Perceptual Pre-weighting and Post-inverse weighting for Speech Coding

Niranjan Shetty and Jerry D. Gibson
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA, 93106
Email: {niranjan, gibson}@ece.ucsb.edu

***Abstract*— We investigate the effect on voice quality of perceptual pre-weighting of the input speech to a codec, and post-inverse weighting the output of the codec. The G.726 adaptive differential pulse code modulation (ADPCM) codec and the AMR narrowband (AMR-NB) code excited linear prediction (CELP) codec are employed in our experiments. The weighting function used has the same form as that of the perceptual weighting function for the analysis-by-synthesis codebook search in AMR-NB. We observe a significant improvement in voice quality at rates of 16 and 24 kbps in the case of G.726 when perceptual weighting is used. When we use pre-weighting with the AMR codec, the unweighted squared error is used within the analysis-by-synthesis codebook search loop, and we find that the quality of the pre-weighted approach is comparable to the quality achieved by the standard AMR codec. The proposed pre-weighting method requires an additional bit-rate of 1.35 kbps to communicate the linear prediction (LP) coefficients of the original speech input to the decoder.**

I. Introduction

Perceptual weighting in the analysis-by-synthesis codebook search is a common feature of CELP-based coders and provides an improved quality relative to using mean square error (MSE) alone, through shaping of the error spectrum so that it is masked by the speech spectrum envelope. The origins of the function and form of the perceptual weighting filter can be traced back to the noise spectral shaping technique used in adaptive predictive coders [1], [2].

In this paper, we propose and investigate an alternative method that involves pre-weighting the input speech before it is processed through the codec, and then post-inverse weighting at the decoder. In our proposed model, we use the perceptual weighting filter as the pre-processor, and the inverse of the perceptual weighting filter as the post-processor. We show that this implementation results in a perceptual weighting of the end-to-end error envelope that can be effective in keeping the error spectrum below that of the input speech spectrum across the frequency band of interest. The proposed method was integrated with the G.726 and the AMR-NB codecs. In the case of G.726, the proposed method offers a significant improvement in voice quality with a small increase in bit rate. In the case of AMR-NB, the voice quality obtained using the proposed method is similar to the quality obtained in default operation of the AMR standard codec, with the potential for reducing computational complexity in the codebook search at the cost of a small increase in bit-rate of 1.35 kbps.

This work was supported by the California Micro Program, Applied Signal Technology, Dolby Labs, Inc. and Qualcomm, Inc., and by NSF Grant Nos. CCF-0429884 and CNS-0435527.

Adaptive pre-filtering and post-filtering have been employed in lossless audio coding [3], where a psycho-acoustic model is used as a basis for evaluating a set of LP coefficients that are used in the pre-filter, and are transmitted to the receiver for post-filtering. Relative to the perceptual audio coder (PAC), psychoacoustic pre- and post-filtering is said to provide a clear improvement for speech [4]. Another method that uses a pre-processor based on a psychoacoustic model for removing perceptual irrelevancy, defined as components of the speech input that cannot be detected by the ear, has been proposed and studied in [5]. The main contribution of our work lies in investigating the use of a pre-filter and post-filter based on the perceptual weighting function to add perceptual weighting to waveform-following speech coders and to move the perceptual weighting outside the analysis-by-synthesis codebook search for CELP speech coders.

The paper is organized as follows. In Section II, we demonstrate how pre-weighting and post-inverse weighting achieves a perceptual weighting of the error. In Sections III and IV, we describe the implementation and results for the proposed method used along with the G.726 and AMR narrowband speech coders respectively.

II. Pre-weighting principle

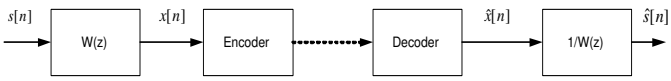


Fig. 1. Block diagram for proposed method using pre-weighting and post-inverse-weighting

Consider the system block diagram shown in Fig. 1. In the figure, $s[n]$ is the input speech, $x[n]$ is the pre-weighted speech input, $\hat{x}[n]$ is the pre-weighted speech output and $\hat{s}[n]$ is the output speech after post-inverse weighting. From the block diagram, we can write

$$S(z).W(z) = X(z) \tag{1}$$

and

$$\hat{X}(z).\frac{1}{W(z)} = \hat{S}(z) \tag{2}$$

Let $e[n] = x[n] - \hat{x}[n]$ denote the coding error for the pre-weighted speech. Therefore, in the z-domain

$$X(z) - \hat{X}(z) = E(z) \tag{3}$$

Substituting Eq. (1) and Eq. (2) in Eq. (3)

$$W(z)[S(z) - \hat{S}(z)] = E(z) \tag{4}$$

Thus, we see that through the use of pre-weighting and post-inverse weighting, we can achieve a perceptual shaping of the error envelope. The question we address now is whether the perceptual weighting is effective in improving the rate versus quality performance of a straightforward waveform coder such as G.726 ADPCM, and whether this pre-weighting and post-inverse weighting can be as effective in code-excited liner predictive coders as having perceptual weighting inside the analysis-by-synthesis loop.

III. Pre-weighting for G.726 ADPCM

The G.726 ADPCM speech codec [6] is a waveform coder that converts a 64 kbps A-law or μ-law pulse code modulated (PCM) waveform to a 40, 32, 24 or 16 kbps bit stream at the encoder and reconstructs the speech at the decoder. To obtain the 64 kbps input for G.726, the input file is first passed through the G.711 codec. Similarly, the output of the G.726 decoder is passed through the G.711 decoder to obtain PCM speech. Within the G.726 encoder, the input A-law/μ-law encoded speech is first converted into uniform quantized speech, before being passed into the adaptive differential pulse code modulation encoder. Similarly, at the receiver, the G.726 decoder is comprised of the ADPCM decoder and a uniform PCM-to-A-law/μ-law converter. In Fig. 2, the processing done by the blocks within the dashed region is referred to as the default operation of the G.726.

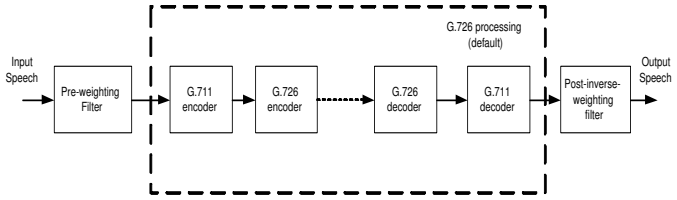


Fig. 2. Block diagram for G.726 processing with pre-weighting and post-inverse-weighting

In the proposed pre-weighting scheme, we use a pre-weighting filter and a post-inverse weighting filter, as shown in Fig. 2. The input speech is pre-weighted before it is passed into the G.726 codec. The G.726 processed speech is passed through the inverse of the weighting filter. The pre-weighting filter used has the form of the perceptual weighting filter used in the AMR codec, and is expressed as

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{5}$$

where γ_1 and γ_2 have values 0.94 and 0.6 respectively. The post inverse weighting filter has the form $1/W(z)$. Note that this process adds a perceptual weighting capability to the existing G.726 codec without modifying the standard codec and without including a much more complicated codebook search loop.

A. Determination of LP coefficents for each sub-frame

The process for obtaining the LP coefficients for each frame or subframe follows the procedure in the AMR-NB standard [7] for rates excluding 12.2 kbps. In the AMR-NB codec, LP analysis is performed once per speech frame (160 samples), using the autocorrelation method with 30 ms asymmetric windows and a look-ahead of 5 ms. The asymmetric window consists of a half Hamming window for the first part, while the second part is a quarter cosine function cycle. The coefficients of the 10^{th} order LP filter are obtained from the autocorrelation values using the Levinson-Durbin algorithm. The LP coefficients are then converted to LSP coefficients and are quantized and interpolated for each subframe.

We explore both frame-based and subframe-based pre-weighting. In frame-based pre-weighting, the LSP coefficients for the frame are converted back into LP coefficients and are used to obtain the pre-weighting function in Eq. (5). In sub-frame based pre-weighting, the set of quantized LSP parameters determined for the frame are used for the 4^{th} sub-frame, while those for the first three sub-frames are interpolated as follows [7]:

$$\begin{aligned} \hat{q}_1^{(n)} &= 0.75\hat{q}_4^{(n-1)} + 0.25\hat{q}_4^{(n)} \\ \hat{q}_2^{(n)} &= 0.5\hat{q}_4^{(n-1)} + 0.5\hat{q}_4^{(n)} \\ \hat{q}_3^{(n)} &= 0.25\hat{q}_4^{(n-1)} + 0.75\hat{q}_4^{(n)} \end{aligned} \tag{6}$$

where $\hat{q}_1$, $\hat{q}_2$, $\hat{q}_3$, and $\hat{q}_4$ are the quantized LSP vectors for each of the 4 subframes that comprise a frame, and the superscript n denotes the current frame. The quantized and interpolated LSP vectors are then converted back to LP coefficients, and are used in determining the pre-weighting filter. For the post-inverse weighting filter, only the quantized LSP coefficients for a frame need to be transmitted to the decoder. They can then be interpolated at the decoder in case of subframe-based pre-weighting, converted back to LP coefficients and employed in the post-inverse weighting filter. The LSP vector for each 20 ms frame are quantized using 27 bits. This translates to an additional coding rate requirement of 1.35 kbps.

B. Experimental Results

A comparison of frame-based and subframe-based pre-weighting reveals that the PESQ-MOS values for both frame-based and subframe-based processing are close. However, pre-weighting on a frame basis results in the presence of vertical striations in the spectrogram of the pre-weighted speech, that are not observed when the pre-weighting is done on a sub-frame basis. One of the reasons for using interpolated LP coefficients for each subframe in the case of speech coders is to smooth out the transients that are caused due to changes in LP coefficients from frame to frame [8]. This is one possible explanation for the striations observed in the case of the pre-weighted speech when the processing is done on a frame basis.

Therefore, in our experiments, subframe-based processing is adopted.

For our experiments, we used a narrowband speech file of duration 96 seconds comprised of 6 pairs of male sentences and 6 pairs of female sentences. PESQ-MOS [9] and informal listening tests were used to evaluate the quality of the processed speech files. In evaluating the PESQ-MOS, the speech file was split into 8 second long files, and the PESQ-MOS value was evaluated for each file. The average PESQ-MOS values were then computed over the 12 pairs of speech files.

TABLE I
AVERAGE PESQ-MOS VALUES FOR G.726 ADPCM FOR DEFAULT AND PRE-WEIGHTED OPERATION

	16 kbps	24 kbps	32 kbps	40 kbps
Default	2.87	3.41	3.78	3.98
Pre-weighted	3.40	3.75	3.97	4.10

The term default condition is used to indicate G.726 without pre-weighting, and we observe, from Table I, that pre-weighting results in an improvement in PESQ-MOS values for each rate supported by G.726. The improvement in PESQ-MOS values increases with a decrease in rates, ranging from a MOS increase of 0.12 at the maximum supported rate of 40 kbps, to an increase of 0.53 at the lowest rate of 16 kbps. Further, the PESQ-MOS for the pre-weighted case is close to the PESQ-MOS under the default operation for the next higher rate. In listening to the processed speech files, the pre-

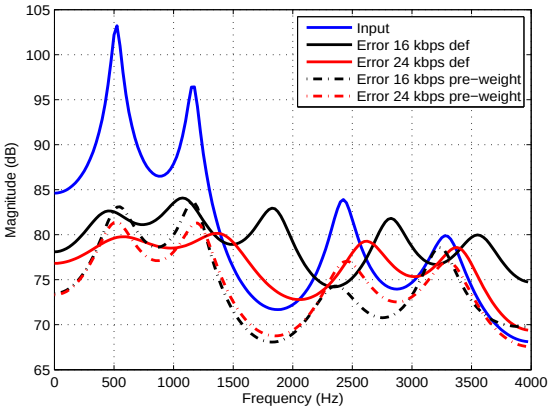
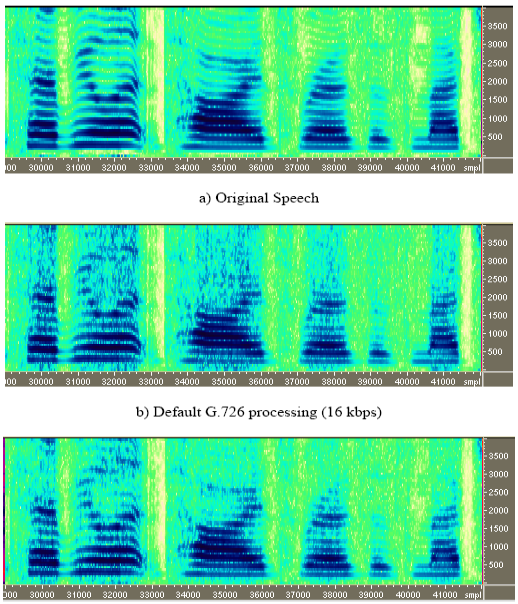


Fig. 3. Spectral envelopes for input and processed speech for G.726 rates 16 ans 24 kbps, for default and pre-weighted processing

weighted case corresponding to the rate of 16 kbps sounds relatively coarser compared to default 24 kbps coded speech, while being significantly better than default 16 kbps decoded speech. The pre-weighted speech corresponding to a rate of 24 kbps has a mild coarseness relative to default operation at 32 kbps. The pre-weighted speech corresponding to the rate of 32 kbps sounds mildly coarser in 4 sentences out of a total of 24 sentences, relative to default processing at 40 kbps.

Figure 3 shows the spectral envelopes for a sample male voiced frame of input speech and the error for the default case and the pre-weighted operation, for rates 16 and 24 kbps. From the figure, we see that the default case error spectrum for both



a) Original Speech

b) Default G.726 processing (16 kbps)

c) Pre-weighted processing (16 kbps)

Fig. 4. Spectrogram for a portion of the input speech and processed speech using default and pre-weighted processing respectively for G.726 at 16 kbps

the 24 and 16 kbps rates is significantly greater than the input speech spectrum in several frequency bands. However, the pre-weighting and post-inverse weighting results in a shaping of the error envelope that corresponds well with the input speech envelope, with the error spectrum falling below that of the input speech across the full band. The result is a noticeable improvement of the reconstructed speech as indicated by the PESQ-MOS values in Table 1.

Figure 4 contains the spectrograms for a section of the original speech and processed speech for default and pre-weighted codecs corresponding to the 16 kbps encoding rate. From the figure, the granular distortions in the reconstructed speech for the default codec are clearly evident in the spectrogram. The proposed pre-weighting scheme has a much cleaner looking spectrogram due to including perceptual weighting that is not part of the G.726 standard codec. As we move toward higher rates, the spectrograms for default processing have increasingly fewer artifacts, and at a rate of 40 kbps, the spectrograms for the default and pre-weighted speech are similar.

It is important to note here that only 27 bits per every 20 ms frame are needed to send the LP coefficients to the receiver for post-inverse weighting in the pre-weighted case. This translates to an additional rate of 1.35 kbps for pre-weighting. Thus we see that a significant improvement in quality can be attained at the cost of a small increase in bitrate when pre-weighting is used in G.726 ADPCM.

IV. PRE-WEIGHTING FOR THE AMR-NB CODEC

The narrowband AMR codec [7] is based on the CELP method, and encodes speech at 8 different bit rates, ranging from 4.75 kbps to 12.2 kbps. The coder operates on speech frames of size 20 ms (160 samples). In CELP speech synthesis,

two excitation vectors, one each from the fixed and adaptive codebooks respectively, are added and synthesized through a 10^{th} order LP synthesis filter. The optimal excitation vectors from the codebook are chosen at the encoder based on minimizing a perceptually weighted distortion criterion. The perceptual weighting filter is given by Eq. (5). The value of γ_1 is 0.9 for 10.2 kbps and 12.2 kbps, and 0.94 for all other modes.

The proposed pre-weighting and post-inverse weighting for the AMR-NB is the same as that used in Sec. III in that the input to the codec is passed through a pre-weighting filter that has the same form as the perceptual weighting filter, and the decoded output of the codec is passed through an inverse weighting filter. Additionally, for the pre-weighted case, the perceptual weighting conducted as a part of CELP codebook search in the encoder is disabled, and mean squared error (MSE) is used as the criterion to be minimized for choosing the excitation codevector. Since our objective is to evaluate and compare the performance of perceptual weighting, the standard post-filtering operation is disabled in each of these schemes. For the same reason as for the G.726 mentioned earlier, the pre-weighting is conducted on a sub-frame basis. For comparison, we also investigate the case where there is no pre-weighting and MSE is used instead of perceptual weighting in the CELP codebook search. This is referred to as the 'no-weighting' case.

A. Experimental Results

The speech file used and PESQ -MOS evaluation procedure are as described in Section III. Two modes of the AMR codec were used for our experiments: 4.75 kbps and 7.95 kbps. These modes represent the highest and lowest rates among those available in the AMR that employ a γ_1 value of 0.94 in the perceptual weighting filter. Further, the lower rate of 4.75 kbps tends to highlight the improvement due to perceptual pre-weighting, just as in the G.726 experiments described in the previous section.

TABLE II
AVERAGE PESQ-MOS VALUES FOR AMR-NB FOR DEFAULT, NO WEIGHTING AND PRE-WEIGHTING/POST-INVERSE WEIGHTING

AMR-NB 4.75 kbps	Default	3.38
	Pre-weighted	3.38
	No weighting	3.24
AMR-NB 7.95 kbps	Default	3.83
	Pre-weighted	3.82
	No weighting	3.7

The average PESQ-MOS values for processing under default, pre-weighted and no-weighting operation are shown in Table II. For both the 4.75 kbps and the 7.95 kbps cases, the PESQ-MOS values for pre-weighted operation are very close to the PESQ-MOS for the default operation. On listening, the speech files for the default and pre-weighted cases are found to be perceptually similar. Given that the AMR-NB codec is optimized for processing original speech, and not pre-weighted speech, this closeness in performance for the pre-weighted and

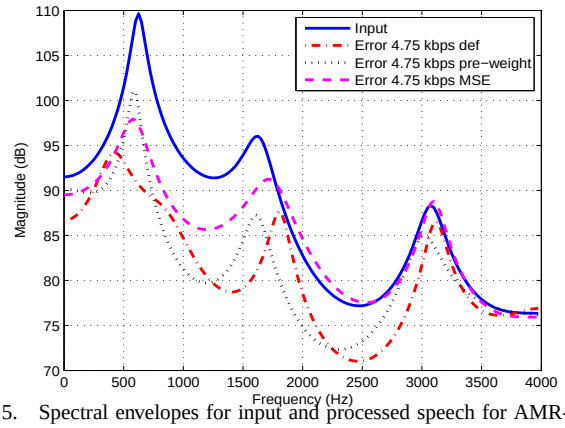


Fig. 5. Spectral envelopes for input and processed speech for AMR-NB for a rate of 4.75 kbps for default, pre-weighting, and no-weighting

default cases is quite remarkable. Both the default and pre-weighted cases have PESQ-MOS values that are better than the no-weighting case, with a difference of about 0.13 in MOS. We observe that the difference in terms of PESQ-MOS and even in terms of quality based on informal listening tests, between the pre-weighted operation and the no-weighting case is not as significant as in the case of G.726. This may be attributed to the observation that in CELP, even when no weighting is used, there is an inherent shaping of the error envelope that roughly follows the speech envelope [2], as seen in Figure 5. However, even thought the shaping does generally follow the input speech spectrum, the error envelope for MSE (without any weighting) rises above the speech envelope for a frequency range of 1700 to 2600 Hz, and hence, this shaping is not sufficient to achieve good quality speech.

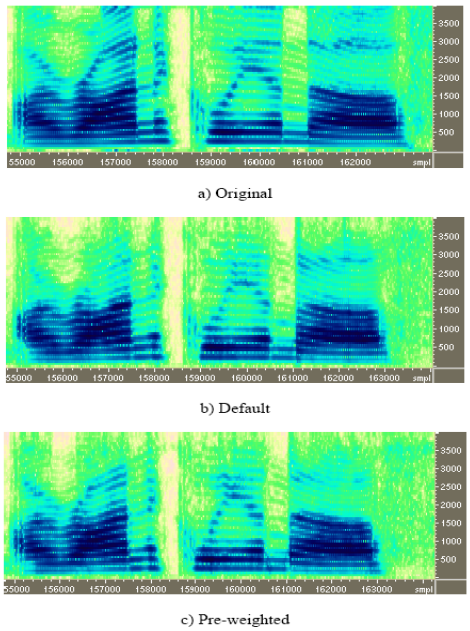


Fig. 6. Spectrogram for a portion of input and processed speech using default and pre-weighted processing respectively for AMR-NB at 4.75 kbps

In analyzing the LP envelope plots for the default and pre-weighted cases in Figure 5, we observe instances where one of them performs better than the other. In the spectral envelopes

for a sample frame of voiced female speech in Figure 5, we observe that the error envelope for the default case touches the input speech envelope between 1500-2000 Hz. Between 2500-3000 Hz, we also see the error envelope for the pre-weighted speech touching the speech envelope. Thus, in terms of this particular speech frame, a clear preference between the standard AMR codec with perceptual weighting and the AMR codec using squared error with pre-weighting is not evident.

In analyzing the spectrograms for the pre-weighted and default codecs over many speech segments, we observe that for most of the speech file used, the spectrograms are similar for the pre-weighted and default cases, with instances when either one of the default or pre-weighted case does better than the other. For example, in the spectrogram for the AMRNB 4.75 kbps processed female speech in Figure 6, the spectrograms for the pre-weighted speech are better organized and retain more spectral content relative to the default case, for frequency range 1500-2500 Hz, and sample range 156000-157000. Whereas comparing the default and pre-weighted spectrograms within the sample range of 161000-163000, and a frequency range of 2000-3000Hz, we find that the default case has clearer pitch harmonics relative to the pre-weighted case.

B. Prediction Gain evaluation for Pre-weighted Speech

The use of pre-weighting for the AMR-NB eliminates the need for weighting within the analysis-by-synthesis (AbS) loop in the CELP encoding process. Since the AbS loop is executed multiple times during the codeword search for each subframe, the proposed method results in a saving in computational complexity. Since the LP coefficients used for pre-weighting and the LP coefficients used in encoding the pre-weighted speech are different, the pre-weighting method requires that the speech LP coefficients be communicated to the receiver. This requires an additional bit-rate of 1.35 kbps.

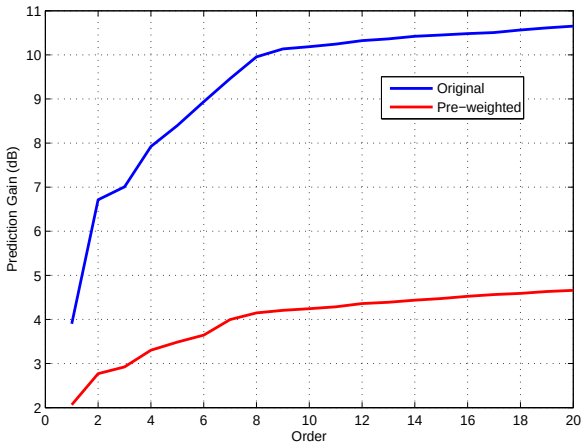


Fig. 7. Prediction gain for AMR-NB 4.75 kbps for the original and pre-weighted speech

The AMR-NB codec is a linear prediction codec designed to encode unprocessed speech, and hence the performance gain due to linear prediction may be less for the perceptually pre-weighted inputs for a given predictor order. To investigate this possibility, we calculate the prediction gain for the pre-weighted speech and for the original speech as a function of prediction order as shown in Figure. 7. The prediction gain is averaged over 50 frames of male and female voiced speech.

We see that the prediction gain for the original speech increases by about 6 dB when the predictor order is increased from 1 to 10, with a maximum prediction gain of 10.2 dB for the 10th order predictor. For the same increase in prediction order, the prediction gain for the pre-weighted speech increases by only about 2 dB and the 10th order predictor achieves a maximum prediction gain of under 4.5 dB. This suggests that savings in bit-rate and in complexity may be possible for the pre-weighted operation by using a lower-order predictor for the pre-weighted speech within the codec.

V. Conclusion

The pre-weighting and post-inverse weighting method proposed in this paper results in a significant improvement in voice quality for ADPCM coded narrowband speech with a small increase in bit rate of 1.35 kbps. Alternately, for the same voice quality, the proposed method results in a reduction in bit-rate of 6.65 kbps for each rate of the G.726 codec above 16 kbps. When used with the AMR-NB codec, the proposed method achieves the same quality as the AMR-NB default decoding without employing perceptual weighting in the analysis-by-synthesis codebook search loop. The proposed method comes with the advantage of reduced computational complexity, since weighting and inverse-weighting is done only once per speech subframe, compared to CELP coding where the perceptual weighting is performed multiple times within the AbS loop in determining the excitation codevectors. Further, pre-weighting reduces the correlation between the signal input to the codec, and may allow for a lower order predictor to be used in the codec. This advantage comes at the cost of an additional bit-rate requirement of 1.35 kbps, which is necessary for sending the parameters needed for post-inverse weighting at the decoder.

References

[1] B. S. Atal and M. R. Schroeder, “Predictive Coding of Speech Signals and Subjective Error Criterion,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 3, pp. 247–254, June 1979.

[2] J. D. Gibson, T. Berger, T. Lookabaugh, and R. L. Baker, *Digital Compression for Multimedia*. Morgan Kauffman, San Francisco, 1998.

[3] G. Schuller, B. Yu, D. Huang, and B. Edler, “Perceptual Audio Coding using Adaptive Pre- and Post-Filters and Lossless Compression,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 379–390, Sept 2002.

[4] B. Edler and G. Schuller, “Perceptual Audio Coding using Adaptive Pre- and Post-Filters and Lossless Compression,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 10, pp. 881–884, 2000.

[5] M. Lahdekorpi, “Perceptual Irrelevancy Removal in Narrowband Speech Coding,” *M.S. Thesis, Tampere University of Technology*, 2003.

[6] ITU-T Recommendation G.726, “40, 32, 24, 16 kb/s Adaptive Differential Pulse Code Modulation(PCM),” 1990.

[7] 3GPP, “TS 26.090 V6.0.0,” *Adaptive Multi-Rate (AMR) speech codec: Transcoding functions*, Dec. 2004.

[8] W. C. Chu, *Speech Coding Algorithms*. Wiley-Interscience, 2003.

[9] ITU-T Recommendation P.862, “PESQ: An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” Feb. 2001.