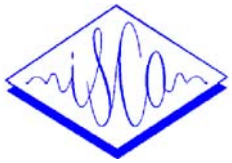

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LOW DELAY ANALYSIS/SYNTHESIS SCHEMES FOR JOINT SPEECH ENHANCEMENT AND LOW BIT RATE SPEECH CODING

Rainer Martin[*], *Hong-Goo Kang, and Richard V. Cox*

AT&T Labs-Research, Speech and Image Processing Services Research Lab
180 Park Avenue, Florham Park, N.J. 07932
martin@ind.rwth-aachen.de, goo|rvc@research.att.com

10.21437/Eurospeech.1999-332

ABSTRACT

In this contribution we discuss methods to reduce the algorithmic delay of joint speech enhancement and low bit rate speech coding algorithms. We introduce a novel overlap/add scheme which can reduce the overall delay of the joint system for some speech coders considerably. The new scheme takes advantage of the fact that some low bit rate coders store a significant amount of look-ahead samples in their input buffer. The look-ahead samples often have less influence on the parameter estimation than the samples in the current frame. Therefore, these samples might be used by the coder without being fully overlapped or reconstructed and with only little effect on the quality of the coded speech. The concept has been successfully used for joint enhancement and coding using the 2.4 kbps versions of the MELP and the WI speech coders.

1. INTRODUCTION

Despite substantial progress in recent years, low bit rate speech coding in the presence of acoustic background noise is still a challenge. Low bit rate speech coders rely mostly on spectral parameters and allocate only very few bits for the representation of the residual signal or phase information. It is therefore not surprising that these coders suffer severe quality and intelligibility losses when a significant level of background noise is present.

The quality of the encoded speech can be significantly improved if the coder is combined with a noise reduction preprocessor [1, 2]. However, besides adding to the computational complexity of the joint system, the noise reduction preprocessor also introduces additional algorithmic delay.

Since low bit rate speech coders already have a relatively large algorithmic delay any additional delay must be kept at a minimum. One way to reduce the additional delay is to shorten the frame length of the enhancement preprocessor and/or use zero padding. This, however, reduces the spectral resolution of the enhancement process. In this paper we propose an overlap/add scheme which is suited to reduce the algorithmic delay of the joint enhancement and coding system to the delay of the coder plus a very small additional delay (< 5 ms). At the same time this scheme allows sufficient frequency resolution in the enhancement process, which leads to superior speech quality compared to enhancement systems with shorter FFT lengths.

[*] This work was carried out while on leave from IND, Aachen University of Technology, D-52056 Aachen, Germany.

2. JOINT SYSTEM APPROACHES

A speech enhancement preprocessor commonly consists of three major components: a spectral analysis/synthesis system, a noise estimation algorithm, and a spectral gain computation algorithm. The analysis/synthesis system is usually realized by means of a windowed FFT/IFFT and an overlap/add algorithm. The length of the analysis window, its type, and the shift of the window from one analysis frame to the next, all have a major impact on the algorithmic delay and the performance of the preprocessor. Typical windows include the Hanning and the Tukey window with a length of 128 or 256 samples.

In this paper we will consider two methods for the overlap/add operation of the preprocessor (see Fig. 1). In method A the preprocessor is completely independent from the speech coder. The output from the enhancement algorithm is the input to the speech coder. In this case the frame size of the enhancement preprocessor can be chosen completely independent from the frame size of the speech coder. Although this approach should not be used for real time processing, it will serve as a reference system in this paper. In method B the final overlap/add operation of the preprocessor is moved into the input buffer of the speech coder which allows a more flexible use of the enhanced data. More specifically, by integrating the overlap/add operation into the speech coder, the overlap/add operation can be modified to reduce the algorithmic delay of the joint system significantly.

3. LOW BIT RATE SPEECH CODERS

To facilitate the discussion of the new overlap/add scheme we briefly outline how the MELP and the WI coder use the samples in their input buffers.

3.1. The MELP Coder

The MELP coder is a mixed excitation vocoder with remarkable speech quality at bit rates around 2 kbps [3]. For each input signal frame of 180 samples the 2.4 kbps MELP coder extracts 10 linear prediction coefficients, 2 gain factors, 1 pitch value, 5 bandpass voicing strength values, 10 Fourier magnitudes, and an aperiodic flag. These parameters are extracted from the input data buffer of the coder as shown in Fig. 2. The input buffer holds the data of the current frame as well as some past samples and one look-ahead frame. We notice that the latest 60 samples of the input buffer are not used for LPC analysis and the computation of the first gain factor. It can be expected that enhancement errors within these samples have a low impact on the overall performance of

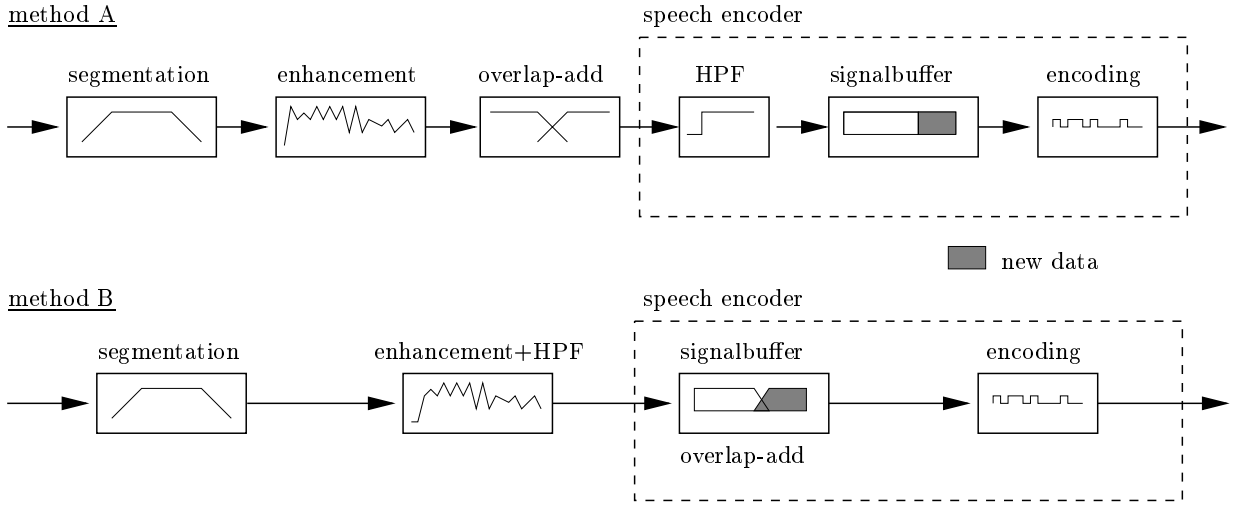

Figure 1: Joint speech enhancement and speech coding. Method A: enhancement preprocessor is independent of coder. Method B: overlap/add of enhancement preprocessor is integrated into speech coder.

the MELP coder. We can exploit this to reduce the delay of the joint enhancement preprocessor and coding system (see Sec. 5).

The MELP coder also uses an IIR highpass filter to remove low frequency noise. In the low delay joint enhancement and coding system (method B) we move this filter into the frequency domain and combine it with the enhancement algorithm. By removing the recursive operations in between the enhancement and the coding the implementation of the new scheme is simplified.

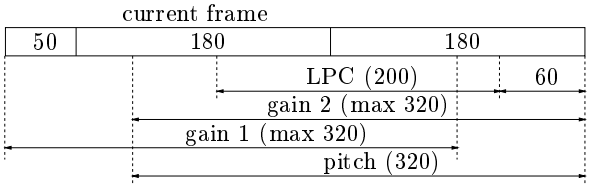

Figure 2: Utilization of data in the input buffer of the MELP coder. Numbers indicate frame sizes.

3.2. Waveform Interpolation Coder

In Waveform Interpolation (WI) coding [4], the excitation signal is represented by an evolving waveform. The evolving waveform is efficiently described by a decomposition into two components by filtering along the time axis. High-pass filtering results in the rapidly evolving waveform (REW) representing the noise-like/unvoiced component of speech. Low-pass filtering results in a slowly evolving waveform (SEW) representing the quasi-periodic voiced component of speech. A low accuracy description of the REW magnitude spectrum at a relatively high update rate is sufficient for good performance. The SEW magnitude spectrum requires an accurate description but a relatively slow update rate. It is well known that the 2.4 kbps WI [5] has better quality than the 4.8 kbps federal standard (FS1016).

Besides the current frame of 200 samples, the 2.4 kbps WI coder keeps one past and one future frame plus 110 look-ahead samples in its input buffer. From this buffer the 2.4 kbps WI coder extracts its parameters as shown in Fig. 3. These look-ahead samples are used for LPC

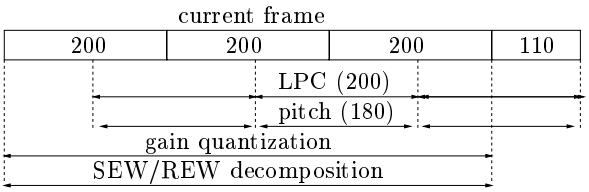

Figure 3: Utilization of data in the input buffer of the WI coder. Numbers indicate frame sizes.

analysis and pitch estimation. However, the residual signals from this part are eventually not used for modelling the excitation information (SEW/REW) and gain parameters. Besides, the pitch value for this part is only used as a temporary value (final pitch for SEW/REW decomposition is estimated in the previous frame).

4. DELAY OF JOINT SPEECH ENHANCEMENT AND SPEECH CODING SYSTEMS

Let M_E, M_C, and M_O denote the frame length of the enhancement preprocessor, the frame length of the speech coder, and the length of the overlapping section of the enhancement preprocessor frames, respectively. Then, the additional delay Δ_E of a noise reduction preprocessor when combined with a speech coder is given by

$$\Delta_E = M_O + \max_{k \in \mathbb{N}} \{[k(M_E - M_O)] \mathrm{mod}(M_C/l)\} \leq M_E \tag{1}$$

for $M_C/l \leq M_E - M_O < 2M_C/l$ and some given $l \in \mathbb{N}$. mod denotes the modulo operator. Since the enhancement and the coder frames might not match in size the variable k must run from $k = 1$ to $k = \infty$ to determine the maximum additional delay. The total algorithmic delay of the joint system is given by $\Delta_T = \Delta_E + \Delta_C$ where Δ_C denotes the algorithmic delay of the speech coder.

The total algorithmic delay of the joint enhancement and coding system is minimized if the frame shift of the noise reduction preprocessor is adapted to the frame size

of the speech coder such that $l(M_E - M_O) = M_C$ with $l \in \mathbb{N}$. This situation is depicted in Fig. 4.

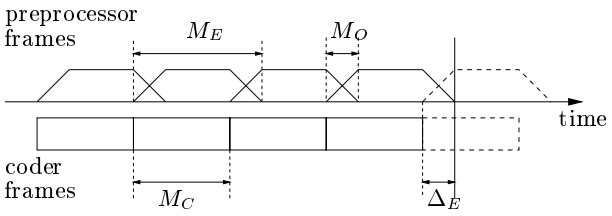

Figure 4: Frame alignment of enhancement preprocessor and speech coder with $M_E - M_O = M_C$.

For a given speech coder and within the framework of Fig. 4 there is usually more than one way to select parameters M_E and M_O and thus an opportunity to balance performance and delay of the joint system. For example for $M_C = 160$ both $M_E = 256$, $M_O = 96$ and $M_E = 128$, $M_O = 48$ are viable solutions. Clearly, the latter solution has less delay but also less frequency resolution on part of the enhancement.

We will discuss these tradeoffs again in the section 5 and also present a solution which allows to reduce the delay while keeping the frequency resolution of the enhancement preprocessor.

5. REDUCTION OF ALGORITHMIC DELAY

The delay of the enhancement algorithm is mainly determined by the spectral analysis/synthesis system. The analysis/synthesis system has to satisfy various conflicting requirements such as sufficient spectral resolution, little spectral leakage, smooth transitions between frames, low delay, and low complexity.

In this section we first look at method A. We stress the usefulness of a tapered synthesis window for a low delay overlap/add scheme. In section 5.2 we show how the input buffer of a parametric coder can be effectively utilized to reduce the additional delay of the enhancement preprocessing to about 1-3 ms while maintaining the spectral resolution in the analysis of the enhancement (method B).

5.1. Method A: Preprocessing Approach

In method A, the delay of the joint system is minimized when the frame advance of the enhancement system (or a multiple thereof) matches the frame advance of the codec. In this case the additional delay due to the enhancement is given by the length M_O of the overlapping sections of adjacent synthesis frames. Reducing the number of overlapping samples M_O, and thus the delay of the joint system has several effects. First, a reduction of the number of overlapping samples will reduce the sidelobe attenuation in the spectral analysis. This leads to increased crosstalk between frequency bins which might complicate the speech enhancement task. Most enhancement algorithms assume that adjacent frequency bins are independent and cannot exploit correlation between bins. Secondly, as the shift between frames is increased transitions between adjacent frames of the enhanced signal become less smooth. The discontinuities arise from the fact that the analysis window attenuates the input signal most at the edges of a frame and spectral estimation errors within a frame tend to spread evenly over the full frame. This leads to larger relative errors at the frame boundaries, and the resulting discontinuities which are most notable for low SNR (0-6 dB) conditions can lead to pitch estimation errors in the speech coder.

These discontinuities are greatly reduced if we use not only an analysis window but also a tapered synthesis window. We found that the square root of the Tukey window

$$w(i) = \begin{cases} \sqrt{0.5(1 - \cos(\frac{\pi i}{M_O}))} & 1 \le i \le M_O \\ \sqrt{0.5(1 - \cos(\frac{\pi (M-i)}{M_O}))} & M - M_O \le i \le M \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

gives good performance when used as an analysis and synthesis window. It results in a perfect reconstruction system if the signal is not modified between analysis and synthesis.

5.2. Method B: Integrated Approach

In the integrated joint enhancement and coding scheme we move the final overlap/add operation of the enhancement system into the input buffer of the speech coder.

Whenever a new signal frame is enhanced, only the part that overlaps with the data already in the input buffer of the speech coder is actually multiplied by the synthesis window and added to the data in the buffer. The non-overlapping part is multiplied by the inverse analysis window prior to the parameter extraction of the coder. After the codec parameters are extracted from the data in the input buffer the non-overlapping part is remultiplied by the analysis window and also multiplied by the synthesis window. After a shift by $M_E - M_O$ samples the input buffer is ready for the next input frame. Since the analysis window has a high attenuation at the frame edges multiplying the signal frames by the inverse analysis filter will greatly amplify estimation errors at the frame boundaries. We therefore leave a small delay of 1-3 ms and do not apply the multiplication with the inverse analysis filter to the last 8-24 samples of the input buffer.

The overlap/add procedure can be summarized as follows, where we assume that speech samples in the right half of the input buffer are more recent than speech samples in the left half:

- multiply input frame with analysis window;
- enhance input frame, compute IDFT;
- multiply left half of the enhanced frame with synthesis window and right half with inverse analysis window;
- add this frame to input buffer of coder;
- extract coder parameters;
- multiply right half of frame in the speech coder input buffer with analysis and synthesis window;
- shift data in input buffer before writing next frame into buffer.

6. LISTENING TESTS

The analysis/synthesis schemes were tested in conjunction with a state-of-the-art speech enhancement algorithm [1] and the 2.4 kbps MELP and WI speech coders. Informal and formal listening test were conducted for clean speech, medium SNR (6-12 dB) car noise, and low SNR (3-6 dB) conditions.

6.1. MELP Coder

For the 2.4 kbps MELP coder and method A we found that $M_E = 256$ and $M_O = 76$ results in good performance for all SNR conditions. However, the additional delay of the enhancement preprocessor is 9.5 ms in this case. Therefore we also conducted tests with the low delay method B which has a delay of 3 ms.

6.1.1. Method A

The preprocessing approach of method A was tested in a formal DAM (Diagnostic Acceptability Measure [6]) quality test and a formal DRT (Diagnostic Rhyme Test [6]) intelligibility test. In both tests the performance of the enhancement preprocessor in combination with the MELP coder was tested. The frame length of the enhancement preprocessor was $M_E = 256$. We compared an overlap/add version with $M_O = 128$, a Hanning analysis window and a rectangular synthesis window with a lower delay version with $M_O = 76$ and square-root Tukey analysis and synthesis windows (method A).

As expected the mean DAM test scores showed no quality reduction for clean speech and a very small reduction (within the standard error of the test) for noisy low SNR speech. The mean DRT test scores revealed no loss of intelligibility for noisy low and medium SNR speech and a slight intelligibility reduction for clean speech. This intelligibility reduction visible in the mean scores was also within the standard errors of the test.

6.1.2. Method B

A/B listening tests were carried out with clean and noisy speech (car noise, SNR about 10 dB) and 6 expert listeners. In these tests we compared the approach of method B with a delay of 3 ms to a method A system with 9.5 ms delay (Sec. 5.1). The same analysis and synthesis windows as for method A were used. For clean speech, listeners did not favour one of the systems. For noisy speech, listeners reported that it was often difficult to decide in favour of one of the two systems. Table 1 lists the result of the test in terms of percent of the total number of votes. Additional experiments which were conducted at a lower SNR condition (3-6 dB) gave similar results. In summary, we find a small performance degradation for noisy speech but at the same time the additional delay of the joint system has been reduced considerably.

	9.5ms (A)	3ms (B)	no prefer.
clean speech	28%	28%	44%
car noise (10 dB)	44%	28%	28%

Table 1: Results of an informal listening test comparing versions with 9.5 ms and 3 ms delay for the MELP coder.

6.2. WI Coder

As for the MELP coder we selected a frame length of $M_E = 256$ for the WI coder. Method A with an overlap of $M_O = 56$ results in 7 ms of delay. We compared this 7 ms system with a 3 ms system using method B. The processed speech samples were evaluated by means of an informal listening test for clean and car noise (5 and 10 dB) conditions. Simulation data consisted of 8 files (2 male and 2 females), and each file had a duration of 3-4 seconds. As before, six expert listeners were selected for this test. Table 2 summarizes the results. We find a very small preference for the system with the larger delay. The difference, however, is small enough to make the 3 ms system a serious alternative to the 7 ms system.

	7ms (A)	3ms (B)	no prefer.
clean speech	33.3	31.3	35.4
car noise (10 dB)	37.5	33.3	29.2
car noise (5 dB)	39.6	31.3	29.1

Table 2: Results of an informal listening test comparing versions with 7 ms and 3 ms delay for the WI coder.

7. CONCLUSIONS

We have shown that the delay of joint speech enhancement and speech coding systems can be reduced considerably if the overlap/add operation of the enhancement system is integrated into the speech coder. A novel overlap/add scheme was proposed for two different speech coders which led to solutions with less delay without a significant loss of quality. The application of this method to other speech coders requires a careful consideration of the coder parameter estimation procedures, especially the use of the look-ahead data.

8. ACKNOWLEDGMENTS

We thank Prof. David Malah for making his speech enhancement code available.

9. REFERENCES

[1] D. Malah, R. Cox, and A. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1999.

[2] R. Martin and R. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop on Speech Coding*, 1999.

[3] A. McCree, K. Truong, E. George, T. Barnwell, and V. Viswanathan, "A 2.4 KBIT/S MELP Coder Candidate for the New U.S. Federal Standard," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 200–203, 1996.

[4] W. B. Kleijn and J. Haagen, "Waveform interpolation for speech coding and synthesis," W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pp. 175-208, Elsevier Science Publishers, Amsterdam, 1995.

[5] W. B. Kleijn, Y. Shoham, D. Sen and R. Hagen, "A low-complexity waveform interpolation speech coder," *Proc. Int. Con f. Acoust. Speech Sign. Process.*, Atlanta, vol. I, pp. 212-215, 1996.

[6] S. Quackenbush, T. Barnwell III, and M. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.