IDIAP RESEARCH REPORT

Audio Coding Based on Long Temporal Contexts

Petr Motlicek [*] Hynek Hermansky [*]
Harinath Garudadri [+]
Naveen Srinivasamurthy [+]

IDIAP–RR 06-30

avril 2006

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

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 06-30

Audio Coding Based on Long Temporal Contexts

Petr Motlicek Hynek Hermansky Harinath Garudadri
Naveen Srinivasamurthy

avril 2006

Résumé. We describe novel audio coding technique designed to be utilized at medium bit-rates. Unlike classical state-of-the-art audio coders that are based on short-term spectra, our approach uses relatively long temporal segments of audio signal in critical-band-sized sub-bands. We apply auto-regressive model to approximate Hilbert envelopes in frequency sub-bands. Residual signals (Hilbert carriers) are demodulated and thresholding functions are applied in spectral domain. The Hilbert envelopes and carriers are quantized and transmitted to the decoder. Our experiments focused on designing audio coder to provide broadcast radio-like quality audio around $10-20$kbps. Objective quality measures indicate comparable performance with the 3GPP-AMR speech codec standard for both speech and non-speech signals.

1 Introduction

State-of-the-art speech coding techniques that generate toll quality speech typically exploit the short-term predictability of speech signal in the $20-30$ms range. This short-term analysis is based on the assumption that the speech signal is stationary over these segment durations. Techniques like Linear Prediction (LP), which is able to efficiently approximate envelopes of short-term power spectra by Auto-Regressive (AR) model, are applied.

However, speech signal is quasi-stationary and carries information in its dynamics. Such information is not adequately captured in classical LP based approaches. Some considerations that motivated us to explore novel architectures are mentioned below :

- The signal dynamics are described by a sequence of short-term vectors : Many issues come up, e.g., windowing, proper sampling of short-term representation, time-frequency resolution compromises, ...
- There are situations where LP provides a sub-optimal filter estimate. In particular, when modelling voiced speech, LP methods can be adversely affected by spectral fine structure.
- The LP based approaches do not respect many important perceptual properties of hearing (e.g., non-uniform critical-band representation).
- Conventional LP techniques are based on linear model of speech production, thus are at a disadvantage for encoding non-speech signals (e.g., music, speech in background, ...).

The paper describes new audio coding technique that employs AR modelling applied to approximate the instantaneous energy (squared Hilbert envelope) of relatively long-term critical-band-sized sub-band signals. We have shown in our earlier work that the information carried by the approximated envelopes in sub-bands is sufficient to design very low bit-rate speech coder generating intelligible output of synthetic quality [1]. In this work, we focus on efficient coding of residual information (Hilbert carriers) to achieve higher quality of the re-synthesized signal. Our goal is to design an audio coder that provides broadcast radio-like quality around $10-20$kbps.

The paper is organized as follows : In Section 2, we give basic description of the proposed encoder. In Section 3, the decoding-side is described. Section 4 describes the experiments we conducted to validate the approach using objective quality measurements.

2 Encoding

State-of-the-art speech coding systems are based on processing of individual short-term frames ($20-30$ms), and rely on speech-specific source-system model [2]. The spectral envelopes are typically estimated by Temporal-Domain Linear Prediction (TDLP) technique [3], in which the power spectrum of each short-term frame is approximated by the AR model. Temporal evolution of the spectral envelopes of speech is thus captured in the 2-dimensional time-frequency plane. However, as we are interested in coding both speech and non-speech signals, it is beneficial to consider alternatives to source-system models and TDLP. Furthermore, the complexity of current audio coders based on LP approach becomes very high, with little room for additional substantial improvements (quality versus efficiency).

Recently, new technique utilizing LP to model temporal envelopes of input signal has been proposed [4]. More precisely, Hilbert envelope (squared magnitude of an analytic signal), which yields a good estimate of instantaneous energy of the signal, can be parameterized by Frequency Domain Linear Prediction (FDLP). FDLP represents frequency domain analogue of the TDLP.

The FDLP technique can be summarized as follows : To get an all-pole approximation of the squared Hilbert envelope, first the Discrete Cosine Transform (DCT) is applied to a given audio segment. Next, the autocorrelation LP technique is applied to the DCT transformed signal. The Fourier transform of the impulse response of the resulting all-pole model approximates the squared Hilbert envelope of the signal. The FDLP technique is explained in further detail in [4].

Just as LP fits an all-pole model to the power spectrum of the input signal, FDLP fits an all-pole

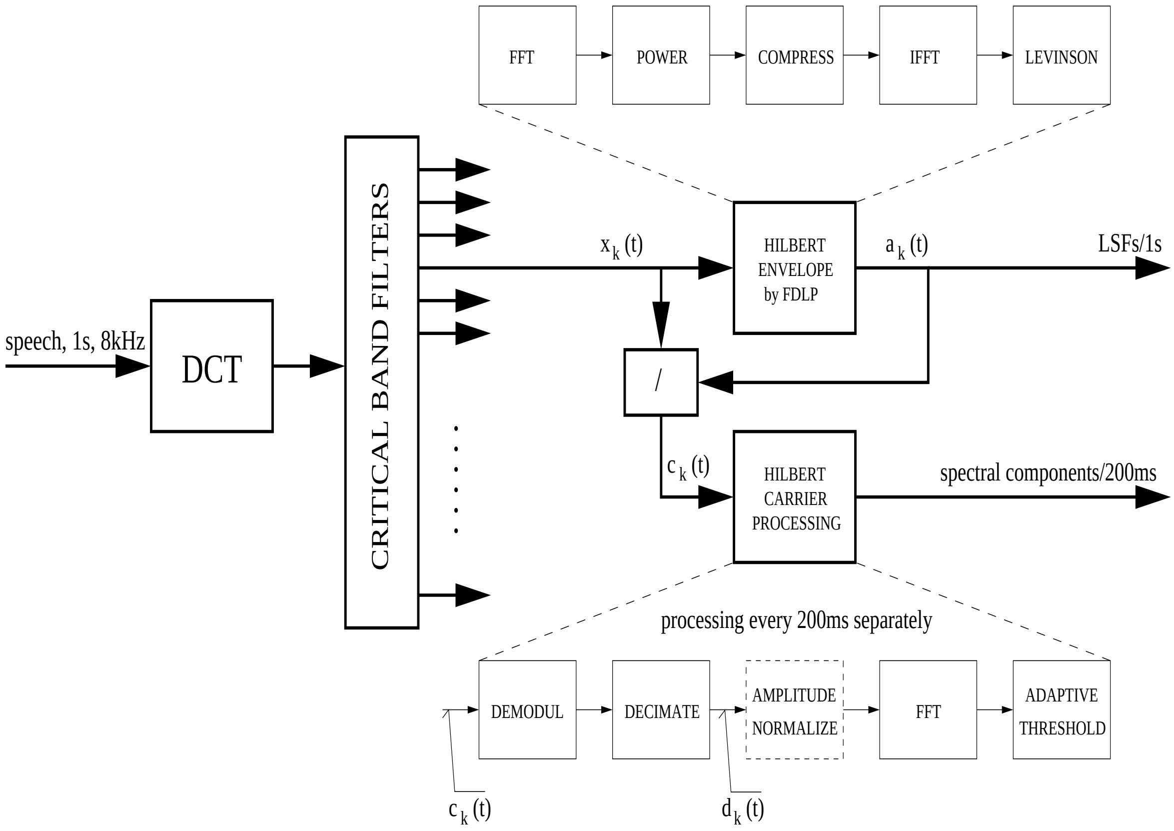

FIG. 1 – *Simplified structure of the proposed encoder.*

model to the squared Hilbert envelope of the signal. As discussed later, this approach can be exploited to approximate temporal envelope of the signal in individual frequency sub-bands. This presents an alternate representation of signal in the 2-dimensional time-frequency plane that can be used for audio coding operating on bit-rates comparable to traditional TDLP based speech coders.

2.1 Parameterizing temporal envelopes in critical sub-bands

First, the signal is divided into 1000ms long non-overlapping temporal segments which are transformed by DCT into the frequency domain, and later processed independently. To emulate auditory-like frequency selectivity of human hearing, we apply N_{BANDs} Gaussian functions (N_{BANDs} denotes number of critical sub-bands), equally spaced on the Bark scale with standard deviation $\sigma = 1$Bark and center frequency F_k, to derive sub-segments of the DCT transformed signal. The Bark scale from Perceptual Linear Prediction (PLP) analysis [5] is applied. FDLP technique is performed on every sub-segment of the DCT transformed signal that represents the frequency range of the sub-band (its time-domain equivalent obtained by inverse DCT is denoted as $x_k(t)$, where k means frequency sub-band). Resulting approximations of Hilbert envelopes in sub-bands are denoted as $a_k(t)$.

2.2 Spectral transform linear prediction

Well-known properties of LP that would normally apply to power spectra of the signal (such as better fitting of peaks than dips) apply in the case of FDLP to Hilbert envelopes. In order to control the balance between modelling peaks and dips of the envelope, Spectral Transform Linear Prediction (STLP) technique [6] is used.

2.3 Excitation of FDLP in frequency sub-bands

To reconstruct the signal in each critical-band-sized sub-band, the additional component – Hilbert carrier $c_k(t)$ is required. $c_k(t)$ represents time-domain signal. Modulating $c_k(t)$ with approximated temporal envelope $a_k(t)$ in each critical sub-band yields the original $x_k(t)$ (see e.g., [7] for mathematical explanation).

As apparent, $c_k(t)$ is analogous to excitation signal in TDLP. Utilizing $c_k(t)$ leads to perfect reconstruction of $x_k(t)$ in sub-band k and, after combining the sub-bands, in perfect reconstruction of the overall input signal.

We experimented in past with various approaches to efficiently encode $c_k(t)$ on very low bit-rates [1]. In this work, we focus on improving the quality for speech and audio signals targeting a bit-rate around $10 - 20$kbps.

2.3.1 Processing Hilbert carriers

FDLP approach provides set of LP coefficients that have similar characteristics in all sub-bands. It would be convenient for the subsequent processing if all Hilbert carriers $c_k(t)$ were also similar. This can be achieved by demodulating $c_k(t)$ (shifting Fourier spectrum of $c_k(t)$ from F_k to 0 Hz). Since modulation frequency F_k of each sub-band is known, we exploit standard procedure to demodulate $c_k(t)$ through the concept of *analytic signal* $z_k(t)$. $z_k(t)$ is the complex signal that has zero-valued spectrum for negative frequencies. To demodulate $c_k(t)$, we perform scalar multiplication $z_k(t).c_k(t)$. Demodulated carrier in each sub-band is lowpass filtered, to preserve only the down-shifted spectral components, and down-sampled. Frequency width of the lowpass filter as well as the down-sampling ratio is given by the frequency width of the Gaussian window (the cutoff frequencies correspond to 40dB decay in magnitude with respect to F_k) for a particular critical sub-band. The resulting time-domain signal (denoted as $d_k(t)$) represents demodulated and down-sampled Hilbert carrier $c_k(t)$ that has similar properties in all sub-bands (e.g., relevancy of its Fourier spectral components). $d_k(t)$ is a complex sequence, because its Fourier spectrum is not conjugate symmetric. Perfect reconstruction of $c_k(t)$ from $d_k(t)$ can be done by reversing all the pre-processing steps.

Since Hilbert carriers $c_k(t)$ change with fundamental frequency of the input signal, they are split into 200ms long non-overlapping sub-segments and processed independently.

2.3.2 Encoding of demodulated Hilbert carriers

The Hilbert envelopes $a_k(t)$ and complex valued demodulated Hilbert carriers $d_k(t)$ carry all the information necessary to entirely reconstruct $x_k(t)$. If the original Hilbert envelope is used to derive $d_k(t)$, then $|d_k(t)| = 1$, and only the phase information from $d_k(t)$ would be required for perfect reconstruction. However, since FDLP yields only approximation of the original Hilbert envelopes, $d_k(t)$ in general will not be perfectly flat and both components of complex sequence are required.

The coder implemented is an "adaptive threshold coder" applied on Fourier spectrum of $d_k(t)$, independently in each sub-band, where only the spectral components whose magnitudes are above the threshold are transmitted. The threshold is dynamically adapted to meet a required number of transmitted spectral components, described later in Section 4.1. The quantized value of both the magnitude and the phase for each selected spectral component are transmitted.

The whole encoder is depicted in Fig. 1.

3 Decoding

In order to reconstruct the input signal, the carrier $c_k(t)$ in each critical sub-band needs to be re-generated and then modulated by temporal envelope $a_k(t)$ obtained using FDLP.

A general scheme of the decoder, given in Fig. 2, is relatively simple. It inverts the steps performed at the encoder. The decoding operation is also applied on each (1000ms long) input segment independently. The decoding steps are :

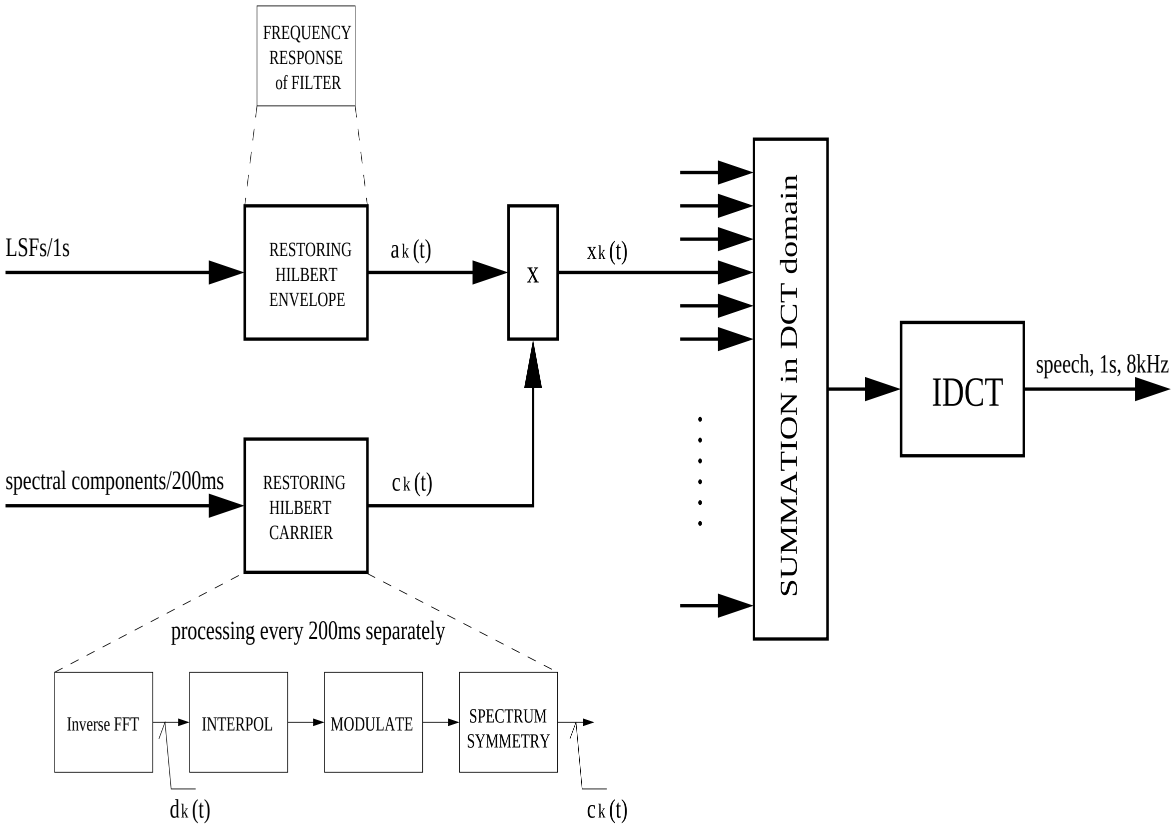

Fig. 2 – *Simplified structure of the proposed decoder.*

- Signal $d_k(t)$ is reconstructed using inverse Fourier transform from transmitted complex spectral components. $d_k(t)$ is then up-sampled to the original rate and modulated on F_k (i.e., its Fourier spectrum is frequency-shifted and post-processed to be conjugate symmetric). This results in the reconstructed Hilbert carrier $c_k(t)$.
- Temporal envelope $a_k(t)$ is reconstructed from transmitted AR model coefficients. The temporal trajectory $x_k(t)$ is obtained by modulating $c_k(t)$ with $a_k(t)$.

The above steps are performed in all frequency sub-bands. Finally :

- The temporal trajectories $x_k(t)$ in each critical sub-band are projected to the frequency domain by DCT and summed.
- A "de-weighting" window is applied to compensate for the effect of Gaussian windowing of DCT sequence at the encoder.
- Inverse DCT is performed to reconstruct 1000ms long output signal (segment).

Fig. 3 shows time-frequency characteristics of the proposed coder generated from randomly selected speech example.

4 Experiments

All experiments were performed with speech and audio signals sampled at $F_s = 8$kHz. We used decomposition into $N_{BANDs} = 13$ critical sub-bands, which roughly corresponds to partition of one sub-band per 1bark.

FDLP approximating Hilbert envelope in each frequency sub-band $a_k(t)$ is represented by Line Spectral Frequencies (LSFs). Previous informal subjective listening tests, aimed at finding proper approximations $a_k(t)$ of temporal envelopes, showed that for coding the 1000ms long audio segments, the "optimal" AR model $N_{LSFs} = 20$ [1]. Corresponding LSFs then can be quantized using $N_{BITs} = 4$ bits per LSF, and we achieve bit-rates $\sim$ 80bps per critical sub-band.

When using conventional autocorrelation all-pole method for deriving the FDLP AR models, the Hilbert envelope peaks seem overemphasized and the decoded signal sounds reverberant. This is especially true when low model order is involved in FDLP. STLP de-emphasizes the peaks and thus significantly reduces this reverberation. In our experiments, we use STLP compression factor $r = 0.1$.

4.1 Objective quality tests on Hilbert carrier

We used Itakura-Saito (I-S) distance measure [8] to experiment with threshold values on Fourier spectrum of $d_k(t)$ for reconstructing the Hilbert Carriers $c_k(t)$ at the decoder.

The measure was used to evaluate performance as a function of variable number of Fourier spectral components for the reconstruction of $c_k(t)$ (this number is always constant over all sub-bands) while keeping all other parameters constant.

The performance was tested on a sub-set of TIMIT – standard speech database [9], containing 380 speech sentences sampled at $F_s = 8$kHz. A total of about 20 minutes of the data was used. The experimental procedure was at follows :

- I-S measure was performed on short-term frames (30ms frame-length, 7.5ms frame-skip).
- Encoded sentences were compared to original sentences measuring the I-S distance between them. The lower values of I-S measure indicate smaller distance and better speech quality.
- As suggested in [10], to exclude unrealistically high spectral distance values, 5% of frames with the highest I-S distances were discarded from the final evaluation. This method ensures a reasonable overall measure of performance.

Fig. 4 shows the mean I-S distance value computed over all 380 sentences of TIMIT DB sub-set as a function of the number of Fourier spectral components used to reconstruct spectrum of demodulated Hilbert carrier $d_k(t)$ in each critical sub-band. The I-S distance shows marked improvement when the number of spectral components increased from 30 to 80. The error bars indicate standard deviations.

In order to develop a heuristic appreciation of the I-S distances, we repeated the above objective test with 3GPP-Adaptive Multi Rate (AMR) speech codec at 12.2kbps [11] on the same database, and show the results in Fig. 4. Similarly to the proposed coder, 5% of the frames with the highest I-S distance were discarded. We can see that if $d_k(t)$ is reconstructed from $\sim$ 65 Fourier spectral components (in each critical sub-band, per 200ms), the proposed coder achieves similar performance with respect to the chosen objective measure. Informal subjective results showed that the speech quality was comparable to that of AMR 12.2, while the quality for music signals was noticeably better.

5 Conclusions

A novel variable bit-rate audio coding technique based on processing relatively long temporal segments of audio signal in critical-band-sized sub-bands has been proposed and evaluated. The coder architecture allows to easily control the quality of reconstructed sound and the final bit-rate, thus making it suitable for variable bandwidth channels.

We describe experiments focused on efficient representation of excitation signal for the proposed FDLP coder. Such parameter setting does not indeed correspond to "optimal" approach (e.g., we use uniform spectral parameterization of Hilbert carriers in all sub-bands, ...). We expect that further improvements are possible when the proposed technique is applied in conjuction with the currently dominant analysis-by-synthesis coding schemes.

Preliminary experiments show the promise of a coder capable of encoding speech and music signals at an average data rate of $10 - 20$kbps.

6 Acknowledgments

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.

Références

[1] Motlicek P., Hermansky H., Garudadri H., Srinivasamurthy N., "Speech Coding Based on Spectral Dynamics", *technical report IDIAP-RR 06-05*, <http ://www.idiap.ch>, January 2006.

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

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

[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 Quality", *Prentice-Hall, Advanced Reference Series*, Englewood Cliffs, NJ, 1988.

[9] Fisher W. M, et al., "The DARPA speech recognition research database : specifications and status", *In Proc. DARPA Workshop on Speech Recognition*, pp. 93-99, February 1986.

[10] Hansen J. H. L., Pellom B., "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms", *In Proc. of ICSLP*, Vol. 7, pp. 2819-2822, Sydney, Australia, December 1998.

[11] 3GPP TS 26.071, "AMR speech CODEC", General description, <http ://www.3gpp.org/ftp/Specs/html-info/26071.htm>.

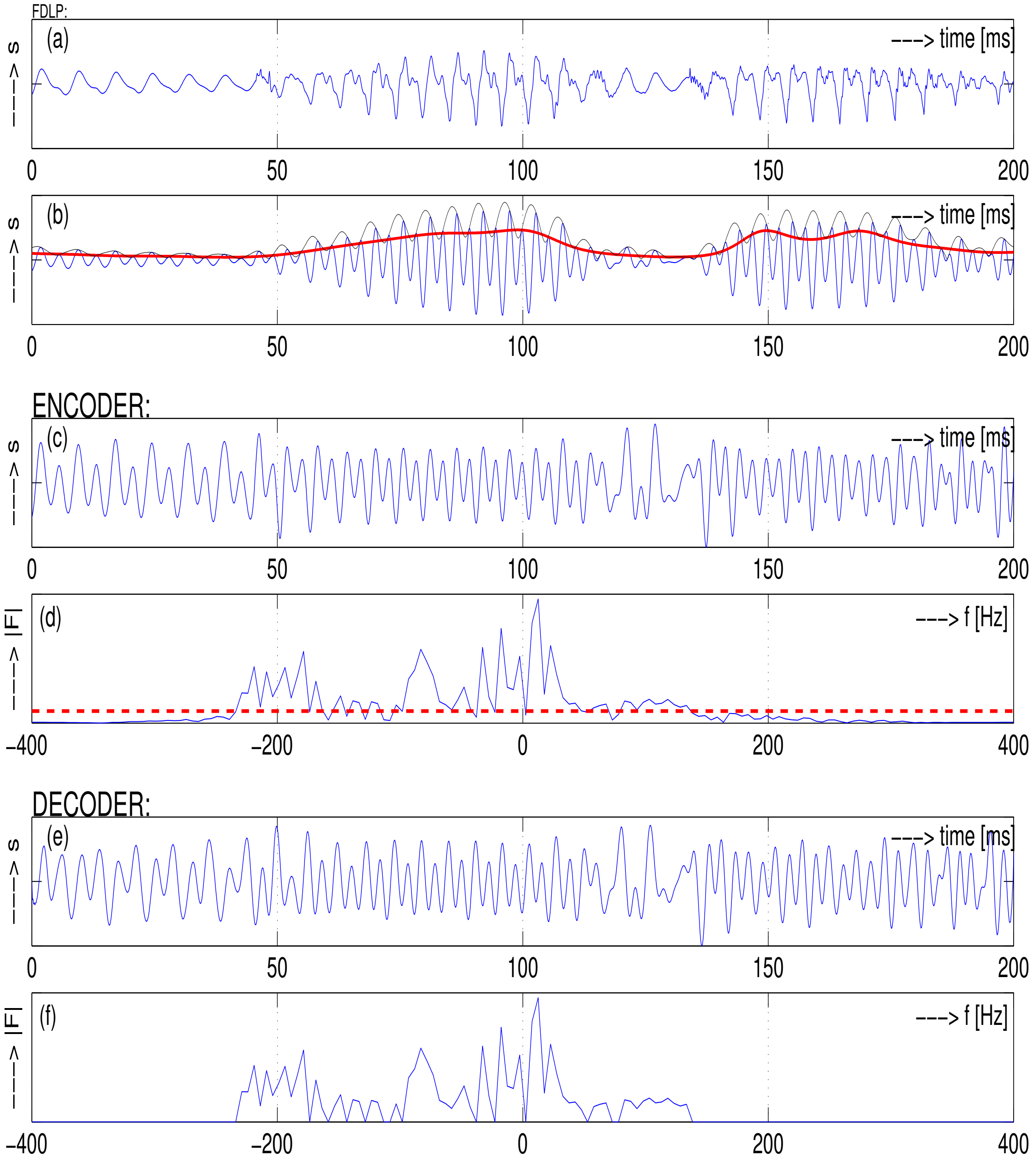

Fig. 3 – *Time-Frequency characteristics generated from randomly selected speech example : (a)* $200ms$ *segment of the input signal. (b)* $x_3(t)$ *sequence (frequency sub-band* $k = 3$*, center frequency* $F_3 = 351Hz$ *), thin upper line represents original Hilbert envelope, solid upper line represents its FDLP approximation. (c) Original Hilbert carrier* $c_3(t)$*. (d) Magnitude Fourier spectral components of the demodulated Hilbert carrier* $d_3(t)$*, the solid line represents the selected threshold. (e) Reconstructed Hilbert carrier* $c_3(t)$ *in the decoder. (f) Magnitude Fourier spectral components of* $d_3(t)$ *post-processed by adaptive threshold.*

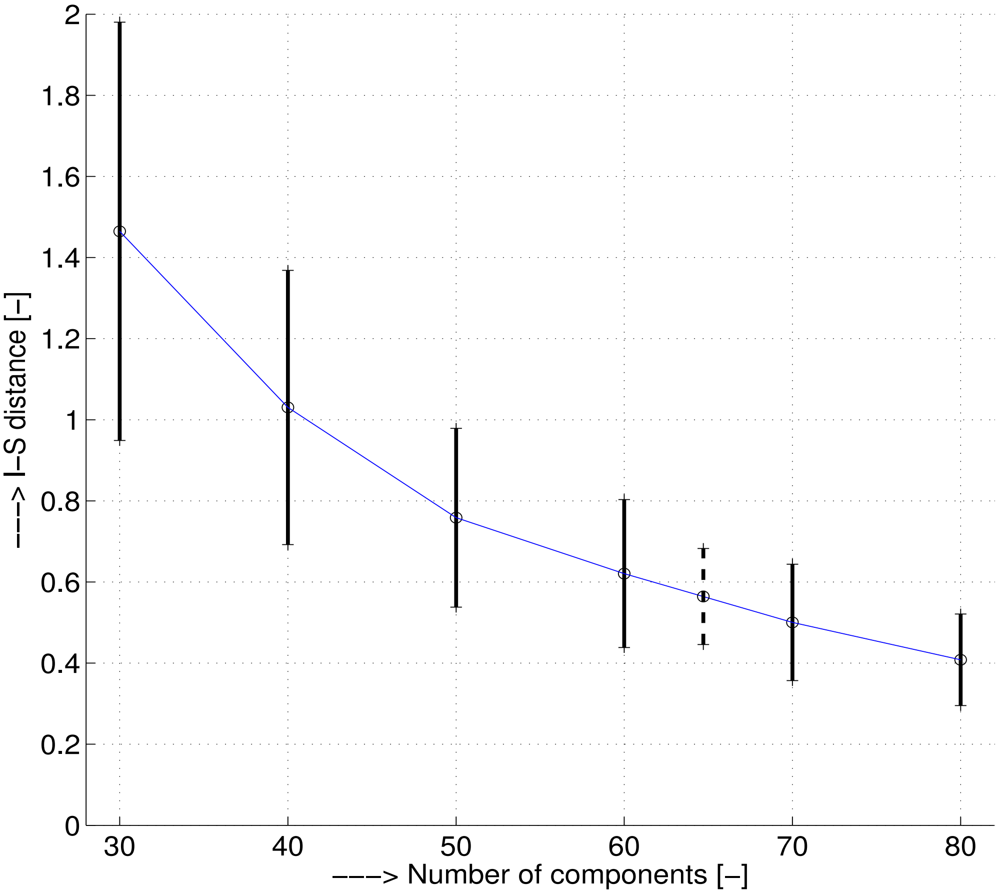

Fig. 4 – *Global mean I-S distance measure of the proposed coder as a function of the Fourier spectral components used to reconstruct* $d_k(t)$ *in each critical sub-band. Dashed line shows the performance of the 3GPP-AMR speech codec at* 12.2*kbps.*