IDIAP RESEARCH REPORT

Non-uniform QMF Decomposition for Wide-band Audio Coding based on Frequency Domain Linear Prediction

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

IDIAP–RR 07-43

October 2007

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

Rapport de recherche de l'IDIAP 07-43

Non-uniform QMF Decomposition for Wide-band Audio Coding based on Frequency Domain Linear Prediction

Petr Motlicek Sriram Ganapathy Hynek Hermansky
Harinath Garudadri

October 2007

Résumé. This paper presents a new technique for perfect reconstruction non-uniform QMF decomposition developed to increase efficiency of a generic wide-band audio coding system based on Frequency Domain Linear Prediction (FDLP). The base line FDLP codec, operating at high bit-rates ($\sim$136 kbps), exploits an uniform QMF decomposition into 64 sub-bands followed by sub-band processing based on FDLP. Here, we propose a non-uniform QMF decomposition into 32 frequency sub-bands obtained by merging 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. Such frequency decomposition, when employed in the FDLP audio codec, results in a bit-rate reduction of 40% over the base line. We also describe the complete audio codec, which provides high-fidelity audio compression at $\sim$66 kbps. In subjective listening tests, the FDLP codec outperforms MPEG-1 Layer 3 (MP3) and achieves similar qualities as MPEG-4 AAC+ standard.

1 Introduction

Frequency Domain Linear Prediction (FDLP) forms an efficient technique to model the temporal envelopes in frequency sub-bands [1]. The first compression technique based on FDLP was developed for coding narrow-band speech at 8 kHz [2]. Employed FDLP gives compression efficiency due to the predictability of the temporal evolution of spectral envelopes in frequency sub-bands.

This compression technique was later extended to high quality audio coding [3], where an input audio signal is decomposed into N frequency sub-bands. Temporal envelopes of these sub-bands are then approximated using FDLP applied over relatively long time segments (e.g. 1000 ms).

Since the FDLP model does not represent the sub-band signal perfectly, the remaining residual signal (carrier) is further processed and its frequency representatives are selectively quantized and transmitted. Efficient encoding of the sub-band residuals plays an important role in the performance of the FDLP codec and this is largely dependent on frequency decomposition employed.

Recently, an uniform 64 band Quadrature Mirror Filter (QMF) decomposition (analogous to MPEG-1 architecture [4]) was employed in the FDLP codec. This version of the codec achieves good quality of reconstructed signal compared to the older version using Gaussian band decomposition [3]. The performance advantage was mainly due to the increased frequency resolution in lower sub-bands and critical sub-sampling resulting in the minimal number of FDLP residual parameters to be transmitted. However, a higher number of sub-bands resulted in higher number of AR model parameters to be transmitted. Hence, the final bit-rates for this version of the codec were significantly higher ($\sim$136 kbps) and therefore, the FDLP codec was incompetent with the state-of-the-art audio compression systems.

In this paper, we propose a non-uniform QMF decomposition to be utilized in the FDLP codec. The proposed sub-band decomposition provides a good compromise between fine spectral resolution for low frequency sub-bands and lesser number of FDLP parameters to be encoded. The idea of non-uniform QMF decomposition has been known for nearly two decades (e.g. [5, 6]). Since it provides the advantages of non-uniform frequency resolution, perfect reconstruction and critical sub sampling, it is widely used in audio coding (e.g. [7]). However, the proposed sub-band decomposition tries to simulate the human auditory critical band filters and, when employed in the FDLP codec, it provides significant bit-rate/quality improvement.

Other benefits of employing non-uniform sub-band decomposition in the FDLP codec are :

- As psychoacoustic models operate in non-uniform (critical) sub-bands, they can be advantageously used to reduce the final bit-rates.
- In general, audio signals have lower energy in higher sub-bands (above 12 kHz). Therefore, the temporal evolution of the spectral envelopes in higher sub-bands require only small order AR model (employed in FDLP). A non-uniform decomposition provides one solution to have same order AR model for all sub-bands and yet, reduces the AR model parameters to be transmitted.
- The FDLP residual energies are more uniform across the sub-bands and hence, similar post-processing techniques can be applied in all sub-bands.

The proposed technique becomes the key part of the high-fidelity audio compression system based on FDLP for medium bit-rates. Objective quality tests highlight the importance of the proposed technique, compared to the version of the codec exploiting uniform sub-band decomposition. Finally, subjective evaluations of the complete codec at $\sim$66 kbps show its relative performance compared to the state-of-the-art MPEG audio compression systems (MP3, AAC+) at similar bit-rates.

2 Non-uniform frequency decomposition

In the proposed sub-band decomposition, the 64 uniform QMF bands are merged to obtain 32 non-uniform bands. Since the QMF decomposition in the base line system is implemented in a tree-like structure (6-stage binary tree [3]), the merging is equivalent to tying some branches at any particular stage to form a non-uniform band. This tying operation tries to follow critical band decomposition in

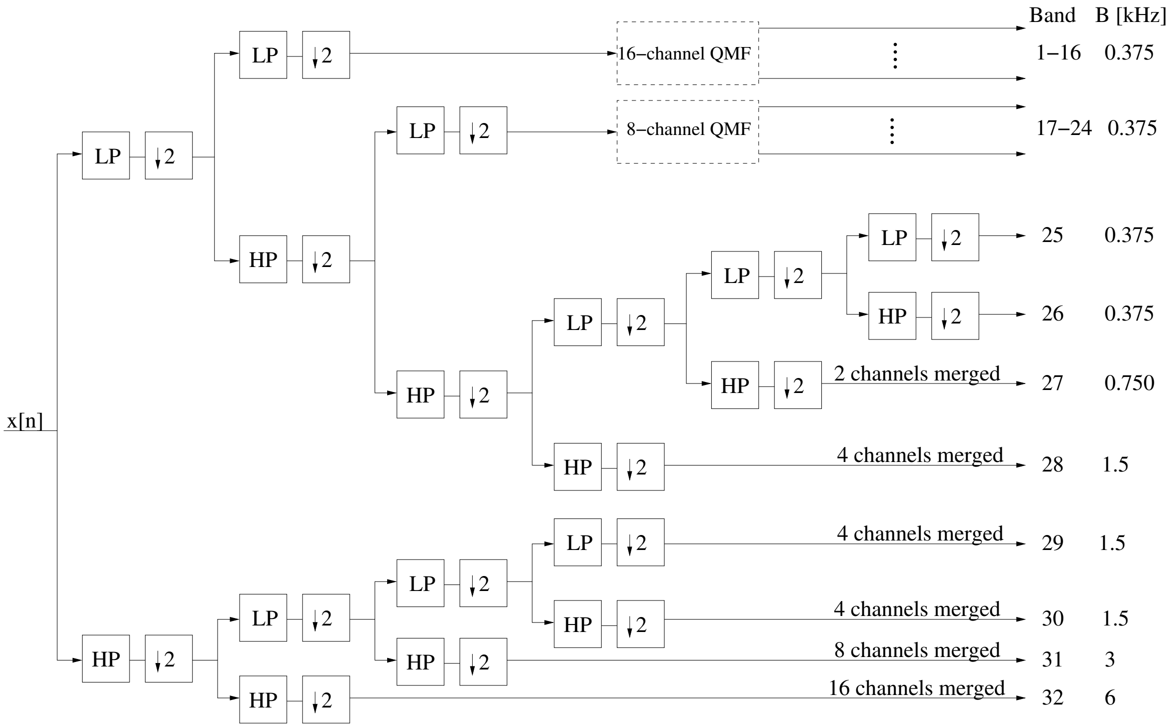


Fig. 1 – *The 32 channel non-uniform QMF derived using 6-stage network. Input signal $x[n]$ is sampled at 48 kHz. LP and HP denote Low-Pass and High-Pass band, respectively. $\downarrow 2$ denotes down-sampling by 2. B denotes frequency bandwidth.*

the human auditory system. This means that more bands at higher frequencies are merged together while maintaining perfect reconstruction. The graphical scheme of the non-uniform QMF analysis bank, resulting from merging 64 bands into 32 bands, is shown in Figure 1. The frequency widths of merged QMF bands are given in the right column in Figure 1.

Application of non-uniform QMF decomposition is supported by the Flatness Measure (FM) of the prediction error power E_i (energy of the residual signal of AR model in each sub-band i) computed across N QMF sub-bands. FM is defined as :

$$FM = \frac{Gm}{Am}, \tag{1}$$

where Gm is the geometric mean and Am is the arithmetic mean :

$$Gm = \sqrt[\frac{1}{N}]{\prod_{i=1}^{N} E_i}, \quad Am = \sum_{i=1}^{N} E_i. \tag{2}$$

If the input sequence is constant (contains uniform values), Gm and Am are equal, and $FM = 1$. In case of varying sequence, $Gm < Am$ and therefore, $FM < 1$.

We apply flatness measure to determine uniformity of the distributions of the prediction errors E_i across the QMF sub-bands. Particularly, for a given input frame, vector $\mathbf{E_u} = (E_1, E_2 \ldots E_N)$ obtained for uniform QMF decomposition ($N = 64$) is compared with the vector $\mathbf{E_n} = (E_1, E_2 \ldots E_N)$ containing the FDLP prediction errors for non-uniform QMF decomposition ($N = 32$). For each 1000 ms frame, FM is computed for the vectors $\mathbf{E_u}$ and $\mathbf{E_n}$. Figure 2 shows the flatness measure versus frame index for an audio sample. In case of uniform QMF decomposition, E_i is relatively high in the

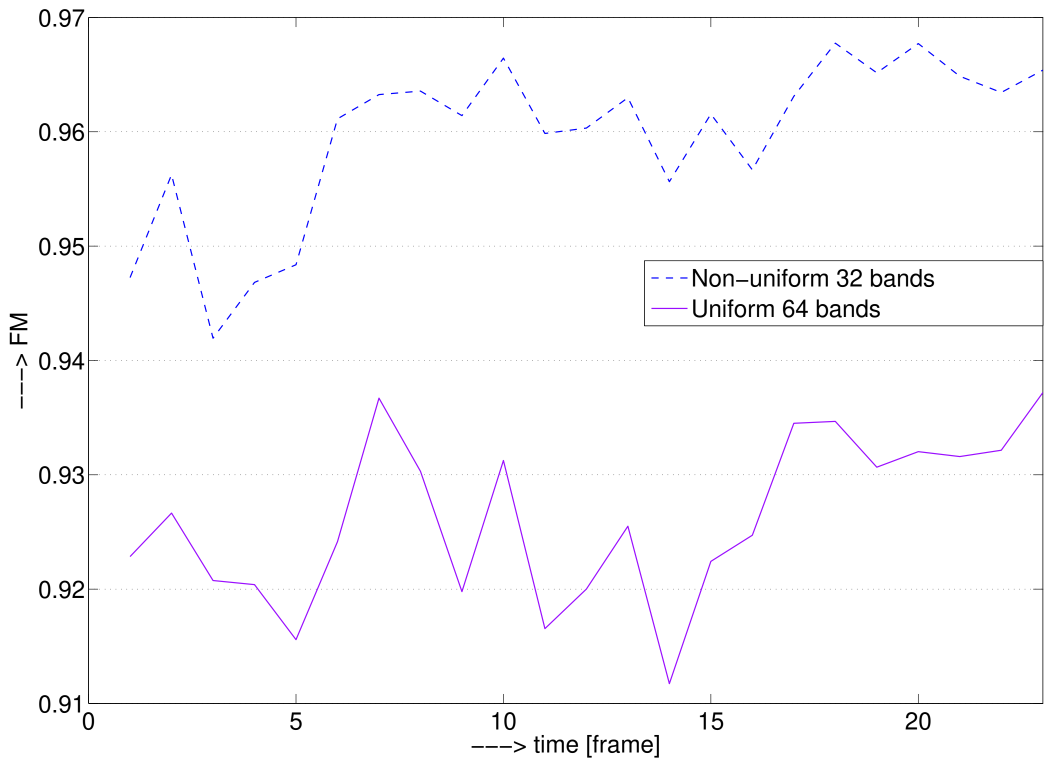


Fig. 2 – *Comparison of the flatness measure of the prediction error* **E** *for 64 band uniform QMF and 32 band non-uniform QMF for an audio recording.*

lower bands and low for higher bands. In case of non-uniform QMF analysis, E_i at higher bands is comparable to those in the lower frequency bands. Such FM curves can be seen for majority of audio samples.

A higher flatness measure of prediction error means that the degree of approximation provided by FDLP envelope is similar in all sub-bands. Therefore, non-uniform QMF decomposition allows uniform post-processing of FDLP residuals.

3 Structure of the FDLP codec

FDLP codec is based on processing long (hundreds of ms) temporal segments. As described in [3], the full-band input signal is 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 Discrete Fourier Transform (DFT) domain. Its magnitude spectral parameters are quantized using VQ, as well. Since a full-search VQ in this high dimensional space would be computationally infeasible, the split VQ approach is employed. Although this is a suboptimal approach, it reduces computational complexity and memory requirements to manageable limits without severely affecting the VQ performance. Phase spectral components of sub-band residuals are Scalar Quantized (SQ). Graphical scheme of the FDLP encoder is given in Figure 3.

In the decoder, shown in Figure 4, quantized spectral components of the sub-band carriers are reconstructed and transformed into time-domain using inverse DFT. The reconstructed FDLP envelopes

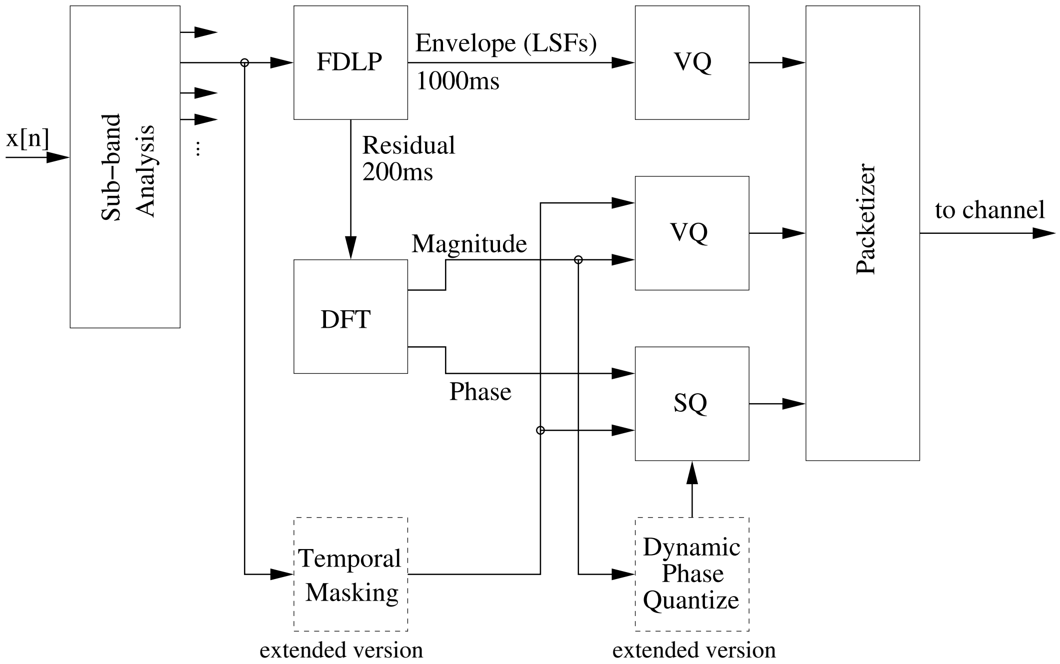


Fig. 3 – *Scheme of the FDLP encoder (dashed blocks refer to the extended version described in Section 4).*

(from LSF parameters) are used to modulate the corresponding sub-band carriers. Finally, sub-band synthesis is applied to reconstruct the full-band signal.

3.1 Objective evaluation of the proposed algorithm

The qualitative performance of the proposed non-uniform frequency decomposition is 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. ODG is an impairment scale which indicates the measured basic audio quality of the signal under test on a continuous scale from -4 (very annoying impairment) to 0 (imperceptible impairment). The test was performed on 18 challenging audio recordings sampled at 48 kHz. These audio samples form part of the MPEG framework for exploration of speech and audio coding [9]. They are comprised of speech, music and speech over music recordings.

The objective quality performances are shown in Table 1, where we compare the base line FDLP codec exploiting 64 band uniform QMF decomposition (QMF_{64}) at 136 kbps with the FDLP codec exploiting the proposed 32 band non-uniform QMF decomposition (QMF_{32}) at 82 kbps. Although, the objective scores of QMF_{32} are degraded by 0.2 compared to QMF_{64}, the bit-rate reduces significantly by around 40%. QMF_{32} is further compared to QMF_{64} operating at reduced bit-rates 88 kbps (bits for the sub-band carriers are uniformly reduced). In this case, the objective quality is reduced significantly. The block of quantization, described in [3], was not modified during these experiments.

4 The complete codec at 66 kbps

For subjective listening tests, the FDLP codec (described in Section 3) exploiting non-uniform QMF decomposition (described in Section 2) is further extended with a perceptual model, a block

bit-rate [kbps]	136	88	82
system	QMF_{64}	QMF_{64}	QMF_{32}
	Uniform	Uniform	Non-Uniform
ODG Scores	-1.04	-2.02	-1.23

Tab. 1 – *Mean objective quality test results provided by PEAQ (ODG scores) over* 18 *audio recordings : the base line FDLP codec at* 136 *kbps, the base line codec operating at reduced bit-rates* 88 *kbps, and the codec exploiting proposed 32 band non-uniform QMF decomposition at* 82 *kbps.*

performing dynamic phase quantization and a block of noise substitution, as shown in Figures 3 and 4.

Perceptual model : Perceptual model, described in [10], is based on temporal masking phenomena. Temporal masking is a property of the human ear, where the sounds appearing for about $100-200$ ms after a strong temporal signal get masked due to this strong temporal component. Masking thresholds determined by the model are used in controlling the bit-allocation for the DFT parameters of the sub-band residuals.

Dynamic phase quantization : It is found that the phase spectral components are uncorrelated. The phase components have a distribution close to uniform, and therefore, have high entropy. To prevent excessive consumption of bits required to represent phase coefficients, 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 (explained in [3]). Only the spectral phase components whose magnitudes are above a threshold are transmitted using high resolution SQ. 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 operation has a minimum impact on the quality of reconstructed audio.

Perceptual quantization based on temporal masking together with dynamic phase quantization and noise substitution techniques increase compression efficiency by around 20%. Final version of the FDLP codec operates at $\sim$66 kbps.

5 Subjective evaluations

The qualitative performance of the complete codec utilizing proposed non-uniform QMF decomposition is evaluated using MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening tests [11] performed on 8 audio samples from MPEG audio exploration database [9]. We compare the subjective quality of the following codecs :

- Complete version of the FDLP codec at $\sim$66 kbps.
- LAME - MP3 (MPEG 1, layer 3) at 64 kbps [12]. Lame codec based on MPEG-1 architecture is currently considered the best MP3 encoder at mid-high bit-rates and at variable bit-rates.
- AAC+, v1 at $\sim$64 kbps [13]. The AAC+ coder is the combination of Spectral Band Replication (SBR) [14] and Advanced Audio Coding (AAC) [15] and was standardized as High-Efficiency AAC (HE-AAC) in Extension 1 of MPEG-4 Audio [16].

The cumulative MUSHRA scores (mean values with 95% confidence) are shown in Figures 5(a) and (b). MUSHRA tests were performed independently in two different labs (with the same setup). Figure 5(a) shows mean scores for the results from both labs (combined scores for 18 non-expert listeners and 4 expert listeners), while Figure 5(b) shows mean scores for 4 expert listeners in one lab. These figures show that the FDLP codec performs better than LAME-MP3 and closely achieves subjective results of AAC+ standard.

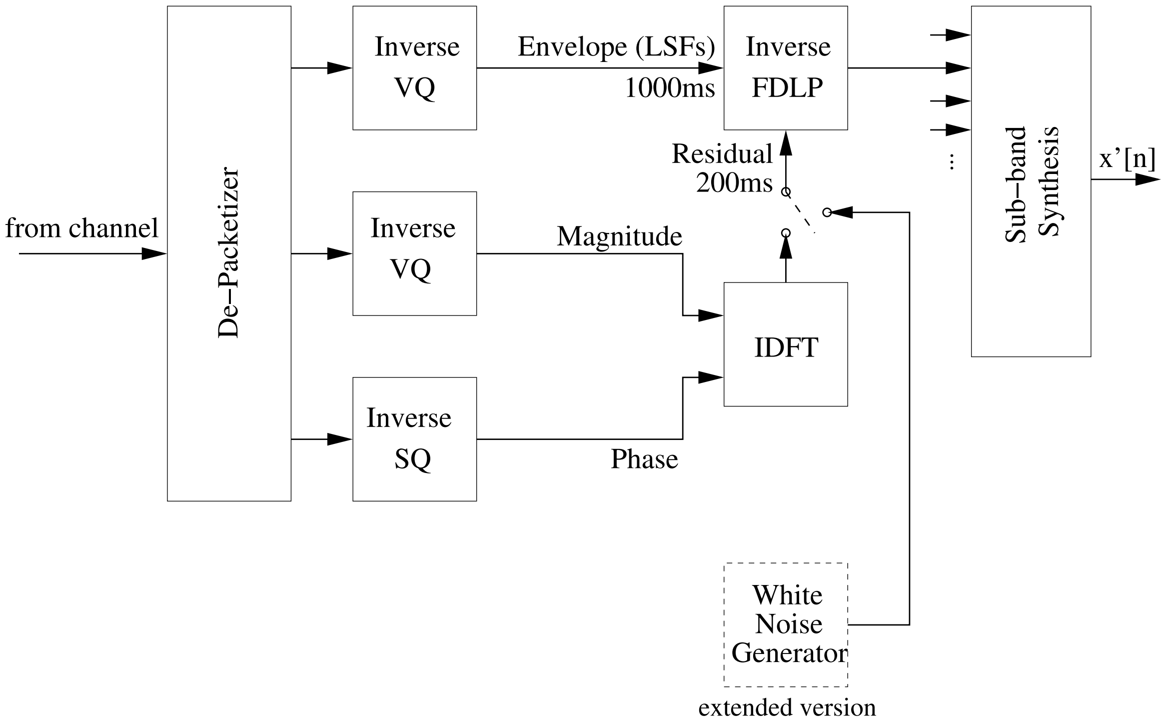


Fig. 4 – *Scheme of the FDLP decoder (dashed block refers to the extended version described in Section 4.)*

6 Conclusions and Discussions

A new technique for non-uniform frequency decomposition is presented, which is employed in the FDLP wide-band audio codec. The resulting QMF sub-bands closely follow the human auditory critical band decomposition. According to objective quality results, the new technique provides bit-rate reduction of about 40% over the base line, which is mainly due to transmitting few spectral components from the higher bands without affecting the quality significantly. Subjective evaluations, performed on the extended version of the FDLP codec, suggest that the complete FDLP codec operating at ∼66 kbps provides better audio quality than LAME - MP3 codec at 64 kbps and gives competent results compared to AAC+ standard at ∼64 kbps. FDLP codec does not make use of compression efficiency provided by entropy coding and simultaneous masking. These issues are open for future work.

Références

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

[2] P. Motlicek, H. Hermansky, H. Garudadri, N. Srinivasamurthy, "Speech Coding Based on Spectral Dynamics", Proceedings of TSD 2006, LNCS/LNAI series, Springer-Verlag, Berlin, pp. 471-478, September 2006.

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

[4] D. Pan, "A tutorial on mpeg audio compression", IEEE Multimedia Journal, vol. 02, no. 2, pp. 60-74, Summer 1995.

[5] A. Charbonnier, J-B. Rault, "Design of nearly perfect non-uniform QMF filter banks", *in Proc. of ICASSP*, New York, NY, USA, April 1988.

[6] P. Vaidyanathan, "Multirate Systems And Filter Banks", Prentice Hall Signal Processing Series, Englewood Cliffs, New Jersey 07632, 1993.

[7] G. Theile, G. Stoll, M. Link, "Low-bit rate coding of high quality audio signals", *in Proc. 82nd Conv. Aud. Eng. Soc.*, Mar. 1987, preprint 2432.

[8] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, "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, Lausanne, Switzerland, July 2007.

[10] S. Ganapathy, P. Motlicek, H. Hermansky, H. Garudadri, "Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction", *Tech. Rep., IDIAP*, RR 07-48, October 2007.

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

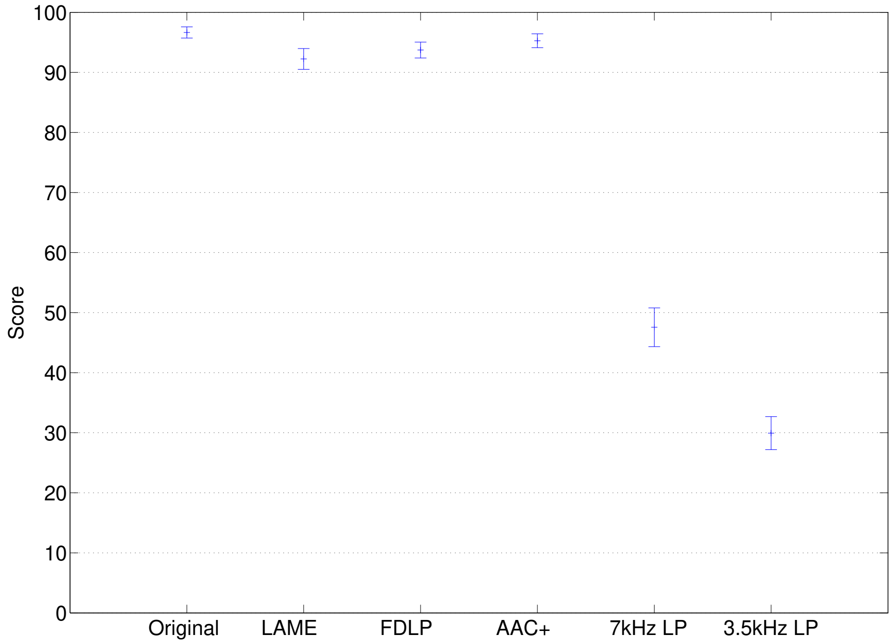
[12] LAME MP3 codec : <*http ://lame.sourceforge.net*>

[13] 3GPP TS 26.401 : "Enhanced aacPlus General Audio Codec", General Description.

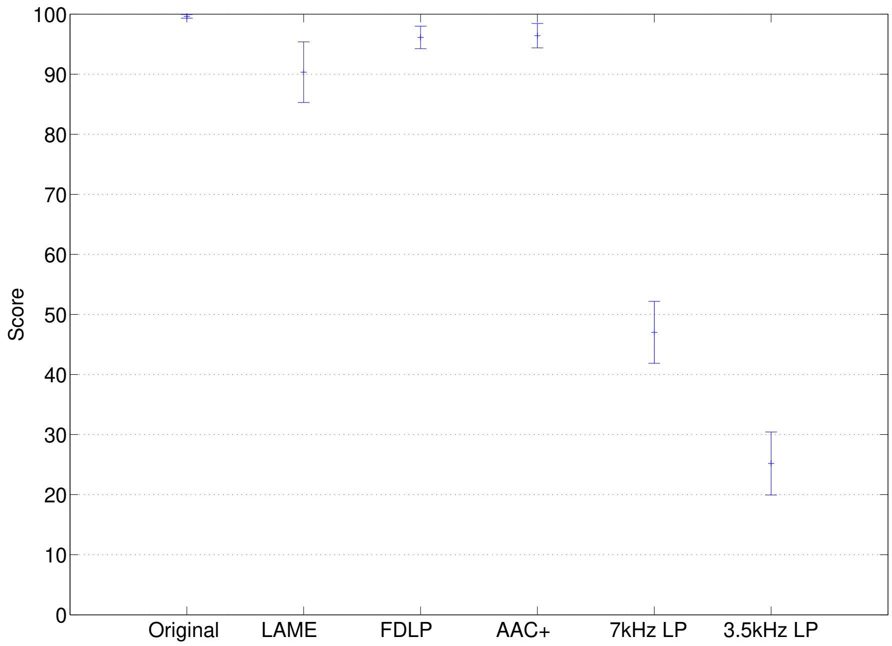
[14] M. Dietz, L. Liljeryd, K. Kjorling, O. Kunz, "Spectral Band Replication, a novel approach in audio coding", in AES 112th Convention, Munich, DE, May 2002, Preprint 5553.

[15] 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. 789814, October 1997.

[16] 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.



(a) 22 listeners.



(b) 4 expert listeners.

FIG. 5 – *MUSHRA results for 8 audio samples using three coded versions (FDLP, AAC+ and LAME MP3), hidden reference (original) and two anchors (7 kHz and 3.5 kHz low-pass filtered).*