Non-Uniform Speech/Audio Coding Exploiting Predictability of Temporal Evolution of Spectral Envelopes*

Petr Motlicek[12], Hynek Hermansky[123], Sriram Ganapathy[13], and Harinath Garudadri[4]

[1] IDIAP Research Institute,
Rue du Simplon 4, CH-1920, Martigny, Switzerland
{motlicek,hynek,ganapathy}@idiap.ch
[2] Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, Brno, 612 66, Czech Republic
[3] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[4] Qualcomm Inc., San Diego, California, USA
hgarudad@qualcomm.com

Abstract. We describe novel speech/audio coding technique designed to operate at medium bit-rates. Unlike classical state-of-the-art 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 speech/audio coder to provide broadcast radio-like quality audio around $15-25$kbps. Obtained objective quality measures, carried out on standard speech recordings, were compared to the state-of-the-art 3GPP-AMR speech coding system.

Key words: Audio coding, audio signal processing, linear predictive coding, modulation coding, lossy compression

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 [1]. This short-term analysis is based on the assumption that the speech

* This work was partially supported by grants from ICSI Berkeley, USA; the Swiss National Center of Competence in Research (NCCR) on "Inter active 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.

signal is stationary over these segment durations. Techniques like Linear Prediction (LP), which is able to efficiently approximate short-term power spectra by Auto-Regressive (AR) model [2], are applied.

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

- When the signal dynamics are described by a sequence of short-term vectors, many issues come up, like windowing, proper sampling of short-term representation, time-frequency resolution compromises, etc.
- There are situations where LP provides a sub-optimal filter estimate. In particular, when modeling 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 have difficulties encoding non-speech signals (e.g., music, speech in background, etc.).

Over the past decade, research in speech/audio coding has been focused on high quality/low latency compression of wide-band audio signals. However, new services such as Internet broadcasting, consumer multimedia, or narrow-band digital AM broadcasting are emerging. In such applications, new challenges have been raised, such as resiliency to errors and gaps in delivery. Furthermore, many of these services do not impose strict latency constraints, i.e., the coding delay is less important as compared to bit-rate and quality requirements.

This paper describes a new coding technique that employs AR modeling applied to approximate the instantaneous energy (squared Hilbert envelope (HE)) of relatively long-term critical-band-sized sub-band signals. It has been shown in our earlier work that based on approximating the envelopes in sub-bands we can design very low bit-rate speech coder giving intelligible output of synthetic quality [3]. In this work, we focus on efficient coding of residual information (Hilbert carriers (HCs)) to achieve higher quality of the re-synthesized signal. The objective quality scores based on Itakura-Saito (I-S) distance measure [4] and Perceptual Evaluation of Speech Quality (PESQ) [5] are used to evaluate the performance of the proposed coder on challenging speech files sampled at 8kHz.

The paper is organized as follows: In Section 2, a basic description of the proposed encoder is given. 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

New techniques utilizing LP to model temporal envelopes of input signal have been proposed [6, 7]. More precisely, HE (squared magnitude of an analytic sig-

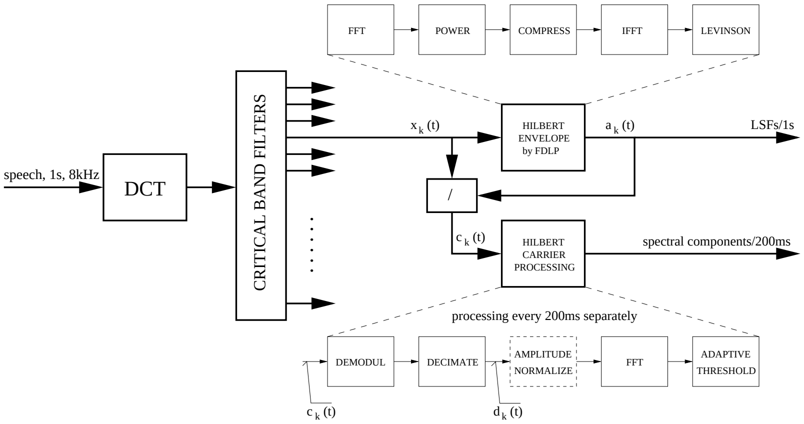

Fig. 1. *Simplified structure of the proposed encoder.*

nal), which yields a good estimate of instantaneous energy of the signal, can be parameterized by Frequency Domain Linear Prediction (FDLP) [7]. FDLP represents frequency domain analogue of the traditional Time Domain Linear Prediction (TDLP) technique, in which the power spectrum of each short-term frame is approximated by the AR model.

The FDLP technique can be summarized as follows: To get an all-pole approximation of the squared HE, 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 HE of the signal.

Just as TDLP fits an all-pole model to the power spectrum of the input signal, FDLP fits an all-pole model to the squared HE 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.

2.1 Parameterizing temporal envelopes in critical sub-bands

The graphical scheme of the whole encoder is depicted in Fig. 1. First, the signal is divided into 1000ms long temporal segments which are transformed by DCT into the frequency domain, and later processed independently. In order to avoid possible artifacts at segment boundaries, 10ms overlapping is used.

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. FDLP technique is performed on every sub-segment of the DCT transformed signal (its time-domain equivalent obtained by inverse DCT is denoted as $x_k(t)$, where k denotes frequency sub-band). Resulting approximations of HEs in sub-bands are denoted as $a_k(t)$.

2.2 Excitation of FDLP in frequency sub-bands

To reconstruct the signal in each critical-band-sized sub-band, the additional component – Hilbert carrier (HC) $c_k(t)$ is required (residual of the LP analysis represented in time-domain). Modulating $c_k(t)$ with approximated temporal envelope $a_k(t)$ in each critical sub-band yields the original $x_k(t)$ (refer [8] for mathematical explanation).

Clearly, $c_k(t)$ 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.

Processing Hilbert carriers (HCs): For convenience in processing and encoding, we need the sub-band carrier signals to be low-pass. 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 employ 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 low-pass filtered and down-sampled. Frequency width of the low-pass filter as well as the down-sampling ratio is determined using 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 HC $c_k(t)$. $d_k(t)$ 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 HCs $c_k(t)$ are quite non-stationary, they are split into 200ms long sub-segments (10ms overlap for smooth transitions) and processed independently.

Encoding of demodulated HCs: Temporal envelopes $a_k(t)$ and complex valued demodulated HCs $d_k(t)$ carry the information necessary to reconstruct $x_k(t)$. If the original HE 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 HEs, $|d_k(t)|$ 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 having magnitudes 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 values of magnitude and phase for each selected spectral component are transmitted.

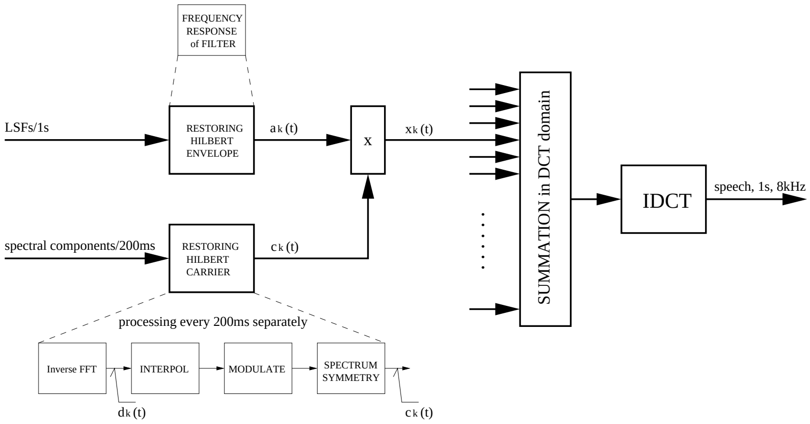
FREQUENCY RESPONSE of FILTER
LSFs/1s
RESTORING HILBERT ENVELOPE
a k (t)
x
x k (t)
SUMMATION in DCT domain
IDCT
speech, 1s, 8kHz
spectral components/200ms
RESTORING HILBERT CARRIER
c k (t)
processing every 200ms separately
Inverse FFT
INTERPOL
MODULATE
SPECTRUM SYMMETRY
d k (t)
c k (t)

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

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 graphical 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: **(a)** Signal $d_k(t)$ is reconstructed using inverse Fourier transform of transmitted complex spectral components. $d_k(t)$ is then up-sampled to the original rate and modulated on sinusoid at F_k (i.e., its Fourier spectrum is frequency-shifted and post-processed to be conjugate symmetric). This results in the reconstructed HC $c_k(t)$. **(b)** 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: **(a)** The temporal trajectories $x_k(t)$ in each critical sub-band are projected to the frequency domain by DCT and summed. **(b)** A “de-weighting” window is applied to compensate for the effect of Gaussian windowing of DCT sequence at the encoder. **(c)** Inverse DCT is performed to reconstruct 1000ms long output signal (segment). Fig. 3 shows time-frequency characteristics of the proposed coder for a randomly selected speech sample.

4 Experiments

All experiments were performed with speech 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 bark.

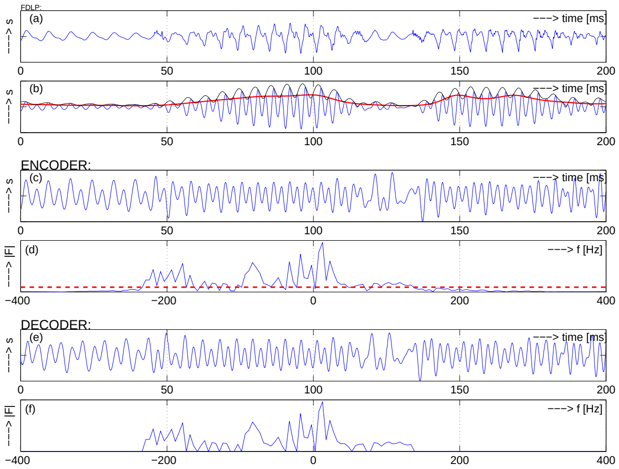

Fig. 3. *Time-Frequency characteristics generated from randomly selected speech sample: (a)* $200ms$ *segment of the input signal. (b)* $x_3(t)$ *sequence (frequency sub-band* $k = 3$*, center frequency* $F_3 = 351Hz$ *), thin upper line represents original HE, solid upper line represents its FDLP approximation.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

FDLP approximating HE in each frequency sub-band $a_k(t)$ is represented by Line Spectral Frequencies (LSFs). Previous informal subjective listening tests, aimed at finding sufficient approximations $a_k(t)$ of temporal envelopes, showed that for coding the 1000ms long audio segments, the "optimal" AR model is of order $N_{LSFs} = 20$ [3].

4.1 Objective quality tests on HC

We used Itakura-Saito (I-S) distance measure [4] as a simple method together with ITU-T P.862 PESQ objective quality tool [5] to adjust the threshold values on Fourier spectrum of $d_k(t)$ for reconstructing HCs $c_k(t)$ at the decoder.

These measures were 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 fixing all other parameters.

The performance was tested on a sub-set of TIMIT – speech database [9], containing 380 speech sentences sampled at F_s = 8kHz. A total of about 20 minutes of speech was used for the experiments.

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

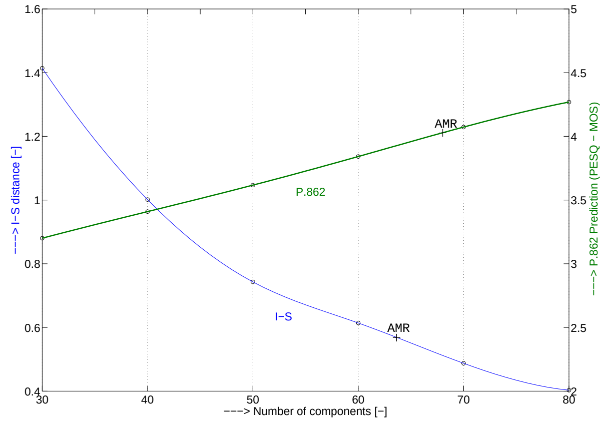

Fig. 4. *Global mean I-S distance measure of the proposed coder as a function of the number of Fourier spectral components used to reconstruct* $d_k(t)$ *in each critical sub-band. "+" marks the performance of the 3GPP-AMR speech codec at* 12.2*kbps.*

discarded from the final evaluation. This method ensures a reasonable measure of overall performance.

PESQ scores were also computed for the reconstructed signal. The quality estimated by PESQ corresponds to the average user perception of the speech sample under assessment PESQ – MOS (Mean Opinion Score).

Fig. 4 shows the mean I-S distance value as well as mean PESQ score computed over all TIMIT DB sub-set as a function of the number of Fourier spectral components used to reconstruct spectrum of demodulated HC $d_k(t)$ in each critical sub-band. Both objective quality measures show marked improvement when the number of spectral components is increased from 30 to 80.

We repeated the above objective tests with 3GPP-Adaptive Multi Rate (AMR) speech codec at 12.2kbps [11] on the same database, and show the results in Fig. 4. The results indicate that if $d_k(t)$ is reconstructed from ~ 65 Fourier spectral components (in each critical sub-band, per 200ms), the proposed coder achieves similar performance to AMR codec with respect to the chosen objective measures. 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.

In this paper, we do not discuss the quantization block and entropy coder. However, in additional informal experiments, LSFs describing temporal envelopes $a_k(t)$ as well as the selected spectral components of $d_k(t)$ were quantized (split VQ technique). These preliminary experiments show the promise of a coder in encoding speech and music signals at an average bit-rate of $15 - 25$kbps.

5 Conclusions

A novel variable bit-rate speech/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. The coding technique representing input signal in frequency sub-bands is inherently more robust to losing chunks of information, i.e., less sensitive to dropouts. This can be of high importance for any Internet protocol service.

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, uniform quantization of LSFs, simple Gaussian decomposition, etc). All these would be the direction of future research in improving the proposed coder. To convert the proposed speech/audio coding technique into a real application, formal subjective tests need to be made both on speech and music recordings.

References

1. Spanias A. S., "Speech Coding: A Tutorial Review", *In Proc. of IEEE*, Vol. 82, No. 10, October 1994.
2. Makhoul J., "Linear Prediction: A Tutorial Review", *in Proc. of IEEE*, Vol. 63, No. 4, April 1975.
3. 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.
4. Quackenbush S. R., Barnwell T. P., Clements M. A., "Objective Measures of Speech Quality", *Prentice-Hall, Advanced Reference Series*, Englewood Cliffs, NJ, 1988.
5. ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", ITU, Geneva, Switzerland, 2001,
6. 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.
7. 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.
8. 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.
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>.