Beerends J. G. ,Colomes C., Keyhl M., Stoll G., Brandenburg K., Feiten B., "PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality", J. Audio Eng. Soc., vol. 48, pp. 3-29, 2000.

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

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

[11] 3GPP TS 26.401 : "Enhanced aacPlus general audio codec ; General Description".

[12] Brandenburg K., Kunz O., Sugiyama A., "MPEG-4 Natural Audio Coding", *Signal Processing : Image Communication*, vol. 15, no. 4, pp. 423-444, January 2000.

[13] Dietz M., Liljeryd L., Kjorling K., Kunz O., "Spectral Band Replication, a novel approach in audio coding", in AES 112th Convention, Munich, DE, May 2002, Preprint 5553.

[14] Bosi M., Brandenburg K., Quackenbush S., Fielder L., Akagiri K., Fuchs H., Dietz M., Herre J., Davidson G., Oikawa Y., "ISO/IEC MPEG-2 Advanced Audio Coding", J. Audio Eng. Soc., vol. 45, no. 10, pp. 789814, Oct. 1997.

[15] ISO/IEC, "Coding of audio-visual objects Part 3 : Audio, AMENDMENT 1 : Bandwidth Extension", ISO/IEC Int. Std. 14496-3 :2001/Amd.1 :2003, 2003.