MIXED-DOMAIN CODING OF SPEECH AT 3 KB/S

Juan Carlos De Martin

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

Allen Gersho

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

ABSTRACT

We present a speech coding algorithm called mixed-domain residual coding (MDRC) wherein a prototype pitch cycle in each frame of the speech residual is coded in the time-domain while interpolation of the residual signal is performed in the frequency-domain. A novel quantization scheme takes into account time scaling and differentially codes successive prototypes with a closed-loop perceptually-weighted search. A fixed-rate (3.15 kb/s) implementation of MDRC achieves quality better or comparable to higher rate coders such as FS 1016 CELP and IMBE.

1. INTRODUCTION

Severe limitations of CELP-based speech coding schemes at bit-rates below 4 kb/s have led to the study of a new and promising approach to speech coding based on *waveform interpolation*. Important contributions of this type are reported in [1], [2], [3], and [4]. These coders exploit the redundancy of voiced speech by extracting selected pitch cycles as prototype waveforms from the prediction residual and interpolating between them to reconstruct the missing cycles. Each prototype pitch cycle is assumed to fully represent the local character of the residual signal with features that are expected to evolve smoothly in successive cycles.

Interpolation between prototypes has been performed in the time-domain with reportedly good quality, but with methods which are fairly heuristic and difficult to reproduce. Frequency-domain interpolation schemes are reported to perform very well, but they are often characterized by high complexity.

Coding of the cycles is generally done in the frequency domain, where often only the spectral magnitudes are retained; but spectral magnitudes alone do not yield natural sounding, high quality speech and spectral phases are expensive to code [7] and difficult to model well. While established and mature coding schemes have achieved good results by modeling the phase, we believe that effective ways of quantizing the cycles in the time-domain could yield even better results, entirely bypassing the problem of phase modeling and coding.

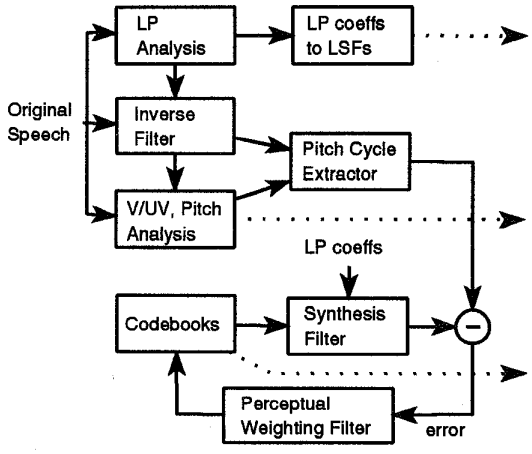


Fig. 1 Encoder Block Diagram

We propose a speech coding algorithm, called *mixed-domain residual coding* (MDRC), first reported in [10], which delivers high quality at a low bit rate by adequately modeling the evolution of pitch cycles in the frequency-domain while effectively quantizing the prototype cycles in the time-domain. The interpolation technique is an adaptation of the scheme described in [6], where it is employed in the context of sinusoidal transform coding (STC). This technique simplifies the extraction of the cycles and does not require any computationally-expensive alignment. The quanti-

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.

0-7803-3192-3/96 $5.00©1996 IEEE

zation is a multistage, variable-dimension scheme, that exploits the quasi-periodicity of successive cycles and preserves the perceptually important features of the waveform. The quality of speech coded with the mixed-domain coder is comparable to or better than the higher rate coders FS 1016 CELP and IMBE.

2. THE MDRC MODEL

In MDRC, each 20 ms frame is processed by the encoder to yield a voicing decision, an open-loop pitch period estimation and an all-pole LP model. A residual frame is obtained by inverse-filtering the input frame. If the frame is voiced, a pitch cycle is extracted at the beginning of the residual frame and coded. The decoder synthesizes the output frame by interpolating between two successive pitch cycles and then passing the result through the LP-synthesis filter defined by the decoded parameters. Unvoiced frames are coded by a conventional CELP scheme.

Pitch-sized DFTs of the current and previous cycles are fed to the interpolation module. The evolution from one cycle to the other is then modeled by a suitable interpolation of the magnitudes of the harmonics and a cubic spline interpolation of their instantaneous phase ([5], [6]).

Since the pitch period is variable, due attention is given to the "death" and "birth" of harmonics. The cubic interpolation of the phases allows for good tracking of changes in the instantaneous frequency of the harmonics. Moreover, the overall method assures continuity at the boundaries and does not require cycles alignment.

Interpolation schemes working on the speech waveform itself, rather than the residual signal, have to track two disjoint evolutions at the same time (excitation and vocal tract), and thus requiring a high frame update (or cycle extraction) rate. Operating on the residual, however, allows a lower frame update rate, since now the interpolation is tracking the evolution of the excitation only, while the evolution of the vocal tract is tracked by interpolating the LP coefficients. Very good quality is then possible with update rates in the range of 20–25 ms, as opposed to 10–12 ms.

This model has been tested by replacing the original voiced segments of a set of speech files with the voiced segments synthesized according to the mixed-domain model. We found that the resulting speech is perceptually almost indistinguishable from the original. Thus, the only source of quality degradation is in the quantization of the prototype pitch cycles. We next describe our quantization technique.

3. QUANTIZATION OF PITCH CYCLES

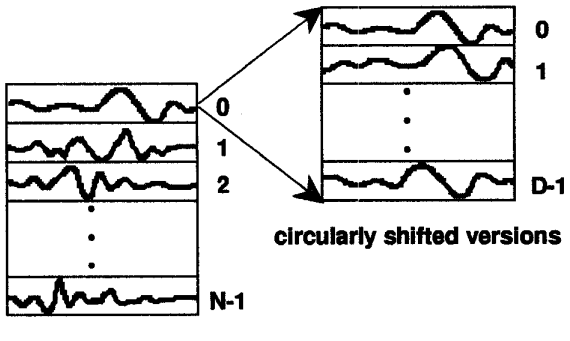
A 3-stage quantization scheme is employed. The first stage selects a cyclically shifted version of the previously quantized and time-scaled cycle. The second stage approximates the first-stage quantization residual with a suitably-placed single impulse, and the third stage quantizes the 2nd stage residual with a trained codebook.

The selection of the above three components and corresponding gains is done sequentially in a closed-loop fashion with perceptual weighting. The state of the synthesis filter, which depends on the interpolation yet to be performed, is approximated by extrapolating backwards from a given prototype cycle at the beginning of the current frame. Due to the variable dimensionality of the vectors, time-scaling is employed.

The first-stage shift can be either determined by means of a closed-loop search and then transmitted, or computed at the decoder based on the knowledge of the pitch contour and the position of the previous cycle. The latter scheme saves 7 bits/frame, at the cost of increased decoder complexity and a slightly suboptimal performance (mostly during transitions).

3.1. First-cycle quantization

In [1], where a 3-stage quantization scheme was used as well, the variable-dimensionality of the cycles was not addressed and the issue of how to code the *first* cycle of each segment was not discussed. Since we are differentially representing successive cycles, suitable representation of the starting cycle in a voiced segment is important to effectively handle a voicing onset.



The first cycle of each voiced segment is coded in two stages. The first stage is a shape-gain VQ stage (7 bits for the shape, 5 bits for the gain), specifically designed to model prototype cycles. The search is conducted on the entries of the codebook and all their circularly shifted versions.

The index for the best codevector as well as the optimal shift are sent to the receiver. The second stage then approximates the resulting residual, searching a trained codebook.

3.2. Fast search

The search for the best codevector *and* for the optimal shift entails a huge increase in complexity if done straightforwardly. However, dramatic savings can be achieved by exploiting the redundancy present in the convolution of shifted versions of the same vector with the same (fixed) impulse response, making the scheme entirely feasible.

When considering the cost of computing the convolution, the energy of the convolved vector and the scalar product of the convolved vector with the target vector, the computational balance between the direct versus simplified approaches, in floating-point operations, for a codebook of size N plus all cyclical shifts of its codevectors is as follows:

- Direct approach: $N \times (k^3 + 5k^2)$;
- Fast approach: $N \times (5k^2 + 4k)$,

where N is the size of the codebook, and k its dimension. For the sake of comparison, the cost of a direct search of a kN-size codebook is: $kN \times (k^2 + 5k)$.

3.3. Bit allocation

We call *mode I* (independent) the quantization scheme used for first-cycles, while *mode P* (predictive) denotes the scheme used in all other (voiced) cases. In the case of closed-loop determination of the first-stage shift, the global bit allocation is delined in the table below.

Voiced Frames			
Mode I		**Mode P**	
LSFs	24	LSFs	24
Pitch	7	Pitch	7
V/UV	1	V/UV	1
Cycle CB	7	Prev. Cycle	7
Shift	7	Impulse	7
Trained CB	9	Trained CB	9
Gains	10	Gains	15
Total	**65**	**Total**	**70**

In our current configuration of the MDRC coder, with open-loop computation of the first-stage shift, we specify the excitation with 1,550 bits/s (1,650 bit/s for first-cycle frames.) With a split-VQ (5-5) 24 bits/frame quantization of the LSFs, 7 bits for the pitch period, and 1 bit for the voiced/unvoiced flag, the overall bit rate for voiced speech becomes 3,150 bits/s (3,250 bit/s for first-cycle frames.)

4. CODEBOOK TRAINING

A modification of the Generalized Lloyd Algorithm, suggested by the characteristics of the signal, is employed to train the first-cycle codebook.

There are two main issues: the variable dimensionality of the cycles in the training set, and the nature of the cycles themselves. To address the former, the target codebook is fixed-dimension, and all the cycles are expanded or compressed to that dimension according to the instantaneous pitch period. The dimension of the codebook has to lie between the minimum and the maximum pitch period values, and will be determined by the choice of the expansion/compression technique and the desired bit-rate. If expansion is more easily, or more efficiently, performed than compression, a higher value (e.g. 128) will be a good choice, while a smaller value (e.g., 64) will be of value if saving a few bits per frame is important.

The nature of the cycles comes into play when forming the partitions. We are training a *cycle*–codebook, and the training must select representative cycles, not just waveforms that happen to minimize a distortion measure. Therefore, each cycles in the training set is first *aligned* to the candidate codevector and *then* the distance is computed. As a consequence, clusters of aligned cycles are formed and the resulting centroids tend to "look like" cycles, since they are formed by summation of aligned waveforms.

In an attempt to have even more "natural" cycles, the cycle closest to what would be the classic centroid can be selected as new centroid. However, we have found that with this method larger codebooks become necessary, otherwise the personality of the speaker is affected. On the other hand, with a sufficiently large codebook, this scheme could provide very high quality.

5. PITCH AND VOICING ESTIMATION

The residual signal is subjected to an open-loop pitch estimation procedure by an autocorrelation-based algorithm, and the resulting trajectory is then smoothed to produce a pitch candidate for the current frame. A tentative voiced/unvoiced decision is made by a method similar to that employed in the LPC-10e standard. The final decision about pitch and voicing is then made by a rule-based algorithm heuristically designed to avoid, among the other things, pitch halving and doubling,

abnormal pitch jumps, and unnatural voicing patterns. Given the nature of the interpolation scheme, transition frames are better dealt with if labeled as voiced.

6. CODING OF UNVOICED FRAMES

Unvoiced frames can be coded with a variety of techniques. In the present configuration of our coder, each 20-ms frame is divided into four 5-ms subframes which are coded by a traditional CELP scheme, using 6-bit for the shape and 3 bits for the corresponding gain.

Unvoiced Frames	
LSFs	24
V/UV	1
Shapes	24
Gains	12
Total	**61**

At this moment, we are using the same quantization scheme for LP coefficients for all frames, voiced and unvoiced. As pointed out, for example, in [11], 24 bits are too many for quantizing unvoiced spectra. In a future variable rate version of the mixed-domain coder, substantial bit-rate savings could be achieved by allocating about 10 bits/frame for spectral quantization of unvoiced frames, instead of the current 24 bits/frame.

7. PERFORMANCE

For subjective quality evaluation, informal A-B comparison tests were performed using the 4.15 kb/s Inmarsat-M IMBE coder ([8]) and the 4.8 kb/s FS 1016 CELP coder ([9]), as reference coders. The speech coded with the mixed-domain coder outperformed or matched the speech coded with the other techniques.

8. CONCLUSIONS

We have presented a mixed-domain speech coding algorithm based on the interpolation of cycles of the voiced residual. We have described a suitable interpolation technique and an effective quantization scheme. An implementation of the algorithm delivers high speech quality, comparable to higher rate FS 1016 CELP and IMBE coders. Variations of the quantization and interpolation scheme and other improvements to the mixed-domain coder are being explored to further enhance quality and robustness.

9. 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, 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.

[8] Digital Voice Systems, "Inmarsat-M Voice Codec, Version 2," *Inmarsat-M specs*, Inmarsat, February 1991.

[9] J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, "The DoD 4.8 kbps Standard (Proposed Federal Standard 1016)," in B. S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, Kluwer Academic Publ., 1991, pp. 121-133.

[10] Juan Carlos De Martin and Allen Gersho, "Mixed-Domain Coding and Interpolation of Voiced Speech," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, Maryland, September 1995, pp. 25–26.

[11] Roar Hagen, Erdal Paksoy and Allen Gersho, "Variable Rate Spectral Quantization for Phonetically Classified CELP Coding," *Proc. IEEE ICASSP*, pp. 748–751, 1995.