IDIAP RESEARCH REPORT

Audio Coding Based on Long Temporal Segments: Experiments With Quantization of Excitation Signal

Vijay Ullal [ab] Petr Motlicek [b]

IDIAP–RR 06-46

August 2006

[a] International Computer Science Institute, Berkeley, California, US
[b] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 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 06-46

Audio Coding Based on Long Temporal Segments: Experiments With Quantization of Excitation Signal

Vijay Ullal Petr Motlicek

August 2006

Résumé. In this paper, we describe additional experiments based on a novel audio coding technique that uses an autoregressive model to approximate an audio signal's Hilbert envelope. This technique is performed over long segments (1000 ms) in critical-band-sized sub-bands. We have performed a series of experiments to find more efficient methods of quantizing the frequency components of the Hilbert carrier, which is the excitation found in the temporal audio signal. When using linear quantization, it was found that allocating 5 bits for transmitting the Hilbert carrier every 200 ms was sufficient. Other techniques, such as quantizing the first derivative of phase and using an iterative adaptive threshold, were examined.

1 Introduction

Since the dynamics of speech vary rapidly over time, most speech analysis techniques assume short-term stationarity within the signal [1]. Many existing speech coders use classical linear prediction (LP) to approximate the speech signal's spectral envelope over short frames, usually between 10 and 30 milliseconds. Since LP-based coding uses the source-filter model of speech production, these techniques do not work well for other types of audio signals, such as music [2]. However, the evolution of vocal tract shape can be largely predictable and it may be more efficient to encode longer segments of speech on the order of hundreds of milliseconds, rather than 10-30 ms. Modeling the temporal signal instead of the frequency components of the signal can be performed over longer periods of time [3]. While the Hilbert operator in the time domain can be used to accomplish this modeling, its infinite impulse response introduces problems. Thus, using an autoregressive model is a viable alternative. The coding method we use involves fitting an autoregressive model to the squared Hilbert envelope of a given signal. This technique, called Frequency Domain Linear Prediction (FDLP), is similar to classical linear prediction [4]. While classical linear prediction is used to model a signal's spectral envelope, FDLP is used to model a signal's temporal envelope. The residual information, or excitation signal, is known as the Hilbert carrier. Since encoding the Hilbert envelope is fairly straightforward, the main goal of this research is to find more efficient ways to model and encode the Hilbert carrier.

2 Codec design

2.1 Encoding

FDLP is used to encode an audio signal by estimating the signal's Hilbert envelope using an autoregressive model. As can be seen in Figure 1, the audio signal is divided into 1000 ms non-overlapping segments. The Discrete Cosine Transform (DCT) is applied to a 1000 ms segment so that it is projected into the frequency domain. The DCT transformed signal is then passed through a set of 15 critical-band Gaussian filters, which are spaced equally on the Bark scale. Each filter has a standard deviation of 1 Bark and a center frequency given in the Bark scale used in Perceptual Linear Prediction (PLP) analysis [5]. The autocorrelation LP technique is applied to each sub-segment of the DCT transformed signal, $y_k(f)$, where k represents a given frequency sub-band. In autocorrelation LP, the Fast Fourier Transform (FFT) of the signal is first taken, and then the power of the sequence is found. The Inverse Fast Fourier Transform (IFFT) of the result, which represents the autocorrelation sequence of the original signal, is taken. Finally, the Levinson-Durbin algorithm is applied to the autocorrelation sequence to create a set of LP coefficients. The Fourier transform of the impulse response that contains these coefficients will approximate the squared Hilbert envelope of the signal. In order to balance the effects of the peaks and dips in the envelope, the Spectral Transform Linear Prediction (STLP) technique is used [6]. This is accomplished by setting the power sequence to a compression factor exponent, which can be seen in Figure 1 as "Compress." In our case, the compression factor is set to 0.1 and the FDLP model order is 20; thus, 20 LP coefficients are obtained for each sub-band per 1000 ms. The Line Spectral Frequencies (LSF) that correspond to the LP coefficients are transmitted to the decoder.

The excitation $c_k(t)$, which is the Hilbert carrier, is needed to reconstruct the time domain signal in each critical sub-band. By modulating $c_k(t)$ with the approximated temporal envelope $a_k(t)$, the original temporal signal $x_k(t)$ can be obtained in each sub-band (see e.g., [7] for a mathematical explanation). Thus, the carrier in each sub-band is found through point-by-point division of $x_k(t)$ by $a_k(t)$. Unlike the Hilbert envelope, the carrier changes more quickly in time, so processing is performed every 200 ms. Each Hilbert carrier is demodulated so that its Fourier spectrum, which was once centered on F_k Hz, is now centered at 0 Hz. Demodulation allows for subsequent processing to be done more efficiently since all Hilbert carriers now have similar properties in all sub-bands. The demodulated carrier is then passed through a 200 Hz low-pass filter and decimated to preserve only the shifted components centered at 0 Hz. This technique of demodulation and downsampling is known

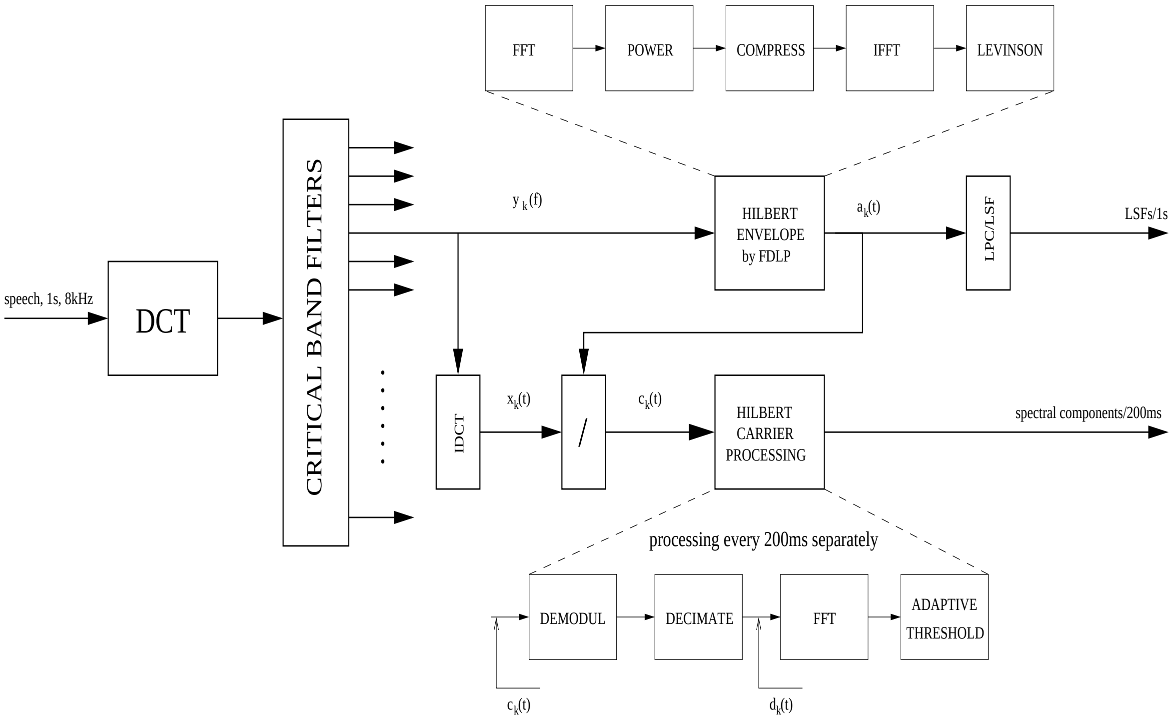

FIG. 1 – *Structure of the encoder*

as heterodyning. An adaptive threshold is applied to the Fourier spectrum of the modulated and decimated carrier, which we call $d_k(t)$. The 40 components with the greatest magnitude are preserved, and the quantized values of the magnitude and phase of these selected components are transmitted to the decoder.

2.2 Decoding

To reconstruct the original signal, the carrier $c_k(t)$ must be modulated by the temporal envelope $a_k(t)$. In order to regenerate $c_k(t)$, the IFFT of the 40 most significant spectral components that were transmitted is taken. The result is then up-sampled to the original rate and modulated so that the spectrum is again centered at F_k Hz. Spectrum symmetry is accomplished in order to obtain a real time domain signal. The temporal envelope $a_k(t)$ is reconstructed by finding the Fourier transform of the impulse response corresponding to the transmitted LSFs. The temporal signal $x_k(t)$ is then found by modulating $c_k(t)$ with $a_k(t)$. After this process has been performed in all frequency sub-bands, the temporal sub-band signals $x_k(t)$ are projected to the frequency domain by the DCT. The resulting signals are then summed and “de-weighted” by the critical-band Gaussian filters used in the encoder. By taking the Inverse DCT, the original 1000 ms temporal input segment is reconstructed.

3 Experiments

The codec was applied to a subset of the TIMIT database that included a total of 380 speech files. Although the original speech files were sampled at 16 kHz, the files that were used in this report were

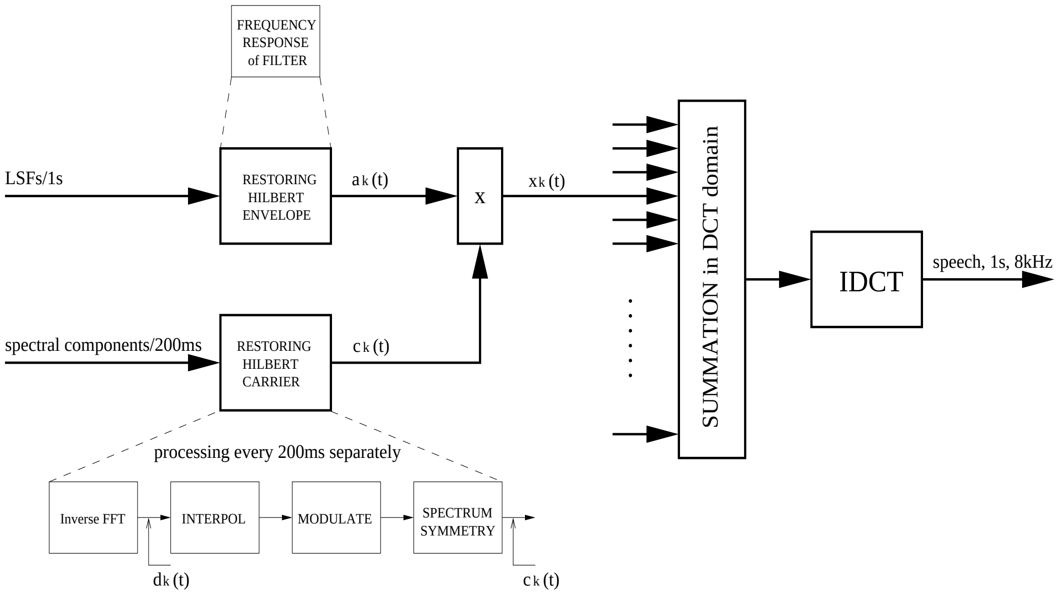

FIG. 2 – *Structure of the decoder*

downsampled to 8 kHz. Ten sentences were spoken by a total of 38 speakers (14 female, 24 male). Each sentence had a duration of 3-5 seconds.

For a given experiment, the Itakura-Saito (I-S) distance between a coded file and the 8kHz-sampled original file was computed every 7.5 ms on a 30 ms frame of speech. The I-S distance measure is a way of relating two LP vectors [8]. A larger distance measure signifies that the LP vectors are less similar and thus, in our case, the speech frame from the coded file is a much more degraded version of the frame from the original file. When comparing a coded file and the original, 95 percent of the I-S distances were taken into account, and the mean of these distances was calculated. Thus, a score could be given to each coded file, and the average over all 380 coded files was then calculated to obtain a distance measure for one type of experiment.

3.1 Quantizing carrier magnitude and phase

As stated earlier, the encoder transmits both the LSFs that model the Hilbert envelope and the most significant spectral components of the Hilbert carrier. In this set of experiments, we chose to transmit the 40 most significant spectral components in each 200 ms sub-segment for each sub-band. The magnitude and the phase of these spectral components were quantized using 2-5 bits for each. Additionally, experiments were run without quantizing either magnitude, phase, or both. The Itakura-Saito distance for each coded file can be seen in Table 3.1. An important pattern to notice is that increasing the number of bits to quantize magnitude while keeping the number of phase bits constant does not significantly improve the signal quality. This supports our initial idea that most of the spectral information is carried in the phase and that we must allocate more bits to quantize it rather than magnitude. Another interesting thing to notice is that the I-S distances are high when quantizing phase with 2 bits. This is due to the fact that one of the bits used to quantize phase is used to determine the sign. When comparing the distance measures from the quantized coded files to the non-quantized coded files, we believe that allocating 2 bits for magnitude and 3 bits for phase (a distance of 3.13) is sufficient. Additionally, subjective tests must be performed to validate this assumption.

In order to ensure that the average I-S distances obtained were not skewed, distributions of the distance measures were calculated for all coded files. An example of four distributions over all files can be seen in Figure 3. All of the histograms show the distribution of I-S distance of coded files when

		Bits of Phase				
		2	3	4	5	NQ
Bits of Mag	2	21.0759	3.3860	2.2822	2.1114	2.0845
	3	21.4106	3.1300	2.0373	1.7960	1.8056
	4	21.4297	3.1537	1.7993	1.8085	1.7944
	5	21.5600	3.1395	1.9998	1.7827	1.7718
	NQ	21.4985	3.1442	1.9933	1.6827	1.6459

Tab. 1 – Global mean I-S distance measure for varying number of bits used to quantize carrier magnitude and phase (NQ means "not quantized")

quantizing magnitude with 2 bits. The distances range from 0 to 50. The top left histogram shows when quantizing phase with 2 bits, the top right with 3 bits, the bottom left with 4 bits, and the bottom right with 5 bits. There is a large difference between the first two histograms and a smaller one between the second and third histograms ; however, quantizing phase with 4 or more bits does not greatly change the histogram.

With the current codec, both magnitude and phase are quantized linearly. An effort was undertaken to reduce the number of bits needed to quantize while still keeping the same quality. One idea was to quantize the phase difference, or first derivative, of the carrier spectral components. Preliminary experiments showed that although such a quantization produced lower I-S distance measures when using 2 bits to quantize phase ; when using 3 or 4 bits, the measures were worse. In order to further understand why quantizing the first derivative of phase did not yield sufficient results, a set of histograms was created to view the distribution of phase differences. The histograms include phase differences from -π to π from sub-bands 4, 8, and 12, over all files in the TIMIT subset. As shown in Figure 4, the distribution of phase difference is fairly uniform for the total of all three sub-bands and even when looking individually at sub-bands 8 and 12. From these figures, it was realized that no bits could be saved by quantizing phase difference.

3.2 Analysis of magnitude spectrum of carrier

Since little progress could be made in directly quantizing the magnitude and phase of the Hilbert carrier spectral components, another issue to further study was the adaptive threshold used to select the most significant components. Currently, we are selecting the adaptive threshold to keep the 40 most significant components in each sub-band ; in effect, having a bit rate of approximately 15 kbps (without side information) if 4 bits are used to quantize the LSFs that model the Hilbert envelope and 5 bits are used to quantize the Hilbert carrier components. However, preserving the same number of components for each sub-band is not efficient since fewer frequency components are needed in the lower sub-bands and more components are needed in the higher sub-bands to maintain good quality. This occurs because the frequency sub-bands follow critical-band widths according to the Bark scale. Thus, the sub-band width is lower at low frequencies and higher at high frequencies.

This assumption was tested by keeping the 40 most significant components in each sub-band for each 200 ms sub-segment and calculating the ratio in decibels of the maximum spectral component to the 40th most significant component. It is important to note that the ratio used in this experiment is not related to the I-S distance measure used earlier. The ratio used in this experiment represents number of decibels within one sub-segment of one sub-band between the most significant and least significant spectral component after adaptive thresholding. Figure 5 shows the distribution of decibel ratios over all sub-segments over the TIMIT subset for sub-bands 4, 8, and 12. As was expected, the average ratio is higher in the lower sub-bands, which means that there are fewer significant components in the lower sub-bands. In the higher sub-bands, the magnitude of the components are more uniform. One can see the average ratios over all sub-bands (2-13) in Figure 6.

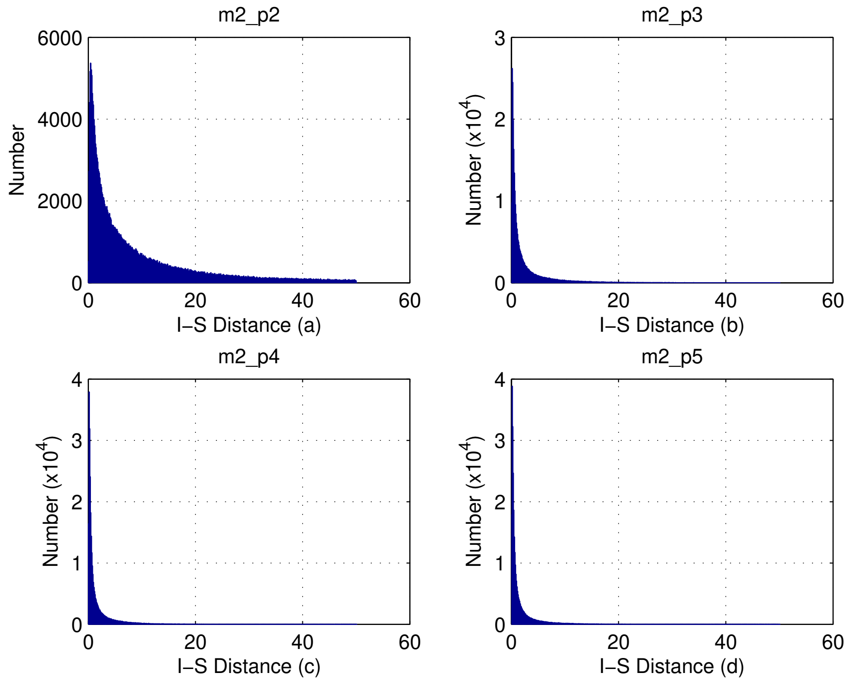

Fig. 3 – Distribution of I-S distance for 380 coded speech files when magnitude is quantized with 2 bits and phase is quantized with (a) 2 bits, (b) 3 bits, (c) 4 bits, (d) 5 bits

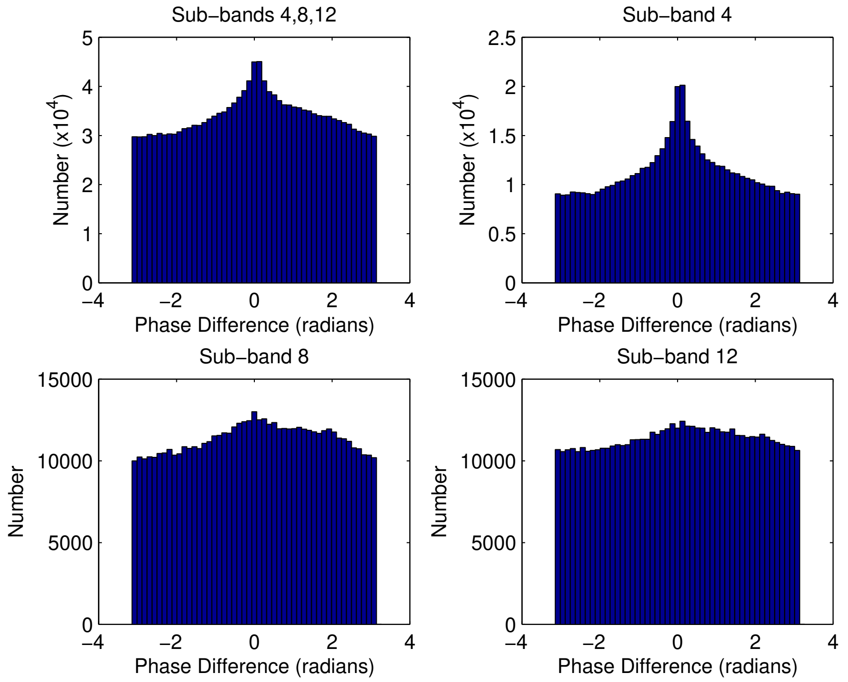

Fig. 4 – Distribution of difference in phase (first order derivative) over 200 ms frames for all files in TIMIT subset

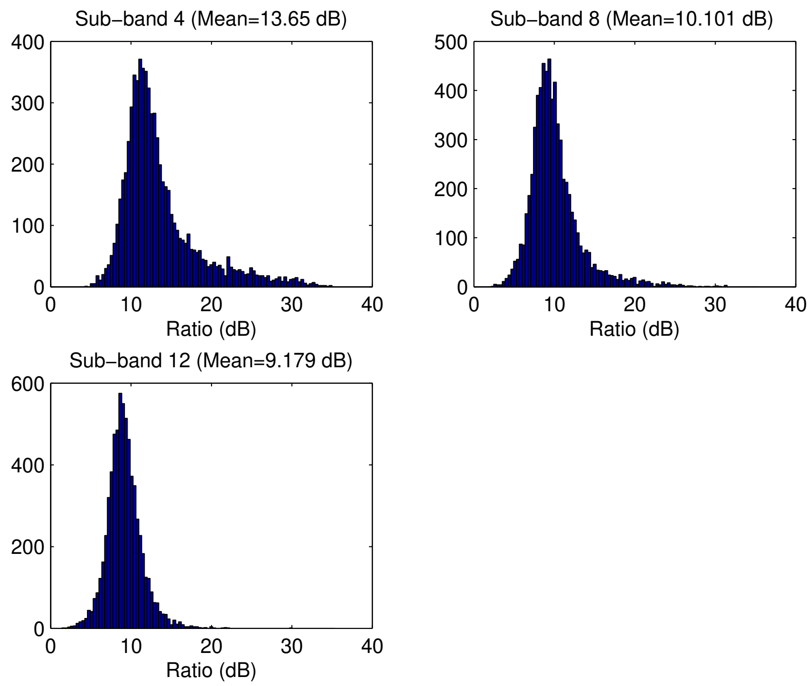

FIG. 5 – Distribution of ratios of highest to lowest spectral components (40 total) over 200 ms frames for all files in TIMIT subset

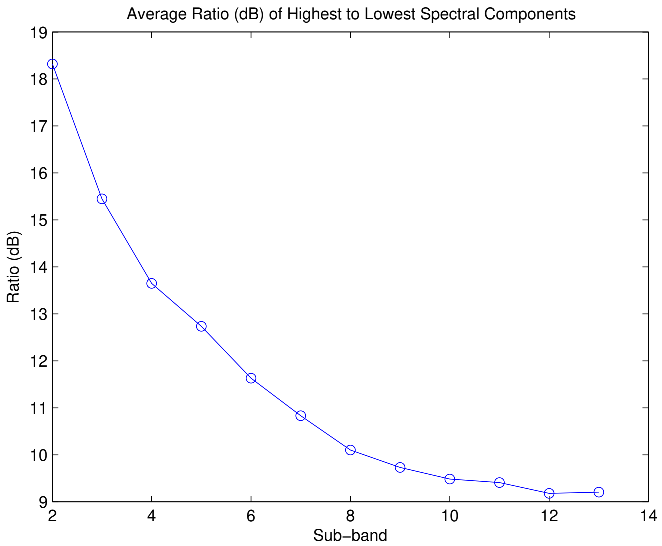

FIG. 6 – Global mean ratio of highest to lowest spectral components (40 total)

4 Conclusions and future work

Several experiments were performed to verify the optimal parameters used within the codec. Additionally, when quantizing the Hilbert carrier components, it was found that allocating 2 bits for magnitude and 3 bits for phase provided a low Itakura-Saito distance measure. Various methods, such as quantizing the first derivative of phase, were attempted, but they were not found to be useful. Thus, other ways to obtain a lower bit rate, such as finding and changing the adaptive threshold, were studied.

In the future, there are several issues which can be further investigated. One such issue includes iteratively finding an adaptive threshold within a sub-band while still keeping the same overall bit rate. This would entail selecting a different number of components within a sub-segment and a sub-band, as opposed to preserving a previously-determined number. Additionally, frequency masking can also be applied to lower sub-bands where there are fewer significant frequency components. Experiments must be run on audio files containing both music and speech, instead of only speech. Other objective quality measures, rather than I-S distance and SNR-based measurements, can also be employed.

5 Acknowledgements

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

Références

[1] Motlicek P., Hermansky H., Garudadri H., Srinivasamurthy N., "Audio Coding Based on Long Temporal Contexts", *technical report IDIAP-RR 06-30*, <http ://www.idiap.ch>, April 2006.

[2] Spanias A. S., "Speech Coding : A Tutorial Review", *In Proc. of IEEE*, Vol. 82, No. 10, October 1994.

[3] Herre J., "Temporal Noise Shaping, Quantization, and Coding Methods in Perceptual Audio Coding : A Tutorial Introduction", AES 17th Int. Conf. on High Quality Audio Coding.

[4] 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.

[5] Hermansky H., "Perceptual Linear Predictive (PLP) Analysis for Speech", *J. Acoust. Soc. Am.*, Vol. 87 :4, pp. 1738-1752, 1990.

[6] 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.

[7] Schimmel S., Atlas L., "Coherent Envelope Detector for Modulation Filtering of Speech", *in Proc. of ICASSP*, Vol. 1, pp. 221-224, Philadelphia, USA, May 2005.

[8] Quackenbush S. R., Barnwell T. P., Clements M. A., "Objective Measures of Speech eQuality", *Prentice-Hall, Advanced Reference Series*, Englewood Cliffs, NJ, 1988.