IDIAP RESEARCH REPORT

LOW-DELAY ERROR RESILIENT SPEECH CODING USING SUB-BAND HILBERT ENVELOPES

Sriram Ganapathy Petr Motlicek
Hynek Hermansky

Idiap-RR-75-2008

DECEMBER 2008

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

LOW-DELAY ERROR RESILIENT SPEECH CODING USING SUB-BAND HILBERT ENVELOPES

Sriram Ganapathy[1,2], *Petr Motlicek*[1], *Hynek Hermansky*[1,2]

[1]IDIAP Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{ganapathy,motlicek}@idiap.ch, hermansky@ieee.org

ABSTRACT

Frequency Domain Linear Prediction (FDLP) represents a technique for auto-regressive modelling of Hilbert envelopes of a signal. In this paper, we propose a speech coding technique that uses FDLP in Quadrature Mirror Filter (QMF) sub-bands of short segments of the speech signal (25 ms). Line Spectral Frequency parameters related to autoregressive models and the spectral components of the residual signals are transmitted. For simulating the effects of lossy transmission channels, bit-packets are dropped randomly. In the objective and subjective quality evaluations, the proposed FDLP speech codec is judged to be more resilient to bit-packet losses compared to the state-of-the-art Adaptive Multi-Rate Wide-Band (AMR-WB) codec at 12 kbps.

Index Terms— Frequency Domain Linear Prediction (FDLP), Speech Coding, Bit-packet loss, Perceptual Evaluation of Speech Quality (PESQ).

1. INTRODUCTION

Conventional approaches to speech coding achieve compression with a linear source-filter model of speech production using the linear prediction (LP) [1]. The residual of this modeling process, which represents the source signal, is transmitted to reconstruct the speech signal at the receiver. On the other hand, modern perceptual codecs [2] typically used for multi-media coding applications are not as efficient for speech content. Furthermore, the reconstruction quality of the state-of-the-art speech codecs are significantly degraded in lossy channel conditions.

In this paper, we propose to exploit the predictability of sub-band spectral components for encoding speech signals. Our approach is based on the assumption that speech signals in sub-bands can be represented as a product of a smoothed Hilbert envelope estimate and fine signal variations in the form of Hilbert carrier. The Hilbert envelopes are estimated using Frequency Domain Linear Prediction (FDLP) [3], which is an efficient technique for auto-regressive modelling of the temporal envelopes of the signal [4]. This idea was first applied for audio coding in the MPEG2-AAC (Advanced Audio Coding) [5], where it was primarily used for Temporal Noise Shaping (TNS).

In the proposed speech codec, the technique of linear prediction in spectral domain is performed on sub-band signals. We use a non-uniform Quadrature Mirror Filter (QMF) bank to derive 5 critically sampled frequency sub-bands. FDLP is applied over short segments (25 ms) of speech signal to estimate Hilbert envelopes in each sub-band. The remaining residual signal (Hilbert carrier) is further processed and its frequency components are quantized. The bit-packets for the FDLP codec contain individual sub-band signals in the form of FDLP envelope parameters and spectral components of the residual signal. At the decoder, the sub-band signal is reconstructed by modulating the inverse quantized residual with the AR model of the Hilbert envelope. This is followed by a QMF synthesis to obtain the speech signal back. The current version of the FDLP speech codec operates at a bit-rate of 12 kbps.

For speech codecs operating in lossy channels, some bit-packets get distorted and hence, the reconstructed signal is degraded. The intelligibility of speech is also affected in cases of severe bit-packet loss (frame dropouts). In order to simulate these channel conditions in speech codecs, bit-packets are dropped randomly at the decoder.

For the proposed FDLP codec, the dropout of bit-packets corresponds to loss of sub-band signals at the decoder. The degraded sub-band signals are recovered from the adjacent sub-bands in time-frequency plane which are unaffected by the channel. The objective and subjective listening tests show that the proposed FDLP codec is more resilient to frame dropouts compared to the state-of-the-art AMR-WB codec at similar bit-rate.

The rest of the paper is organized as follows. Sec. 2 explains the FDLP technique for the auto-regressive modelling of Hilbert envelopes. Sec. 3 describes the proposed low delay speech codec based on FDLP. The sub-band signal reconstruction technique in lossy channels is detailed in Sec. 4. The results of the objective and subjective evaluations are reported in Sec. 5.

2. FREQUENCY DOMAIN LINEAR PREDICTION

Typically, auto-regressive (AR) models have been used in speech applications for representing the envelope of the power spectrum of the signal by performing the operation of Time Domain Linear Prediction (TDLP) [6]. This paper utilizes AR models for obtaining smoothed, minimum phase, parametric models of temporal rather than spectral envelopes. The duality between the time and frequency domains means that AR modeling can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples [3].

For signals that are expected to consist of a fixed number of distinct transients, fitting an AR model constrains the temporal envelope to be a sequence of maxima, and the AR fitting procedure removes the finer-scale detail. This suppression of detail is particularly useful in speech coding applications, where the goal is to extract the general form of the signal by means of a parametric model and to

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.

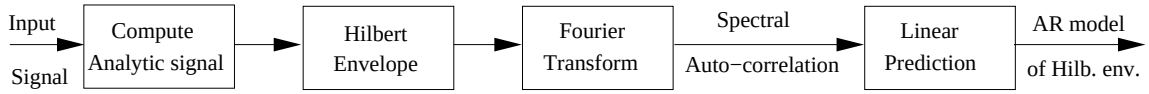

Fig. 1. *Steps involved in FDLP technique for AR modelling of Hilbert Envelopes.*

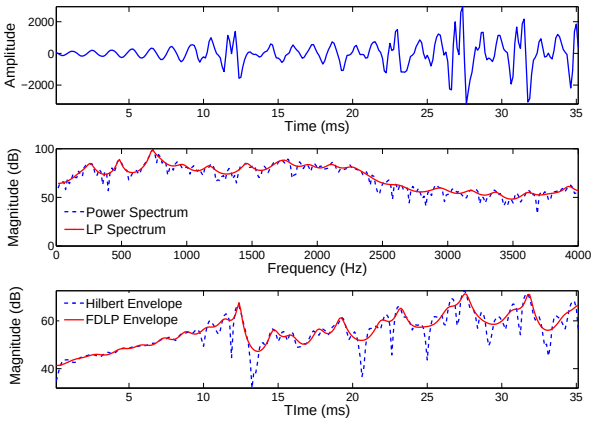

Fig. 2. *Linear Prediction in time and frequency domains for a portion of speech signal*

characterize the finer details with a small number of bits.

The block schematic showing the steps involved in deriving the AR model of Hilbert envelope is shown in Fig. 1. The first step is to compute the analytic signal for the input signal. For a discrete time signal, the analytic signal can be obtained using the Fourier Transform [8]. Hilbert envelope (squared magnitude of the analytic signal) and spectral autocorrelation function form Fourier transform pairs [5]. This relation is used to derive the autocorrelation function of the spectral components of a signal which are then used for deriving the FDLP models (in manner similar to the computation of the TDLP models from temporal autocorrelations [6]).

For the FDLP technique, the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal. This is in exact duality to the approximation of the power spectrum of the signal by the TDLP as shown in Fig. 2.

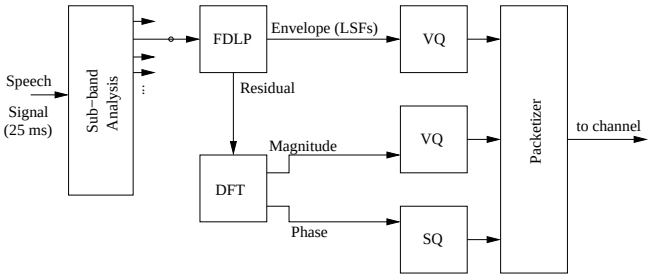

Fig. 3. *Scheme of the FDLP encoder.*

3. FDLP BASED SPEECH CODEC

Short temporal segments (25 ms) of the input speech signal are decomposed into 5 non-uniform QMF 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 residual signal (sub-band Hilbert carrier) is processed in spectral domain and the magnitude spectral parameters are quantized using VQ. Since a full-search VQ in this high dimensional space would be computationally infeasible, the split VQ approach is employed. Although this forms a suboptimal approach to VQ, it reduces computational complexity and memory requirements without severely affecting the VQ performance. Phase spectral components are scalar quantized (SQ) as they are found to have a uniform distribution. Graphical scheme of the FDLP encoder is given in Fig. 3.

In the decoder, shown in Fig. 4, quantized spectral components of the sub-band carrier signals are reconstructed and transformed into time-domain using Inverse Discrete Fourier Transform (IDFT). FDLP residuals in frequency sub-bands above 2 kHz are not transmitted, and they are substituted by white noise in the decoder. The reconstructed FDLP envelopes (from LSF parameters) are used to modulate the corresponding sub-band carriers. Finally, sub-band synthesis is applied to reconstruct the full-band signal. The final version of the FDLP codec operates at $\sim$ 12 kbps.

4. SIGNAL RECONSTRUCTION IN LOSSY CHANNELS

For the proposed FDLP codec, a bit-packet contains information about a single sub-band signal in the form of the Hilbert envelope parameters and the spectral components of the residual signal. Therefore, the loss of bit-packets in a lossy channel refers to the dropout of sub-band signals. Since short-term sub-band signals of

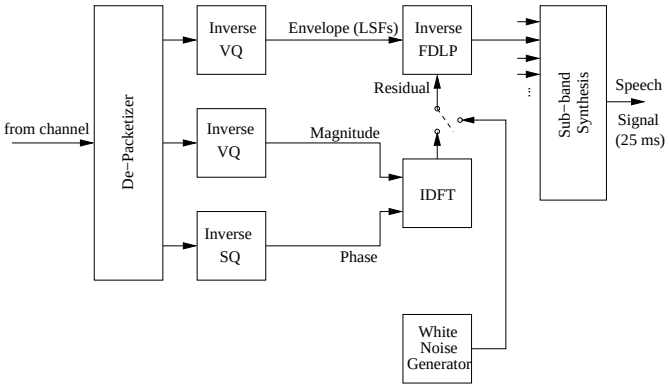

Fig. 4. *Scheme of the FDLP decoder.*

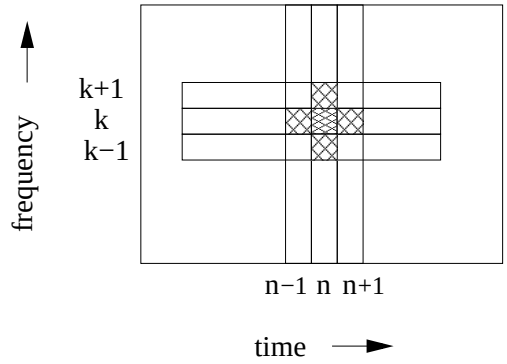

Fig. 5. *Reconstruction of corrupted sub-band signals using adjacent time-frequency sub-bands*

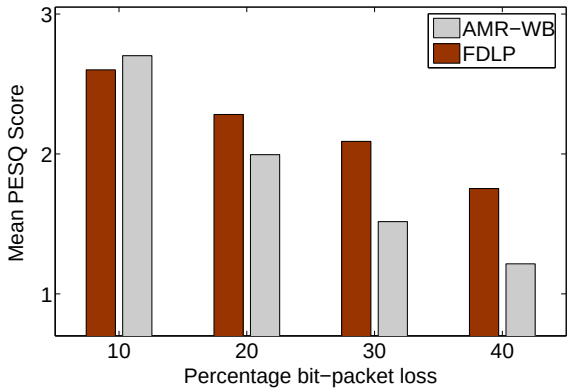

Fig. 6. *Mean objective quality test results provided by PESQ for* 9 *speech files for lossy channel conditions.*

speech are correlated with the neighboring sub-band signals (in the time-frequency plane), the sub-bands corresponding to degraded bit-packets can be reconstructed at the decoder using the adjacent sub-bands which are undistorted. Specifically, we estimate these sub-band signals as a time-frequency average of the contiguous sub-band signals which are unaffected by the channel.

Let $x_{n,k}(t)$ denotes short-term signal in the k^{th} sub-band for n^{th} time frame of a speech signal (Fig. 5). The neighboring sub-bands in the time-frequency plane are denoted as $x_{n,k-1}(t)$, $x_{n,k+1}(t)$, $x_{n-1,k}(t)$ and $x_{n+1,k}(t)$. If the bit-packet corresponding to $x_{n,k}(t)$ is distorted due to lossy channel, it is reconstructed by

$$x_{n,k}(t) = Av.\{x_{n,k-1}(t), x_{n,k+1}(t), x_{n-1,k}(t), x_{n+1,k}(t)\},$$

where $Av.$ denotes operation of averaging. If one of the adjacent sub-bands used in the averaging is also degraded, then it is not included in the mean computation. It is found that such an averaging operation retains the intelligibility of speech although, it introduces colored noise in the reconstructed signal.

5. EXPERIMENTS AND RESULTS

For the purpose of training the VQ codebooks, we use speech files from the subset of TIMIT database [9] which are sampled at 16 kHz. VQ codebooks for the magnitude spectral components of the sub-band residual signals and the FDLP envelope LSFs are trained using 400 speech files spoken by 20 male and 20 female speakers. For the objective and subjective quality evaluations, 9 challenging speech files from [14, 15] are used consisting of clean speech, radio speech and conference room recordings.

5.1. Objective Evaluations

Objective Evaluation is done with the Perceptual Evaluation of Speech Quality (ITU-T P.862 PESQ standard [10]). It is an enhanced perceptual quality measurement for voice quality in telecommunications. The quality estimated by PESQ corresponds to the average user perception of the speech sample under assessment PESQ – MOS (Mean Opinion Score). PESQ scores range from 1.0 (worst) up to 4.5 (best).

bit-rate [kbps]	12	12	12.2
codec	AAC+	FDLP	AMR-WB
PESQ score	3.1	3.3	3.7

Table 1. *Mean objective quality test results provided by PESQ for* 9 *speech files in ideal channel conditions.*

The first set of experiments compare the PESQ scores for ideal channel conditions (without any bit-packet loss). The output quality for the following codecs are compared:

1. The proposed FDLP codec at 12 kbps denoted as FDLP.
2. Enhanced AAC plus [2] at 12 kbps denoted as AAC+.
3. AMR-WB standard [11] at 12.2 kbps denoted as AMR-WB.

Table 1 shows the mean PESQ scores for the 9 speech utterances used for the evaluations. It can be seen that the proposed codec, without the use of standard bit-rate reduction techniques like Huffman coding and psycho-acoustic modules, provides better objective quality than the Enhanced AAC plus codec although the AMR-WB standard provides the best reconstruction speech quality for ideal channel conditions.

The lossy channel experiments are performed using the proposed FDLP codec and the AMR-WB codec. The recovery of the frame dropouts for AMR-WB codec is done as reported in [12]. Fig. 6 shows the mean PESQ score for the 9 speech files as function of the percentage of bit-packet loss.

5.2. Subjective Evaluations

Formal subjective listening tests are performed on ideal channel conditions (0 % bit-packet loss) as well as for the reconstruction with 30% bit-packet loss conditions. We use the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) methodology for subjective evaluation. It is defined by ITU-R recommendation BS.1534 documented in [16]. The mean MUSHRA scores for the subjective listening tests (with 95% confidence interval), using 9 speech files and 5 listeners is shown in Fig. 7. The results shown in this figure justify the objective quality evaluations from the PESQ scores (Table 1). Although the AMR codec performs well for ideal channel

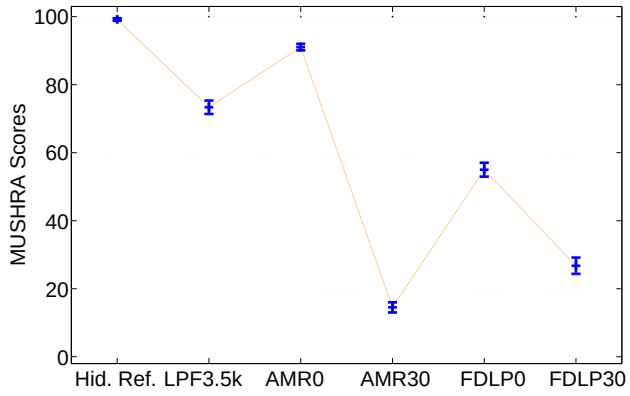

Fig. 7. *MUSHRA scores for* 9 *speech files using* 5 *listeners at* 12 *kbps (FDLP codec for ideal channel conditions (FDLP0), FDLP codec for* 30% *bit-packet loss conditions (FDLP30), AMR-WB codec for ideal channel conditions (AMR0), AMR-WB codec for* 30% *bit-packet loss conditions (AMR30), hidden reference (Hid. Ref.) and* 3.5 *kHz low-pass filtered anchor (LPF3.5k)).*

conditions, the proposed FDLP codec is more resilient to bit-packet losses.

6. CONCLUSIONS

A technique for low-delay error resilient speech coding is proposed which uses auto-regressive modelling of the sub-band Hilbert envelopes. Specifically, the technique of linear prediction in the spectral domain is applied on short segments ($25ms$) of speech signals in QMF sub-bands. The sub-band signals which are degraded in lossy channel conditions are reconstructed by the time-frequency averaging of the adjacent undistorted sub-bands. The objective and subjective evaluations, performed with the current version of the FDLP codec, suggest that the FDLP codec operating at $\sim$ 12 kbps provides more error resilience compared to state-of-art AMR-WB codec at similar bit-rates.

In order to be competent with the AMR-WB codec in ideal channel conditions, the proposed FDLP codec needs further improvement. Furthermore, the current version of the FDLP codec does not utilize the standard modules for compression efficiency provided by entropy coding and simultaneous masking. These form part of the future work.

7. REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *Proc. of the ICASSP*, Vol. 10, Apr. 1985, pp. 937-940.

[2] "Enhanced aacPlus General Audio Codec ",*3GPP TS 26.401* .

[3] M. Athineos, and D. Ellis, "Autoregressive Modeling of Temporal Envelopes", *IEEE Trans. on Signal Proc.*, Vol. 55, pp. 5237 - 5245, Nov. 2007.

[4] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", *Journal of Acoustical Society of America*, vol 105, no 3, pp. 1912-1924, Mar. 1999.

[5] J. Herre and J.D Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)", *Proc. of 101st AES Conv.*, Los Angeles, USA, pp. 1-24, 1996.

[6] J. Makhoul, "Linear Prediction: A Tutorial Review", *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[7] P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Frequency Domain Linear Prediction for QMF Sub-bands and Applications to Audio coding", Proc. of MLMI 2007, LNCS Series, Springer-Verlag, Berlin, 2007.

[8] L.S. Marple, "Computing the Discrete-Time Analytic Signal via FFT", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. 47, pp. 2600-2603, 1999.

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

[10] 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

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

[12] H. G. Hirsch and H. Finster, "The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems", *Proc. of Interspeech* , Sept. 2005, pp. 2697-3000.

[13] ITU-R BS.1284-1: "General methods for the subjective assessment of sound quality", 2003

[14] ISO/IEC JTC1/SC29/WG11, "Framework for Exploration of Speech and Audio Coding", *MPEG2007/N9254*, July 2007, Lausanne, CH.

[15] "Voice Age", *<http://www.voiceage.com/audiosamples.php>*

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