NEW SPEECH ENHANCEMENT TECHNIQUES FOR LOW BIT RATE SPEECH CODING

*Rainer Martin*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, rvc@research.att.com

ABSTRACT

In this paper we present novel solutions for preprocessing noisy speech prior to low bit rate speech coding. We strive especially to improve the estimation of spectral parameters and to reduce the additional algorithmic delay caused by the enhancement preprocessor. While the former is achieved using a new adaptive limiting algorithm for the a priori signal-to-noise ratio (SNR) estimate, the latter makes use of a novel overlap/add scheme. Our enhancement techniques were evaluated in conjunction with the 2400 bps MELP coder by means of formal and informal listening tests.

1. INTRODUCTION

The quest for speech coders with lower bit rates has led to significant improvements of parametric speech coders in recent years. The MELP [1] and the WI [2] coder families, for example, have achieved speech quality ratings which are far superior to earlier parametric coder standards [3]. This notwithstanding, parametric coders still suffer severely from a lack of robustness in harsh acoustic environments. The artifacts introduced by such a coder when operated at medium (6-12 dB) to low (0-6 dB) SNR conditions can be very annoying and can impair intelligibility [3].

Informal and formal listening tests show that significant improvements are obtained when the speech coder is combined with a speech enhancement preprocessor. However, it turns out that the optimization of an enhancement preprocessor for a low bit rate speech coder is quite different from the optimization for other, e.g. listening, purposes.

This contribution presents novel speech enhancement techniques which improve the estimation of codec parameters and reduce the algorithmic delay of the joint system. Throughout this paper we will use the 2400 bps MELP [1] coder to demonstrate our results. In the remainder of this Introduction we briefly review the MELP coder parameters and the standard spectral weighting speech enhancement technique. We will then discuss our improvements to the LPC, the gain, and the pitch estimation as well as a method to reduce the additional algorithmic delay of the enhancement system to about 2-3 ms.

1.1. Parameters of the MELP coder

For each input signal frame of 180 samples the 2400 bps 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. 1. Since not all data in the buffer is used for all parameters we can exploit this to reduce the delay of the joint enhancement preprocessor and coding system (see Sec. 4).

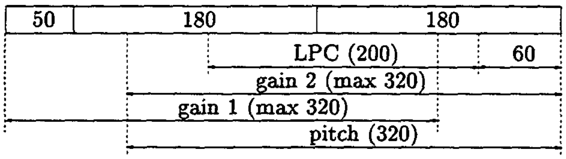

Figure 1: Utilization of data in the input buffer of the MELP coder. Numbers indicate frame sizes.

1.2. Spectral Weighting Speech Enhancement

Speech enhancement algorithms commonly consist of three major components: a spectral analysis/synthesis system (usually realized by means of a windowed FFT/IFFT), a noise estimation algorithm, and a spectral gain computation. The gain modifies only the Fourier magnitudes of an input frame. Noise estimation usually involves some kind of voice activity detection (VAD) [4] or spectral minimum tracking approach [5]. For perceptual reasons it is customary to overestimate the actual noise spectral density.

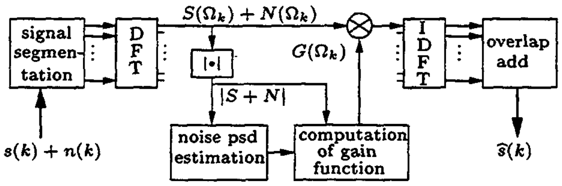

Figure 2: Block diagram of a typical single microphone spectral weighting speech enhancement algorithm.

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

0-7803-5651-9/99/$10.00 ©1999 IEEE.

2. IMPROVEMENT OF SPECTRAL PARAMETER AND GAIN ESTIMATION

After applying a standard speech enhancement technique [6, 7] to the MELP coder we find that most degradations and the loss of intelligibility are due to errors in the spectral parameters. In this section we present a modified spectral weighting rule which allows a better reproduction of the LPC/LSF parameters. The improvements come from an adaptive limiting procedure for the spectral gain factors which are applied to each DFT bin. More specifically we show that while spectral valleys in between formant frequencies are not important for speech *perception* (and thus can be filled with noise to give a better auditory impression) they are important for LPC *estimation.*

Because of its close relation to the Itakura-Saito measure we use the Minimum Mean Square Error Log Spectral Amplitude estimator (MMSE LSA) [6] as the basis of our approach. The MMSE-LSA minimizes $E\{(\log \widehat{A}_k - \log A_k)^2\}$ where A_k denotes the spectral speech amplitude in the kth bin and $\widehat{A}_k$ its optimal estimate. The solution to the minimiziation problem is obtained by applying a gain function $G(\xi_k, \gamma_k)$ [6] to the noisy spectral amplitudes and further improvements are obtained by using a multiplicatively modified MMSE-LSA estimator [4] which accounts for the probability of speech presence. ξ_k and γ_k denote the *a priori* and *a posteriori* SNR values for bin k [8].

It was stressed in [7, 4] that in order to avoid structured 'musical' residual noise and to achieve good audio quality, the *a priori* SNR ξ_k should be limited to values between 0.1 and 0.2. This means that less signal attenuation is applied to bins with low SNR and therefore the noise in the spectral valleys between formants is less attenuated. By limiting the attenuation the annoying 'musical' distortions are largely avoided and the residual noise appears to be very natural. However, by limiting the attenuation in the spectral valleys the overall spectral shape of speech sounds is distorted and thus the estimation of spectral parameters is disturbed. On the other hand, not limiting the ξ_k values introduces noticable fluctuations in the gain factors which result in annoyingly structured residual noise during speech pauses as well as audible distortions of speech for high SNR conditions. The solution to this problem is an adaptive limiting scheme which we outline below.

We utilize a voice activity detector to distinguish between speech+noise and noise only signal frames. Whenever we detect speech pauses we set a preliminary lower limit to $\xi_{minl} = \xi_{minP}$ (e.g. $\xi_{minP} = 0.12$ in this paper) in order to achieve a smooth residual noise. For speech activity the limit ξ_{minl} is set to

$$\xi_{minl} = \xi_{minP} \exp\{-5\}(0.5 + SNR)^{0.65} \qquad (1)$$

and ξ_{minl} limited to a maximum of 0.25. The preliminary limit is then smoothed by means of a first order recursive system

$$\xi_{min}(\lambda) = 0.9\xi_{min}(\lambda - 1) + 0.1\xi_{minl}(\lambda) \qquad (2)$$

to provide for smooth transitions between active and pause segments. λ denotes the frame index and SNR denotes the SNR of the speech sample. The resulting ξ_{min} is then used as a lower limit for ξ_k.

2.1. Experimental Results

The limiting algorithm was added to the speech enhancement system described in [4]. We notice that speech sounds appear to be less noisy with the adaptive limiting procedure while at the same time the background noise during speech pauses is very smooth and natural. Formal DAM and DRT listening tests were conducted to evaluate the benefits of this approach for the joint enhancement and coding system. Table 1 summarizes the mean scores and standard errors for unprocessed noisy speech, for coded speech without enhancement, for enhanced and MELP coded speech with constant ξ_{min}, and for enhanced and MELP coded speech with the adaptive limiting procedure. The SNR of the speech samples ranges between 3 and 6 dB (HMMWV noise). We note that the enhancement improves the performance of the joint system significantly and that the adaptive limiting procedure results in a gain in both the DAM and the DRT scores[1].

experiment	DAM/S.E.	DRT/S.E.
unprocessed	45.0/1.2	91.1/0.37
MELP coded	38.9/1.1	67.3/0.8
enhan.+coded, $\xi_{min} = 0.2$	51.0/0.8	69.5/0.54
enhan.+coded, ξ_{min} adapt.	52.0/0.9	72/0.64

Table 1: DAM and DRT scores and standard errors (S.E.) for experiments described in Sec. 2.1.

3. IMPROVEMENT OF PITCH ESTIMATION

Since the pitch estimation relies on the maxima of the autocorrelation function it can be improved by using a large noise overestimation factor. This, however, requires a separate enhancement branch just for the pitch which might not be justified as the complexity will be notably increased.

4. REDUCTION OF ALGORITHMIC DELAY

Since low bit rate speech coders already have a relatively large algorithmic delay any additional delay must be kept at a minimum. 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 stress the usefulness of a tapered synthesis window in the overlap/add synthesis procedure and show how the input buffer of a parametric coder can be effectively utilized to reduce the additional delay of the enhancement preprocessing to about 2-3 ms.

4.1. Analysis/Synthesis Windows

Considering a simple concatenation of the enhancement and coding algorithms 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

[1] In this test a preliminary version of the limiting algorithm was used which did not account for the SNR and resulted in some small distortions for clean speech.

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. As the shift between frames is increased from the typical half overlap of the synthesis frames (e.g. 128 samples) to the frame shift of the coder (e.g. 180 samples) 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 the frame and 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 conditions can lead to e.g. pitch estimation errors.

The 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

$$w(i) = \begin{cases} \sqrt{0.5(1-\cos(\pi i/M_o))} & 1 \le i \le M_o \\ \sqrt{0.5(1-\cos(\pi(M-i)/M_o))} & M-M_o \le i \le M \\ 1 & \text{otherwise} \end{cases} \qquad (3)$$

gives good performance when used as an analysis and synthesis window. It also results in a perfect reconstruction system if the signal is not modified between analysis and synthesis. Informal listening tests indicate that the quality loss of this scheme is low compared to a Hann windowed, half frame overlap scheme. The quality loss arises mainly from the discontinuities at frame boundaries and from the reduced time resolution. A DRT test revealed no loss of intelligibility for noisy speech and a small intelligibility reduction (93.8 vs. 94.5 for the half frame overlap approach) for clean speech.

4.2. New Low Delay Synthesis Scheme

The additional delay of our enhancement system combined with a MELP coder is still 9.5 ms because the frame size M equals 256 and the frame advance $M - M_o$ equals 180. The algorithmic delay of the joint system can be further reduced if we take explicit advantage of how the MELP coder utilizes the data in its input buffer. Of course, this approach can be applied to other coders as well.

Fig. 1 shows how the MELP coder extracts parameters from the data in its input buffer. The input buffer holds the data of the current frame as well as some past and look-ahead samples. 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 the MELP coder.

We therefore move the final overlap/add operation of the enhancement system into the input buffer of the speech coder with the option to reduce the additional algorithmic delay to about 2-3 ms.

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 computation 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 180 samples the input buffer is ready for the next input frame. Since the analysis/synthesis windows have 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 2-3 ms and do not apply the inverse analysis filter multiplication to the last 16-24 samples of the input buffer. A/B listening tests were carried out with clean and noisy speech (car and HMMWV noise) and 6 expert listeners. In these tests we compared the new approach to the system with 9.5 ms delay of Sec. 4.1. Listeners reported that it was often difficult to decide in favour of one of the two systems. For the HMMWV/car/clean conditions they preferred the 9.5 ms system in 43%/44%/28% of all cases over the 3 ms system. In 24%/28%/28% of the test cases they preferred the 3 ms system and in 33%/28%/44% they had no preference.

5. ACKNOWLEDGMENTS

We thank John S. Collura for his support and Prof. D. Malah for making his speech enhancement code available.

6. REFERENCES

[1] 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.

[2] W. Kleijn and K. Paliwal, eds., *Speech Coding and Synthesis*. Elsevier, 1995.

[3] M. Kohler, "A Comparision of the New 2400 BPS MELP Federal Standard with Other Standard Coders," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 1587–1590, IEEE, 1997.

[4] 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.

[5] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pp. 1182–1185, 1994.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[7] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, April 1994.

[8] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, December 1980.