WIDE-BAND PERCEPTUAL AUDIO CODING BASED ON FREQUENCY-DOMAIN LINEAR PREDICTION

Petr Motlicek[1], *Vijay Ullal*[2], *Hynek Hermansky*[1,3]

[1]IDIAP Research Institute, Martigny, Switzerland
[2]The International Computer Science Institute, Berkeley, CA, USA
[3]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{motlicek,hynek}@idiap.ch, vullal@icsi.berkeley.edu

ABSTRACT

In this paper we propose an extension of the very low bit-rate speech coding technique, exploiting predictability of the temporal evolution of spectral envelopes, for wide-band audio coding applications. Temporal envelopes in critically band-sized sub-bands are estimated using frequency domain linear prediction applied on relatively long time segments. The sub-band residual signals, which play an important role in acquiring high quality reconstruction, are processed using a heterodyning-based signal analysis technique. For reconstruction, their optimal parameters are estimated using a closed-loop analysis-by-synthesis technique driven by a perceptual model emulating simultaneous masking properties of the human auditory system. We discuss the advantages of the approach and show some properties on challenging audio recordings. The proposed technique is capable of encoding high quality, variable rate audio signals on bit-rates below 1bit/sample.

Index Terms— Audio signal processing, data compression, linear predictive coding, Hilbert transform

1. INTRODUCTION

Recently, due to increasing popularity of new wireless and Internet services, parametric coding of speech and audio signals has became a hot topic among researchers as well as in standardization communities. The parametric coding allows for good performance at low bit-rates [1]. It relies on signal models that describe the signal by few physical parameters. Audio coders based on parametric representation give very good quality of the reconstructed signal as long as the model properly obeys properties of the input signal. However, when it does not, the reconstructed signal may be of very low perceived quality. Current requirements on speech and audio coders are extensive, such as allowing for encoding input signal with different sampling rates, performing on various bit-rates, and ability to encode any kind of the input signal (speech, music, signals with mixed audio sources, transient signals), etc. The task becomes extremely difficult for any singular technique and current state-of-the art systems combine several approaches (modes). The coding mode is selected often based on a closed-loop analysis-by-synthesis technique, see e.g., extended AMR wideband encoder [2].

Just as the audio CD stimulated a tremendous boom for the record industry, the booming popularity of the Internet has opened up new avenues for the promotion and distribution of music to consumers. Music over Internet Protocol (IP) generally falls into one of two categories. "Streaming audio" allows visitors to a Web site to hear a selected sample in real time, without the wait of first transferring the entire file to the listener's local hard drive. A "download", on the other hand, means that the music is copied to the user's hard drive, allowing the user to subsequently listen to it without being connected to the Web site.

Furthermore, using IP network as a new service platform, especially telecommunication providers foresee new opportunities emerging, as services over IP reduce costs, maximize bandwidth efficiency and introduce new possibilities for the customer. As a consequence, one may think of other services such as live audio and video streaming applications (e.g., radio and TV broadcast over IP, multicast of a lecture, etc.). All these non-interactive applications require timely delivery, but unlike voice communications services, do not have strict latency constraints. The critical constraints are low error in transmission, breaks in continuity of delivery, and the overall signal quality, but the delay in information arrival is not the most critical issue.

A novel speech coding system that was proposed recently [3], employs predictability of the temporal evolution of spectral envelopes of a speech signal using Frequency-Domain Linear Prediction (FDLP) [4, 5]. Unlike [4], the new technique employs FDLP to approximate relatively long (up to 1000ms) segments of Hilbert envelopes in individual frequency sub-bands. As reported in [3], with a noise excitation, the technique is capable of yielding a highly intelligible but unvoiced speech signal at bit-rates well below 1kbps. In this paper we explore the use of this technique in high-quality speech and audio wide-band multi-rate system codec, to be employed in IP services where longer algorithmic delay is allowed.

The remaining part of this paper is organized as follows: Section 2 contains an overview of the proposed system. The description of the FDLP technique employed to estimate temporal envelopes in critically-band-sized sub-band signals is given in Section 3. Subsequently, the algorithm used to parameterize residual signals is presented in Section 4. Section 5 describes implementation details of the proposed coder, and Section 6 presents experimental results. Section 7 contains discussions and concludes on the work.

2. GENERAL DESCRIPTION OF THE SYSTEM

The presented coding system, depicted in Fig. 1 includes the following steps: **(1)** The input audio signal is split into a number of

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.

(hundreds of ms long) non-overlapping frames. **(2)** Each frame is DCT transformed and partitioned into non-equal length segments to obtain critical-band-sized sub-bands of the DCT transformed signal. **(3)** FDLP approximation is applied on every sub-band by carrying out autocorrelation LPC analysis on the segments of DCT transformed signal, yielding LSP descriptors of FDLP models. Obtained all-pole models approximate squared Hilbert envelopes in sub-bands (described in details in Section 3). **(4)** The associated FDLP-LSP parameters are quantized. **(5)** The residual signal in each sub-band (this signal represents a carrier signal for the FDLP-encoded Hilbert envelope) is split into equal-length non-overlapping segments. **(6)** Each segment is heterodyned to DC range and Fourier transformed to yield spectral components of low-passed sub-band signal carriers. **(7)** Commensurate number of spectral components for each segment in each sub-band is estimated using a psychoacoustic model (employed in analysis-by-synthesis procedure) determining the maximum allowable noise energy in each critical sub-band such that the "distortion" noise remains inaudible. **(8)** Parameters of the selected spectral components are quantized.

In the decoder: **(1)** The residual signal in each sub-band is reconstructed, and **(2)** modulated with corresponding FDLP envelope. **(3)** Individual DCT contributions from each critical sub-band are summed, and **(4)** inverse DCT is applied to reconstruct output signal (frame).

3. ENVELOPE ESTIMATION

This section describes steps involved in the estimation of the temporal envelope in more detail. FDLP principle has been first introduced in [4] as the temporal noise shaping technique for transform coding. In our work, we apply FDLP to approximate relatively long temporal envelopes of the sub-band signal. However, to simplify the notation in this section, we present the full-band version of the technique. The sub-band based technique is identical but applied only to the appropriate parts of the DCT transformed signal.

Let us define the input discrete time-domain sequence as $s(n)$ for time samples $n = 1, \ldots, N$, where N denotes the segment length. Its Fourier power spectrum $P(\omega_k)$ (sampled at discrete frequencies $\omega_k = \frac{2\pi}{N}k$; $k = 1, \ldots, N$) is given as

$$P(\omega_k) = |S(e^{j\omega_k})|^2, \tag{1}$$

where $S(e^{j\omega_k}) = Z\{s(n)\}|_{z=e^{j\omega_k}}$. $Z\{.\}$ stands for the z-transformation. Later, let us use the notation $F\{.\}$ for Discrete Fourier Transform (DFT) which is equivalent to z-transform with $z = e^{j\omega_k}$. It has been shown, e.g., in [6], that classical Temporal-Domain Linear Prediction (TDLP) fits the discrete power spectrum of an all-pole model $\hat{P}(\omega_k)$ to $P(\omega_k)$ of the input signal.

Unlike TDLP, where the time-domain sequence $s(n)$ is modeled by linear prediction, FDLP applies linear prediction on a frequency-domain sequence. In our case, $s(n)$ is first DCT transformed. It can also be viewed as the symmetrical extension of $s(n)$ so that a new time-domain sequence $q(m)$ is obtained ($m = 1, \ldots, 2N$) and then DFT projected. We obtain the real-valued sequence $Q(\omega_k) = F\{q(m)\}$. We then estimate the frequency-domain prediction error $E(\omega_k)$ as a linear combination of $Q(\omega_k)$ consisting of p real prediction coefficients b_i

$$E(\omega_k) = Q(\omega_k) - \sum_{i=1}^{p} b_i Q(\omega_k - i). \tag{2}$$

The b_i are found so that the squared prediction error is minimized [6]. As noted above, in the case of TDLP, minimizing the

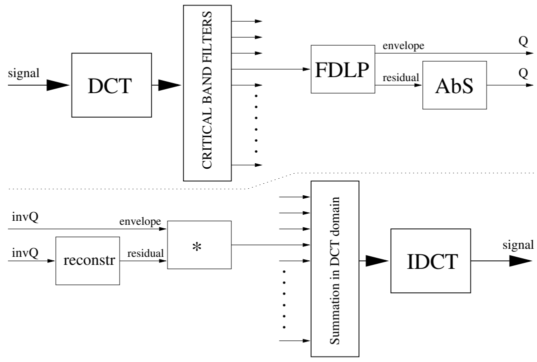

Fig. 1. Generalized scheme of the encoder/decoder with application of closed-loop analysis-by-synthesis technique.

total error is equivalent to the minimization of the integrated ratio of the signal spectrum $P(\omega_k)$ to its model approximation $\hat{P}(\omega_k)$

$$E_{TDLP} \approx \frac{1}{N} \sum_{k=1}^{N} \frac{P(\omega_k)}{\hat{P}(\omega_k)}. \tag{3}$$

In the case of FDLP, we can interpret $Q(\omega_k)$ as a discrete, real, causal, stable sequence (although consisting frequency samples). Its discrete power spectrum will be estimated through the concept of discrete Hilbert transform relationships [7]. $Q(\omega_k)$ can be expressed as the sum of $Q^e(\omega_k)$ and $Q^o(\omega_k)$, denoting an even sequence and an odd sequence, respectively; thus $Q(\omega_k) = Q^e(\omega_k) + Q^o(\omega_k)$. Its Fourier transform

$$\phi(m) = F\{Q(\omega_k)\} = \phi^R(m) + j\phi^I(m), \tag{4}$$

where R and I stand for real and imaginary parts of $\phi(m)$, respectively. It has been shown (e.g., [7]) that $\phi^R(m) = F\{Q^e(\omega_k)\}$ and $\phi^I(m) = F\{Q^o(\omega_k)\}$. By taking the Fourier transform of $Q^e(\omega_k)$, the original sequence $q(m)$ is obtained

$$F\{Q^e(\omega_k)\} = \phi^R(m) = const\ q(m). \tag{5}$$

The relations between $F\{Q^e(\omega_k)\}$ and $F\{Q^o(\omega_k)\}$, called the Kramers-Kronig relations, are given by the discrete Hilbert transform (partial derivatives of real and imaginary parts of an analytic function [8]), thus

$$\phi(m) = \phi^R(m) + j\phi^I(m) = const\big(q(m) + jH\{q(m)\}\big), \tag{6}$$

where $H\{.\}$ stands for Hilbert transformation. Power root $|\phi(m)|^2$ is called the squared Hilbert envelope. Prediction error is proportional to the integrated ratio of $|\phi(m)|^2$ and its FDLP approximation $A(m)^2$

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^2}{A^2(m)}. \tag{7}$$

Eq. 7 can be interpreted in such a way that the FDLP all-pole model fits squared Hilbert envelope of the symmetrically extended time-domain sequence $s(n)$. FDLP models the time-domain envelope in the same way as TDLP models the spectral envelope. Therefore, the same properties appear, such as accurate modeling of peaks rather than dips.

Further, the squared Hilbert envelope $|\phi(m)|^2$ is available and can be modified. Thus, e.g., compressing $|\phi(m)|^2$ by a root function

$[.]^{\frac{1}{r}}$ turns Eq. 7 into

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^{\frac{2}{r}}}{A^{\frac{2}{r}}(m)}. \tag{8}$$

As a consequence, the new model will fit dips more accurately than the original model. This technique has been proposed for TDLP in [9], and is applied also with FDLP in the current work.

In our experiments, the DCT input sequence is weighted by a set of Gaussian windows of variable temporal resolution, spaced following the Bark scale, as described in [3]. Gaussian windows span the whole DCT sequence. Therefore, we can individually exploit FDLP in each critically band-sized sub-band.

4. PROCESSING OF THE RESIDUALS

In the current work, we explore FDLP for high quality, efficient encoding of wide-band audio. Modulating the carrier with the FDLP envelope in each critical sub-band yields the original DCT sequence and allows for lossless reconstruction of the original signal. Thus, the carrier is analogous to the residual signal in TDLP.

Clearly, for coding efficiency, the residual signal representing the Hilbert carrier of the sub-band FDLP envelope cannot be transmitted in its original form, but needs to be efficiently coded to preserve its important components as accurately as possible for proper reconstruction of the coded signal.

To obtain uniform properties of the carriers in all sub-bands, they are first demodulated, i.e., its Fourier spectrum is shifted from center frequency F_k (given by the width of Gaussian window) to 0Hz. The demodulated carriers in each sub-band are low-pass filtered to preserve only the downshifted spectral components, and down-sampled. Frequency width of the low-pass filter as well as the down-sampling ratio is given by the frequency width of the Gaussian window. In our implementation, the cutoff frequencies correspond to 40dB decay in magnitude with respect to F_k, for each critical sub-band. In general, obtained pre-processed carriers are complex sequences. Perfect reconstruction can be accomplished by reversing all the pre-processing steps. Demodulated carriers are encoded in a form of their Fourier spectral components.

4.1. Psychoacoustic model

Since full-encoding of all components of residual signals would be highly inefficient for audio coding applications, we perform a closed-loop Analysis-by-Synthesis (AbS) technique which is used, in conjunction with a psychoacoustic model of simultaneous masking, to quantify the maximum amount of distortion at each point in the time-frequency plane. The model determines the maximum Signal-to-Mask Ratio (SMR), i.e., the largest difference between the intensity of the masking noise and the intensity of the masked tone expressed as Sound Pressure Levels (SPLs). Similarly, as in more conventional AbS coding techniques, SMRs are estimated on a short-term frame basis from the input audio signal. First, local masking thresholds (related to particular spectral peaks in SPL values) are estimated. We exploit the spreading function that has slopes of $+25$dB and -10dB per Bark. A convenient analytical expression is given by [10]

$$\begin{aligned} SF_{dB}(x) = \quad & 15.81 + 7.5(x + 0.474) \\ & -17.5\sqrt{1 + (x + 0.474)^2}\, dB, \end{aligned} \tag{9}$$

where the frequency variable x is in Barks and $SF_{dB}(x)$ in dB. After the thresholds have been identified, they are combined to form a global masking threshold and then combined with the Threshold in Quiet (TIQ). Finally, for each critical sub-band we compute the maximum SMR so that we divide the maximum SPL value of the signal by the estimated masking threshold belonging to a given sub-band. SMRs, obtained for particular time-frequency positions, determine the importance of the critical sub-bands of the input signal, and can be subsequently used for the bit allocation.

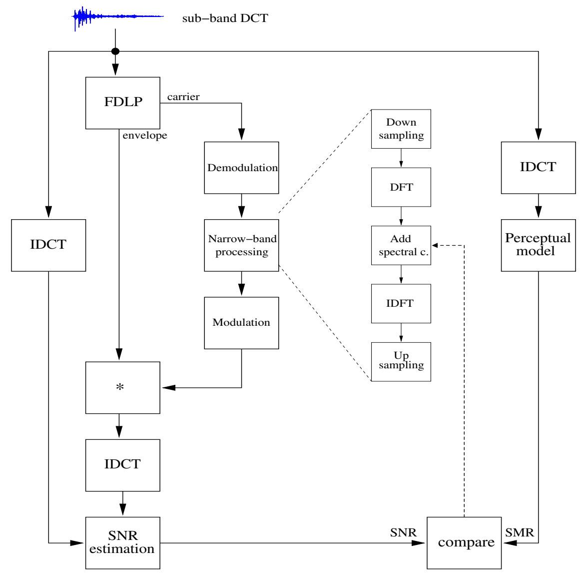

Fig. 2. Graphical scheme of the closed-loop analysis-by-synthesis technique applied on the encoder-side.

4.2. AbS technique

The whole scheme of the closed-loop AbS technique applied on the encoder-side is given in Fig. 2. AbS procedure consists of choosing the combination of parameters whose reconstructed signal is perceptually closest to the analysis signal in a given frequency sub-band. Specifically, short-term frame-based Signal-to-Noise Ratios (SNRs) of such reconstructed signal are compared to estimated SMRs for particular time-frequency positions. Using AbS, we control the number of Fourier spectral components. Spectral components are incrementally added to the Fourier spectrum (starting from the lowest frequencies). SNR of such reconstructed residual modulated by corresponding FDLP envelope is compared to the corresponding SMR value. AbS procedure is terminated when the estimated SNR value outperforms SMR value, which assures that amount of the distortion is below the masking threshold.

5. IMPLEMENTATION DETAILS OF AUDIO CODEC

All experiments were performed with audio signals sampled at $F_s =$ 48kHz. The frequency decomposition was done into $N_{BANDs} = 25$ sub-bands, corresponding to a partition of one sub-band per 1Bark. FDLP with the compressing factor $r = 10$ was exploited to approximate 1000ms long temporal envelopes in each critical sub-band. Corresponding LPCs (for FDLP model order equal to 20) were transformed into LSPs and vector quantized using 4bits per LSP parameter, so that we achieve a bit-rate $\sim$ 80bps per critical sub-band to encode FDLP envelopes. Residual signals corresponding to 1000ms long temporal envelopes were segmented into 200ms chunks and then processed independently (heterodyning, AbS procedure). AbS model returns a perceptually estimated number of spectral components to be used on the decoder-side to reconstruct the residuals in each critical sub-band. Practically, the estimated short-term SMRs and SNRs (in our experiments every 50ms), which coincide with a particular 200ms chunk, are individually evaluated and processed by the logical operation *AND*. Selected spectral components are uniformly quantized. We use 2bits for magnitudes and 3bits for phases. Supplemental information needs to be transmitted, such as quantized energy levels for FDLP envelopes and the residuals, and number of selected spectral components. In the experiments, we control the final bit-rate to be constant over 200ms chunks of the input signal. Although, according to perceptual masking results, AbS returns a different amount of spectral components to be quantized, we can control the total bit-rate by setting constraints on the selected number of encoded parameters. The process can work as a "water-filling" bit-allocation algorithm, i.e., bits are allocated to where they are needed most, according to perceptual results. The configuration was set to run at $\sim$ 30kbps.

On the decoder-side, we invert the steps performed at the encoder. The excitation in each critical sub-band is reconstructed. Processed 200ms chunks are concatenated to form 1000ms long frames and modulated by corresponding FDLP envelope.

6. EXPERIMENTAL RESULTS

The described audio coding system was tested on challenging wide-band audio samples. An attack of castanets is shown in Fig. 3. The top panel is the original waveform sampled at 48kHz. The second panel shows the reconstruction coded with uniform allocation of bits per time-frequency plane (no use of the AbS with psychoacoustic model). The third panel shows the reconstruction coded with FDLP-AbS technique controlled by the psychoacoustic model. In both cases the bit rate was fixed at 30kbps. Any other coding techniques, such as channel coding, were used. As expected, FDLP in conjunction with AbS procedure provides significantly better reconstruction of the sharp castanet attack (with lessened pre-echo effect). Informally, perceptual quality of the AbS reconstruction compared to uniform allocation, was also higher. Our informal experience indicates that AbS technique yields at least halving the bit rates in the FDLP framework. The testing waveforms for informal experiments were different audio signals (containing speech, music, speech between music, speech over music recordings). For comparison, in Fig. 3 we show also the reconstruction coded with AMR WB+ codec [2] performed on approximately same bit rates $\sim$ 30kbps.

7. DISCUSSIONS AND CONCLUSIONS

To conclude, the technique shows promise in several points:
(a) The technique is neither based on linear model of speech production, nor short-term analysis. Therefore, it is applicable for encoding any kind of audio signal with different sampling rates.
(b The technique follows perceptual properties of hearing (e.g., non-uniform critical band representation, simultaneous masking).
(c) FDLP method applied provides coding gain for transient signal types due to good time localization properties.
(d) The final bit-rate (related to required quality) of the reconstructed audio can be easily controlled. We can also simply control length of processed segments of the input signal, which is directly related to latency provided by the coding system (in the range of hundreds of milliseconds).
(e) The coding technique representing the input signal in frequency sub-bands is inherently more robust to losing chunks of information, i.e., less sensitive to dropouts, since the effect of dropping one or several frequency bands is similar to comb-filtering of the signal that has only a minor impact on the quality of the reconstruction signal. This property is informally demonstrated, e.g., by simple demo application which can be found in [11]. This fact can prove very important in IP services.

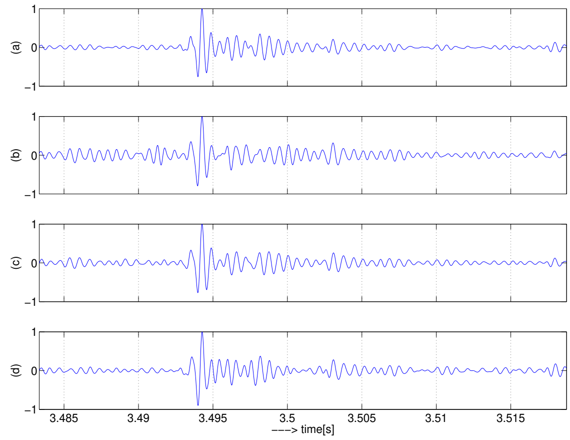

Fig. 3. Example of wide-band FDLP coding of castanets audio file. (a) Original segment; (b) coded using a FDLP technique; (c) coded using a FDLP-AbS technique; (d) coded using AMR-WB+ codec.

8. REFERENCES

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

[2] 3GPP TS 26.290, "Extended AMR Wideband codec", <http://www.3gpp.org/ftp/Specs/html-info/26290.htm>.

[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] 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.

[5] 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.

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

[7] Oppenheim A. V., Schafer R. W., "Discrete-Time Signal Processing", 2nd Ed., Prentice-Hall, NJ, USA, 1998.

[8] Churchill R. V., Brown J. W., "Introduction to Complex Variables Applications", 5th Ed., McGraw-Hill Book Company, NY, USA, 1982.

[9] 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.

[10] Schroeder M. R., Atal B. S., Hall J. L., "Optimizing digital speech coders by exploiting masking properties of the human ear", *in J. Acoust. Soc. Am.*, pp. 1647–1652, 1979.

[11] <http://www.icsi.berkeley.edu/~wooters/QualcommDemo>.