Mixed-Domain Coding and Interpolation of Voiced Speech

Juan Carlos De Martin

Dipartimento di Elettronica
Politecnico di Torino
I-10129 Torino, Italy
demartin@polito.it

Allen Gersho

Dept. of Electrical & Computer Engineering
University of California
Santa Barbara, CA 93106 USA
gersho@ece.ucsb.edu

Abstract

A mixed-domain coding technique is introduced for speech coding below 4 kb/s. In each frame, a pitch cycle is extracted from the LP residual and differentially time-domain coded. The decoder performs frequency-domain interpolation of the residual from the recovered pitch cycles.

1. Introduction

Ever since CELP-type schemes have begun to reveal their limitations in coding voiced speech at low bit rates, much interest has been focused on finding better alternatives. Recently, several important contributions (e.g., [1], [2], [3], [4]) exploit the redundancy of voiced speech by extracting pitch cycles and interpolating between them. Two major issues arise in this approach: (a) how to code the selected pitch cycles, and (b) how to interpolate them.

Coding is generally performed in the frequency-domain, on the assumption that accurate reproduction of only the short-time spectral magnitude is important to human perception. While this is largely correct if we limit ourselves to intelligibility, for synthesizing good natural-sounding speech it is necessary to carefully recreate a suitable short-time spectral phase that maintains a "natural" and smooth evolution with time. Unfortunately, coding the phase directly requires too many additional bits, [7], while finding an effective model for the phase has proven to be elusive. Interpolation has sometimes been performed in the time domain, reportedly with good quality; however, the methods, being fairly heuristic, are not robust and are hard to reproduce. In contrast, very good quality has been achieved with frequency-domain interpolation schemes, but at the cost of very high complexity.

We propose a novel, mixed-domain approach in which time-domain coding of pitch cycles is combined with their interpolation in the frequency-domain. Coding in the time-domain bypasses the phase coding problem and ensures a faithful representation of perceptually significant features; it is also efficient since the quantization can exploit the periodicity of pitch cycles. The frequency-domain interpolation is adapted from [5], [6]. This technique is robust, consistent and relatively simple.

This work was supported in part by the University of California MICRO program, DSP Group, Inc. Speech Technology Laboratories, Echo Speech Corporation, Moseley Associates, Rockwell International Corporation, Texas Instruments, Inc., and Qualcomm, Inc.

2. Coder structure

In our proposed coder the original speech is first passed through an LP inverse filter whose coefficients are computed every 20 ms, the size of the working frame. The output of the filter is subjected to an open-loop pitch estimation procedure by an autocorrelation-based algorithm, and the resulting trajectory is then smoothed. Then, a pitch cycle is extracted from the beginning of the current frame. This cycle is quantized and then transmitted to the decoder. Additional parameters transmitted include the pitch-period, a voiced/unvoiced flag, and the LSFs. The voiced/unvoiced decision is made by a method similar to that employed in the LPC-10e standard.

The decoder rebuilds the quantized pitch cycle waveform, and sends it (together with the previous pitch cycle) to an interpolation module, described later. The excitation frame obtained from the interpolator is then passed through the LP synthesis filter to yield the synthesized speech.

3. Quantization

Each new quantized pitch cycle waveform is made up of three contibutions: *(a)* the previous (quantized) pitch cycle, suitably time scaled, *(b)* a selection from a single-impulse codebook, and *(c)* one from a noise codebook. Each of these also has a corresponding gain factor.

In generating *(a)*, we expand or compress the previous pitch cycle to the current value of the pitch period and shift and scale it optimally. The inclusion of *(b)* ensures that the first pitch cycle at the beginning of every voiced segment of

speech can be generated. Even when a previous pitch cycle is available, *(b)* can be helpful in representing the current pitch cycle. Finally, *(c)* is well-suited to model the residual difference between current and previous pitch cycles.

The selection of the above three components and corresponding gains is done sequentially in a closed-loop fashion with a perceptual weighting of the error. An approximately correct value for the state of the synthesis filter is achieved by forming a backward extended periodic excitation segment and passing it through the synthesis filter. This state computation facilitates the closed-loop search.

A rate of 2,350 bits/s for the excitation can be obtained with the following bit allocation: 5 bits for each gain, 7 bits each for *(a)* and *(b)*, 10 bits for *(c)*, 7 bits for the pitch period, and 1 bit for the voiced/unvoiced flag. With a split-VQ 24 bits/frame quantization of the LSFs, the overall bit rate for the voiced speech becomes 3,550 bits/s.

4. Interpolation

The new excitation frame is synthesized in the interpolation module from the pitch-sized DFTs of the previous and current pitch cycle waveforms. The spectral magnitudes are interpolated linearly in time; for the evolution of the phases in time we employ cubic interpolation [6] [5]. Since the pitch period is variable, due attention is given to the "death" and "birth" of harmonics. The cubic phase interpolation method guarantees smooth transitions at the boundaries, and good tracking of the instantaneous frequency of the waveform.

This method entirely avoids several problems affecting time-domain approaches, such as correct alignment of the extracted pitch cycles. It is also less complex than some other frequency-domain techniques which require upsampling and alignment.

5. Results and conclusions

Coding the pitch cycle waveforms in the time-domain allows representing the perceptually significant features of voiced excitation in an efficient manner. The interpolation algorithm ensures smooth transitions and good instantaneous frequency tracking. Preliminary results indicate that the mixed-domain approach has promise of achieving good quality at a bit-rate below 4 kbit/s, and lends itself to several further improvements that are currently under study.

6. References

[1] W. Bastiaan Kleijn and Wolfgang Granzow, "Methods for Waveform Interpolation in Speech Coding," *Digital Signal Processing*, pp. 215-230, 1991.

[2] Yair Shoham, "High-quality speech coding at 2.4 to 4.0 Kbps based on time-frequency interpolation," *Proc. IEEE ICASSP*, pp. II/167-170, 1993.

[3] Peter Lupini and Vladimir Cuperman, "Spectral Excitation Coding of Speech," *Proc. SBT/IEEE International Telecommunications Symposium*, Brazil, August 1994.

[4] G. Yang and H. Leich and R. Boite, "Voiced Speech Coding at Very Low Bit Rates Based on Forward-Backward Waveform Prediction", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 40–47, January 1995.

[5] Luis B. Almeida and Fernando M. Silva, "Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme," *Proc. IEEE ICASSP*, pp. 27.5.1-27.5.4, 1984.

[6] Robert J. McAulay and Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 4, August 1986.

[7] William A. Pearlman and Robert M. Gray, "Source Coding of the Discrete Fourier Transform," *IEEE Trans. on Information Theory*, vol. IT-24, no. 6, pp. 683–692, November 1978.

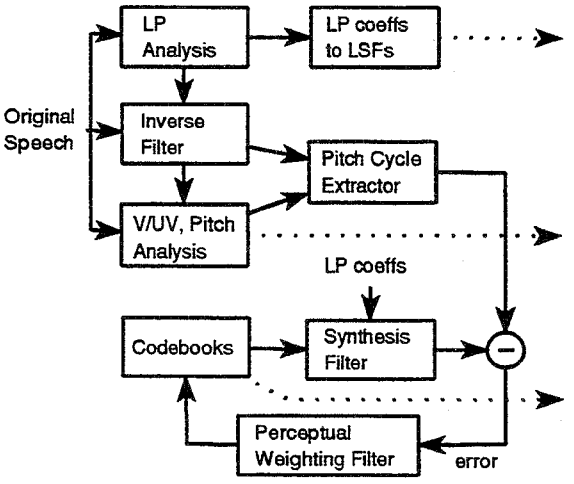

Fig. 1 Encoder Block Diagram