Speech Coding based on Spectral Dynamics

Petr Motlicek[12], Hynek Hermansky[123], Harinath Garudadri[4], and Naveen Srinivasamurthy[4]

[1] IDIAP Research Institute,
Rue du Simplon 4, CH-1920, Martigny, Switzerland
{motlicek,hynek}@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,naveens}@qualcomm.com

Abstract. In this paper we present first experimental results with a novel audio coding technique based on approximating Hilbert envelopes of relatively long segments of audio signal in critical-band-sized sub-bands by autoregressive model. We exploit the generalized autocorrelation linear predictive technique that allows for a better control of fitting the peaks and troughs of the envelope in the sub-band. Despite introducing longer algorithmic delay, improved coding efficiency is achieved. Since the described technique does not directly model short-term spectral envelopes of the signal, it is suitable not only for coding speech but also for coding of other audio signals.

1 Introduction

Intelligibility of the coded speech depends on proper estimation of parameters related to the short-term spectral envelope. Due to inertia of the air-mass in vocal tract cavities, the speech spectral envelope is relatively smooth and the speech signal is short-term predictable. This is used with advantage in techniques such as Linear Prediction (LP) and most of current speech coding techniques employ LP that approximates the envelope of the short-term power spectrum of speech by a spectrum of an all-pole (autoregressive) model. The LP-based speech coding techniques rely on the source-filter model of speech production and usually fail for any other kind of audio signals (several speakers, music, speech with some background, ...). See e.g., [1] for an excellent and comprehensive review.

Classical speech analysis techniques assume short-term signal stationarity. The input signal is divided into short-term frames ($10-30$ms), each containing relatively stationary signal and each being processed independently by techniques such as LP that yield vectors of short-term features. Speech dynamics is represented by a sequence of these vectors, each vector representing a particular configuration of the vocal tract.

However, vocal organs and their neural control mechanism have their own inertia too and subsequently the evolution of vocal tract shapes is also largely predictable. Thus, in terms of efficiency, it might be desirable to capitalize on this predictability and to encode longer temporal context (several hundreds of milliseconds) rather than processing every $(10 - 30\text{ms})$ temporal vectors independently. This is supported by recently reported efficiency of a new generation of modulation spectrum based audio coding techniques [2]. While such an approach obviously introduces longer algorithmic delays, the efficiency gained may justify its deployment in many evolving communications applications.

In this paper, we introduce a new audio coding technique that employs autoregressive modeling applied for approximating the instantaneous energy (Hilbert envelope) of critical-band sized sub-band signals. Unlike in [2], much longer temporal segments (around 1000 ms) are processed at a time. We also propose several initial attempts for proper reconstruction of the audio signal from such encoded Hilbert envelopes.

2 Encoding

2.1 Frequency Domain Linear Prediction

Hilbert envelope (squared magnitude of an analytic signal) can be parameterized by Frequency Domain Linear Prediction (FDLP) [3] that represents frequency-domain analogue of the well-known time-domain Linear Prediction (LP) [4]. Just as LP fits an all-pole model to the power spectrum of the input signal, FDLP fits an all-pole model to the squared Hilbert envelope of the signal.

To get an all-pole approximation of the 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 Hilbert envelope of the signal. The whole technique of deriving all-pole models of sub-band Hilbert envelopes is similar to the technique applied in [3].

2.2 Derivation of parameters of temporal envelopes in frequency sub-bands

In the proposed coding technique, the signal is divided into 1000ms long non-overlapping temporal segments which are transformed by DCT into the frequency domain, and later processed independently. FDLP technique is applied to every sub-segment of the DCT transformed signal that represent the frequency range of the sub-band. We get the approximations of Hilbert envelopes in sub-bands.

N_{BANDs} Gaussian functions (N_{BANDs} denotes number of frequency sub-bands) equally spaced on the Bark scale, with standard deviation $\sigma = 1\text{bark}$, are projected on the linear (Hertz) frequency scale and used as weighting windows on the DCT transformed signal. Therefore, the weighting windows are asymmetric and their width and spacing increases with frequency. The Bark scale from Perceptual Linear Prediction (PLP) analysis [5] is applied.

FDLP applied on each sub-segment from windowed DCT segment yields the approximation of the sub-band Hilbert envelope. The order of the all-pole model that in the case of FDLP controls temporal resolution of the technique depends on the length of the processed signal frame and is chosen experimentally. A graphical representation of the whole technique is given in Fig. 1.

2.3 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 modeling peaks and dips of the envelope, Spectral Transform Linear Prediction (STLP) technique [6] is used.

3 Decoding

FDLP approximates squared Hilbert envelope of the sub-banded temporal trajectory $x_k(t)$ (k determines the sub-band, t is time variable). Estimated Hilbert envelope $a_k(t)$ yields information about modulation of the signal in the particular sub-band. To reconstruct the signal, an additional component, the carrier $c_k(t)$, is required. This carrier is then modulated by the estimated envelope (see e.g., [7] for mathematical explanation).

$a_k(t)$ is approximated by FDLP and described by parameters of the resulting all-pole model. We have so far no explicit method to parameterize $c_k(t)$. However, some attempts for its efficient coding are discussed later in this paper.

3.1 Decoder

A general scheme of the decoder, also given in Fig. 1, is relatively simple and follows backwards the steps performed on the encoder. The decoding operation is also applied on each (1000ms long) input segment independently.

First, the signal $c_k(t)$ that represents Hilbert carrier is generated. The temporal envelope $a_k(t)$ is created from transmitted all-pole model coefficients. Temporal trajectory $x_k(t)$ is created so that $c_k(t)$ is modulated by $a_k(t)$. All these steps are performed for all frequency sub-bands. Finally:

1. Obtained temporal trajectories $x_k(t)$ for each frequency sub-band are projected to the frequency domain by DCT and added together.
2. A "de-weighting" window is applied to alleviate the effect of Gaussian windowing of DCT trajectory in the encoder.
3. Inverse DCT is performed to obtain 1000ms long output signal (segment).

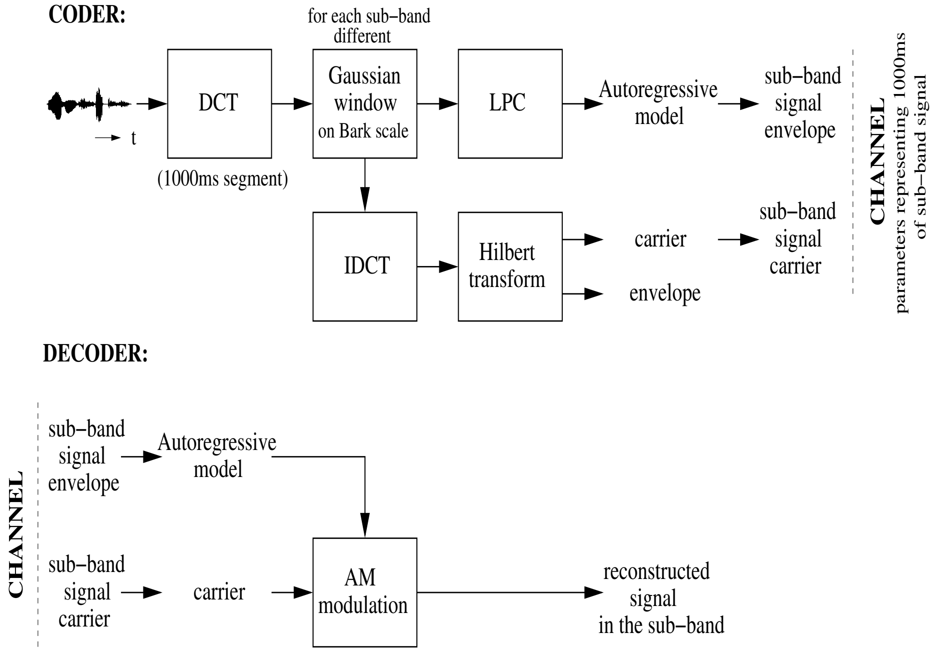

Fig. 1. Graphical scheme of the whole technique for one frequency sub-band.

4 Experiments

When the original temporal envelope $a_k(t)$ as well as Hilbert carrier $c_k(t)$ in all frequency sub-bands are fully preserved, the encoding scheme described in previous sections is lossless. This is analogous to classical residual-excited LP vocoder, where using the unmodified error signal for excitation of LP system yields the original signal.

All experiments were performed with audio signals sampled at $F_s = 8$kHz.

4.1 Representing temporal envelope in the individual frequency sub-bands

We chose $N_{BANDs} = 15$ which roughly corresponds to partition of one sub-band per 1bark. The FDLP estimated Hilbert envelope of each frequency sub-band is described by Line Spectral Frequencies (LSFs). The order of the all-pole models (the same for all sub-bands) was found by informal listening experiments. For coding the 1000ms long audio segments, the order of the model was set to $N_{LSFs} = 20$. We have used scalar quantization with $N_{BITs} = 4$ bits per LSF, which seems to be sufficient (quantization noise is not audible).

When using conventional autocorrelation all-pole method for deriving the FDLP all-pole models, the Hilbert envelope peaks seem overemphasized and the decoded signal sounds reverberant. This is especially true when low model order

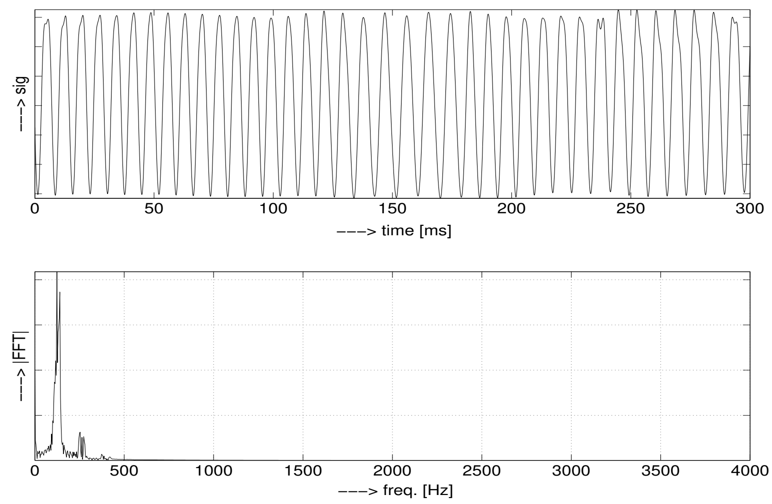

Fig. 2. Temporal trajectory and its spectral representation of $c_k(t)$ for $k = 2$. The frequency of the strongest spectral component is about 112Hz, which corresponds to the center bin of the 2nd sub-band.

is used. STLP can de-emphasize the peaks and thus significantly reduces this reverberation. In our experiments, we use STLP compression factor $r = 0.1$.

4.2 Decoding

In order to reconstruct an input audio signal in the decoder, we need to restore the carrier $c_k(t)$ in each frequency sub-band and to modulate this carrier by the envelope estimated using FDLP. In the first experiments we were dealing with rather effortless approaches producing output signal of *synthetic* quality on very low bit-rates.

1. Generating unvoiced speech: In the simplest approach, the carrier $c_k(t)$ can be substituted by a band-passed white noise. The applied band-pass represents frequency response of the Gaussian window that was applied in the analysis. Since we do not use (and therefore do not need to transmit) any information about the original $c_k(t)$, the bit-rate of the transmission R is given only by the rate necessary for the transmission of the quantized all-pole models. The resulting signal sounds obviously unvoiced (whispered) but is clearly intelligible.

2. Generating voiced speech: First, we have experimented with spectral components located at the frequencies that are integral multiples of some fixed fundamental frequency $F0$. However, as shown in Fig. 2, for voiced signal segments, especially in lower frequency bands, the carrier signal $c_k(t)$ appears to be well structured. Magnitude spectrum of Fourier Transform (FT) of $c_k(t)$ typically contains one dominant spectral component located close to the central frequency of the corresponding sub-band. Therefore, in order to introduce some voiced quality into the reconstructed signal, we tried to substitute $c_k(t)$ by a cosine with the corresponding frequency.

Another (computationally more expensive) approach is to estimate several additional strong spectral components using a "peak-picking" algorithm and to transmit just the corresponding frequencies to regenerate $c_k(t)$ in the decoder.

Subjectively, all these approaches introduce some voicing quality into the resulting audio signal, however, the reconstructed signal sounds quite unnatural and machine-like, similar to buzzy character of speech coded by a simple fixed-pulse excited vocoder without any voicing decision. However, the decoded speech appears clearly intelligible.

Finally, we have informally observed that a simple but possibly sufficient voice detector can be built on peaks and troughs of temporal envelope $x_k(t)$ in any of the low frequency sub-bands (up to $k = 5$). Since the voiced speech segments typically have dominant spectral energy concentration at lower frequencies, peaks of the envelope in lower sub-bands indicate voicing. Informal experiments with mixed excitation based on this criterion appear promising for improving the coded speech quality.

Carriers in higher frequency sub-bands (above $k = 5$) are less structured (more noise-like) and substituting these higher-frequency carriers by a band-pass noise as in the case of unvoiced excitation appears to be possible with only a minor effects on decoded signal quality.

3. Scalar quantization of carrier signal: In order to further improve the quality of the coded signal, we attempted some simple encoding of the original carrier. In this respect we have so far mainly explored scalar quantization of spectral components of $c_k(t)$, as described below. Since the character of the carrier $c_k(t)$ can change more quickly than once during the 1000ms used in estimating the sub-band Hilbert envelopes, we have been working with 200ms long segments of $c_k(t)$. We have observed that quantization of magnitudes by as little as 2 bits and phases by 3 bits seems sufficient for preserving reasonable signal quality. Further, it appears that only a few spectral components located around center frequency of corresponding frequency sub-bands are necessary for a carrier reconstruction. Yet additional thresholding can be applied to suppress very low magnitude spectral components.

5 Initial subjective impressions

The goal of this paper is to describe feasibility and basic principles of the proposed novel technique. However, even at this stage of its evolution, we have already first subjective impressions that may indicate its possible advantages.

5.1 Unvoiced carrier

First experiments were aimed at finding proper approximations of temporal envelopes $a_k(t)$ by FDLP and thus we used only random noise carriers. The sufficient intelligibility was achieved at bit-rates $R = 1.2$kbps (with parameters $N_{BANDs} = 15$, $N_{LSFs} = 20$, $N_{BITs} = 4$). The algorithm provides subjectively much more natural signal than LPC10 standard with only noise excitation at the bit-rate around 2.1kbps. Although the reconstructed signal sounds whispered, it is clearly intelligible. Therefore, we applied this parameterization of temporal envelopes in all subsequent experiments focused on parameterization of the carrier $c_k(t)$.

In addition, other informal experiments indicate that preserving as few as 5 important frequency sub-bands seem not to degrade an intelligibility of reconstructed speech. This means that bit-rates around 400bps are achievable.

5.2 Voiced carrier

Fixed-frequency carrier: The same bit rate of $R = 1.2$kbps can be obtained when $c_k(t)$ is substituted by cosine signal with frequency equal to center bin of corresponding sub-band. This substitution is suitable for fully voiced audio segments. Then magnitude spectra of $c_k(t)$ especially for low frequency bands contain one strong spectral component. The reconstructed signal is well audible but contains strong tonal artifacts.
Estimating frequencies of the carrier: Simple "peak-picking" algorithm performed on top of spectral magnitudes can reduce these artifacts. Subjectively the best results were achieved with scalar quantization of spectral components of $c_k(t)$. Then, especially periodic audio signals (e.g., music) can be encoded into few kbps (around 5kbps), preserving good quality.
Mixed carrier: Reconstructed signal can be noticeably improved when combining the two source-models together.

6 Discussion and Conclusions

This paper describes first experiments with novel audio codec based on spectral dynamics. Although the algorithm introduces rather large algorithmic delays, we believe the technique can find many possible applications.

Here, we have described only preliminary experiments focused mainly on the approximation of temporal envelopes and proposed simple methods how to encode the carrier.

Among possible advantages of this coding technique compared to classical state-of-the art methods based on short-term frames belongs:

- Exploiting predictability of temporal evolution of spectral envelopes of speech spectra allows for efficient transmission and/or storage of intelligible speech.
- The technique is based on independent processing of individual frequency sub-bands. It is therefore inherently suitable for exploiting non-equal frequency resolution of human hearing.

- The technique is not directly based on source-filter model of speech production, thus it is also potentially suitable for coding of non-speech sounds.
- The well structured character of the sub-band carrier signals (discussed in Section 4.2) suggests a potential for its efficient coding, thus allowing for encoding of high-quality audio. Some simple carrier coding schemes have been discussed in this paper, other are a topic of our current interest.
- Though not extensively discussed in this paper, it is straightforward to control the algorithmic delay, the quality of reconstructed sound, the resiliency to drop-outs, and the final bit-rate, making the codec suitable for variable bandwidth channels.
- The reconstruction is based on linear addition of contributions from different frequency sub-bands. Possible loss of data (e.g., due to drop-outs in the transmission) may merely mean loss of data from some sub-bands and some change in the signal quality but does not significantly affect signal intelligibility.

7 Acknowledgments

This work was partially supported by grant from ICSI Berkeley, USA. It was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management (IM)2" as well as DARPA through the "EARS (Effective, Affordable, Reusable Speech-to-Text)" project.

References

1. A. S. Spanias. "Speech Coding: A Tutorial Review", *in Proc. of IEEE*, Vol. 82, No. 10, October 1994.
2. M. S. Vinton, L. E. Atlas. "A scalable and progressive audio codec", *in Proc. of ICASSP*, Vol. 5, pp. 3277-3280, Salt Lake City, USA, May 2001.
3. M. Athineos, H. Hermansky, D. P. W. Ellis. "LP-TRAP: Linear predictive temporal patterns", *in Proc. of ICSLP*, pp. 1154-1157, Jeju, S. Korea, October 2004.
4. J. Makhoul. "Linear Prediction: A Tutorial Review", *in Proc. of IEEE*, Vol. 63, No. 4, April 1975.
5. H. Hermansky. "Perceptual linear predictive (PLP) analysis for speech", *J. Acoust. Soc. Am.*, pp. 1738-1752, 1990.
6. H. Hermansky, H. Fujisaki, Y. Sato. "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. S. Schimmel, L. Atlas. "Coherent Envelope Detector for Modulation Filtering of Speech", *in Proc. of ICASSP*, Vol. 1, pp. 221-224, Philadelphia, USA, May 2005.