PERCEPTUAL WEIGHTING IN LSP-BASED MULTI-DESCRIPTION CODING FOR REAL-TIME LOW-BIT-RATE VOICE OVER IP

Dong Lin, Benjamin W. Wah, and Hang Yu

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois, Urbana-Champaign
Urbana, IL 61801, USA
{dlin,wah,yuhang}@manip.crhc.uiuc.edu

ABSTRACT

This paper focuses on improving an LSP-based multi-description coding (MDC) scheme for concealing losses when real-time CELP-coded speech data is sent over IP networks. In that scheme, LSP vectors are interleaved because they are highly correlated and can be reconstructed by interpolations when packets are delayed or lost. However, excitation vectors are random in nature and must be replicated in order to conceal losses. To avoid increasing the bit rate of the encoded speech, the excitation vectors are generated from longer subframes. This leads to lower quality of the decoded speech because noise from outside formant regions is over-emphasized, and more excitation information is extracted there. To improve quality, we propose in this paper to modify the perceptual-weighting filter (PWF) in the coder in order to adjust the allocation of noise inside and outside formant regions. We further study a method that selects the PWF in such a way that maintains high quality of the LSP-based MDC across different voice streams and loss scenarios. Experimental results on FS-1016 CELP (4800 bps), ITU G.723.1 ACELP (5300 bps), and ITU G.723.1 MP-MLQ (6300 bps) demonstrate noticeable improvements in decoding quality.

1. INTRODUCTION

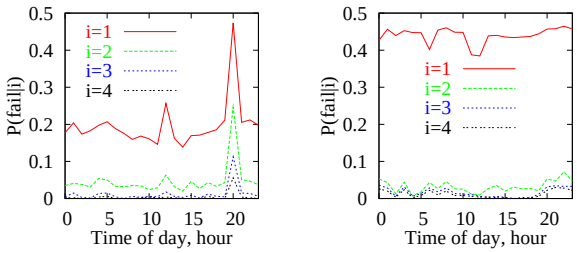
Background. With the rapid growth of the Internet, real-time Voice-over-IP (VoIP) is an attractive alternative to conventional public telephony. A decoded VoIP stream may have degraded quality when packets are lost or delayed because lost packets cannot be retransmitted in real time, and pervasive dependencies in the coded speech may lead to sustained distortions over a number of consecutive frames. These distortions cannot be overcome by source coding methods based on a channel-loss model because losses in the Internet are non-stationary and connection dependent [1].

Existing schemes for robust VoIP transmissions include those with added redundant information for loss concealments and those without. Typical schemes with redundancy include those using FEC codes and those sending extra content-dependent information (like voice indicators and pitch information) to facilitate recovery. The drawback of these schemes is that they incur a large bandwidth overhead. Schemes that use implicit redundancies recover lost speech by waveform substitution and from parameters in the speech signals received. These methods work well when losses are infrequent and the packet size is small, but fail to produce good quality under bursty and frequent losses.

Dissimilar to the single-description coding (SDC) schemes described above, multiple-description coding (MDC) is a popular scheme that enables robust VoIP transmissions without explicit redundancies. To apply MDC in low bit-rate coded speech, we have observed that the LSP coefficients in CELP-coded speech data are highly correlated across speech frames and that the excitation parameters are not. These observations lead to an MDC scheme that interleaves the LSP vectors (resp., that replicates the excitations) of a set of adjacent frames in multiple independent packets (called an interleaving set) [1]. When at least one packet in the interleaving set is received, the scheme allows the LSP information in those lost packets to be reconstructed by interpolations and the excitation information to be copied from the packets received.

Because the excitation information is replicated, the scheme enlarges the size of a subframe when extracting excitations in order to keep the bit rate constant. For example, in applying MDC with an interleaving degree of two to FS-CELP-1016 coded speech, the sender groups each pair of 240-sample frames in the original speech sequence, performs linear prediction analysis, once for each frame in order to generate a 34-bit quantized LSP vector, and dis-

RESEARCH SUPPORTED BY THE MOTOROLA CENTER FOR COMMUNICATIONS, UNIV. OF ILLINOIS, URBANA-CHAMPAIGN.
IEEE WORKSHOP ON MULTIMEDIA SIGNAL PROC., 2005.



a) UIUC-Western China b) UIUC-Slovakia

Figure 1: $Pr(fail|i)$, the probability of bursty losses that cannot be concealed under interleaving factor i, in two round-trip paths between UIUC and Western China (*www.lzu.edu.cn*) and Slovakia (*us.svf.stuba.sk*).

tributes the two LSP vectors to the two packets in an interleaving set. However, instead of generating a 220-bit vector of excitations for eight 60-sample subframes, it extends the subframe size to 120 samples, generates a 110-bit vector of codewords for four 120-sample subframes, and replicates the 110-bit vector to the two packets in the interleaving set.

Due to the use of enlarged subframes in extracting excitations, the decoded speech at receivers has degraded quality, even when all the packets in each interleaving set have been received. In this paper, we study the cause of this degradation and the effect of perceptual weighting on quality. We propose a systematic approach to select PWFs that can be generalized across voice streams and loss scenarios.

Internet traffic. Figure 1 plots $Pr(fail|i)$, the probability of non-concealable packet losses for given interleaving factor i over a 24-hour period for two round-trip connections from the University of Illinois (UIUC) to Western China and Central Europe. The graphs show that $Pr(fail|i)$ drops quickly when i increases; that $i = 2$ works well for the connection to Central Europe and achieves $Pr(fail|i)$ well below 5%; and that $i = 4$ is needed for the connection to Western China. Since the behavior is similar for other sites, our results suggest that an interleaving degree of two or four is sufficient for concealing losses when MDC is used.

Quality measures and test audio files. Besides common objective measures like the *Itakura-Saito Likelihood Ratio* (LR) and the *Cepstral Distance* (CD), subjective quality measures, such as the Mean Opinion Score (MOS), may be important for judging the quality of a decoded speech stream. Since standard MOS tests are difficult to carry out, we use the ITU-T P.862 recommendation, also called the Perceptual Evaluation of Speech Quality (PESQ), as an alternative to the subjective MOS tests.

In our test, we use eight audio files that include male, female and hybrid speeches of different durations. Three typical low-bit-rate speech codecs are used for testing. These include FS-1016 CELP (4.8 kbps), ITU G.723.1 ACELP (5.3 kbps), and ITU G.723.1 MP-MLQ (6.3 kbps).

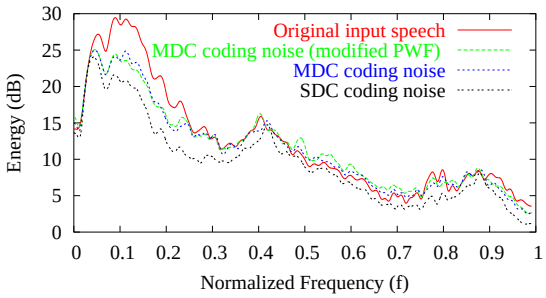


Figure 2: Energy-density spectra of the original stream ($|S(j2\pi f)|^2$), together with the coding noises in SDC ($|D_{S(\ell_0,0,0.8)}(j2\pi f)|^2$), 2-way MDC ($|D_{M(\ell_0,0,0.8)}(j2\pi f)|^2$), and 2-way MDC with modified perceptual weighted filtering ($|D_{M(\ell_0,0,0.94)}(j2\pi f)|^2$) for Audio File 0 coded by FS-1016 CELP. The default γ in FS-1016 CELP is 0.8.

2. IMPROVING EXCITATION QUALITY BY PERCEPTUAL WEIGHTING

Causes for quality degradation. Let $\mathcal{L}$ be the set of all possible loss scenarios, $\mathcal{L}_i$ be the set of loss scenarios under i-way MDC, and $\mathcal{V}$ be the set of all audio files. Given loss scenario $\ell \in \mathcal{L}$, voice stream $v \in \mathcal{V}$, and a coder-dependent PWF parameter γ (to be explained later), we define $s(n)$, $\hat{s}_{M(\ell,v,\gamma)}(n)$, and $\hat{s}_{S(\ell,v,\gamma)}(n)$ to be, respectively, the original speech, the LSP-based MDC decoded speech, and the SDC decoded speech. In particular, let $\ell_0 \in \mathcal{L}$ be the scenario with no loss.

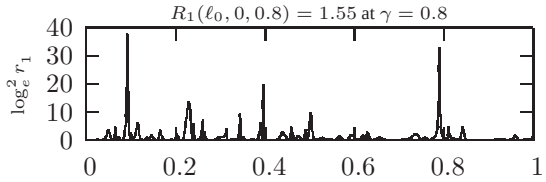
To identify the reasons for quality degradation, we evaluate $d_{M(\ell_0,v,\gamma)}(n) = s(n) - \hat{s}_{M(\ell_0,v,\gamma)}(n)$ (the coding noise of MDC under no loss) and $d_{S(\ell_0,v,\gamma)}(n) = s(n) - \hat{s}_{S(\ell_0,v,\gamma)}(n)$ (the coding noise of SDC under no loss). Figure 2 illustrates the energy-density spectra of the original stream and SDC's and MDC's coding noises (MDC with modified PWF is discussed later).[1] It shows that MDC's coding noise is much larger than that of SDC and has non-uniform increases across different frequencies.

To evaluate the quality of the MDC decoded speech, we compare it against two baselines. First, $r_1(f,\ell_0,v,\gamma) = \frac{|D_{M(\ell_0,v,\gamma)}(j2\pi f)|^2}{|D_{S(\ell_0,v,\gamma)}(j2\pi f)|^2}$ denotes the relative coding noise of MDC with respect to that of SDC at f. The increase over the entire spectrum can be evaluated by computing the geometric mean of r_1 over all f. This can be computed by taking the logarithm of each term and integrating:

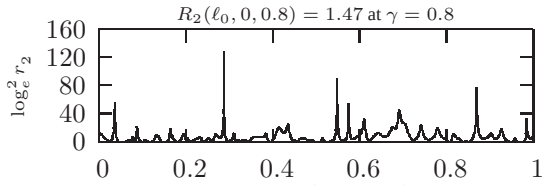
$$R_1{}^2(\ell_0,v,\gamma) = \int_0^1 \log_e^2 r_1(f,\ell_0,v,\gamma)df. \qquad (1)$$

Second, we evaluate $r_2(f,\ell_0,v,\gamma) = \frac{|\hat{S}_{M(\ell_0,v,\gamma)}(j2\pi f)|^2}{|S(j2\pi f)|^2}$, the energy of the MDC decoded speech at every f with respect to that of the original signal. We then compute the geometric mean over all f by first taking the logarithm of

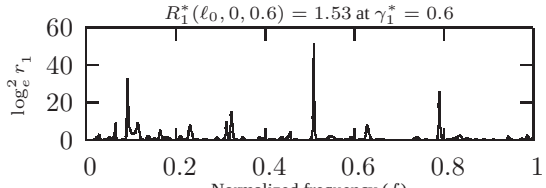
[1] The capitalized form of a symbol represents its Fourier Transform.



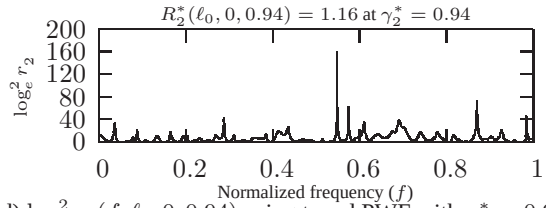
a) $\log_e^2 r_1(f, \ell_0, 0, 0.8)$ using original PWF with $\gamma = 0.8$



b) $\log_e^2 r_2(f, \ell_0, 0, 0.8)$ using original PWF with $\gamma = 0.8$



c) $\log_e^2 r_1(f, \ell_0, 0, 0.6)$ using tuned PWF with $\gamma_1^* = 0.6$



d) $\log_e^2 r_2(f, \ell_0, 0, 0.94)$ using tuned PWF with $\gamma_2^* = 0.94$

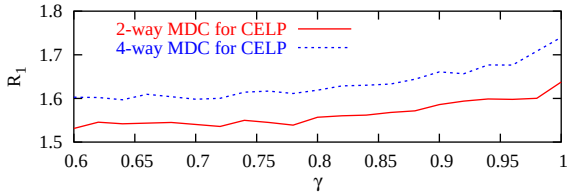
Figure 3: Density spectra showing the relative degradation of 2-way LSP-based MDC under no loss (ℓ_0) using the default $\gamma = 0.8$ and the best $\gamma_1^* = 0.6$ and $\gamma_2^* = 0.94$ in its PWF for Audio File 0 ($v = 0$) coded by FS-1016 CELP.

r_2 and then evaluating R_2^2:

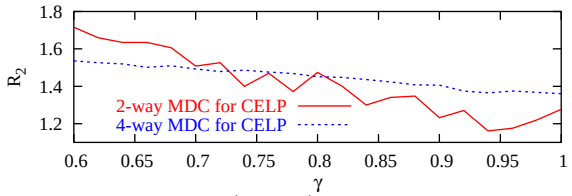
$$R_2{}^2(\ell_0, v, \gamma) \quad = \quad \int_0^1 \log_e^2 r_2(f, \ell_0, v, \gamma) df. \qquad (2)$$

Figure 3 illustrates, for the spectrum in Figure 2, the values of $\log_e^2 r_1(f, \ell_0, 0, 0.8)$ and $\log_e^2 r_2(f, \ell_0, 0, 0.8)$ over all f and those of $R_1(\ell_0, 0, 0.8)$, and $R_2(\ell_0, 0, 0.8)$. Figure 3a further shows that $\log_e^2 r_1(f, \ell_0, 0, 0.8)$ is large at $f = 0.1$, 0.4 and 0.8, which are exactly the spectral peaks in Figure 2.

According to speech perception principles, formants that generally correspond to spectral peaks have greater perceptual importance than spectral valleys. Hence, the noise-masking threshold should be higher in regions of spectral peaks than those of spectral valleys. Figure 3a shows that the noise energies of MDC in formant regions is excessive as compared to those of SDC. This over-emphasis leads to higher distortions inside formant regions because, under a fixed bit constraint on coding excitation information, more information will be extracted from outside formant regions and less from inside formant regions. The imbalance of



a) $R_1(\ell_0, 0, \gamma)$ for CELP



b) $R_2(\ell_0, 0, \gamma)$ for CELP

Figure 4: Effect of changing γ of PWF on R_1 and R_2 in 2-way and 4-way MDC under no loss (ℓ_0) and Audio File 0 ($v = 0$) coded by FS-1016 CELP.

noise energies also indicates that the standard PWF used in SDC is not tuned properly for MDC. Figure 3b shows that there are more distortions with respect to the original signal at high frequencies. This is common in many compression methods that discard details in the high-frequency band, leading to more distortions in this band.

Modified PWF. In most LPC speech coders, noise allocation is achieved by minimizing their perceptually weighted mean-square errors, or noises, between $s(n)$ and $\hat{s}_{S(l,v,\gamma)}(n)$. This weighting is done through a filter with parameter γ that shapes noises differently, depending on whether they are near spectral peaks or valleys. As an example, the PWF $W(z)$ for FS-1016 CELP is defined as $W(z) = \frac{A(z)}{A(z/\gamma)}$, where γ is fixed at 0.8, and $A(z)$ is the LPC analysis filter. As γ gets closer to one, the concaves will be shallower because the filter will be flatter, and noises near each peak are given more importance. In contrast, when γ is further away from one, formant-region noises will be suppressed more, and excitations near valleys will be more finely coded. Figure 4 illustrates the value of R_1 and R_2 when γ is changed in FS-1016 CELP.

As the relative coding noise of MDC is large at the spectral peaks of its energy-density spectrum (Figure 3a), we increase the γ of PWF in order to code the formant-region noises more finely. The change in γ causes R_1 to decrease and R_2 to increase (Figure 4). It further reshapes the energy-density spectra of the relative MDC coding noise (Figure 3c) and the relative MDC signal (Figure 3d). Comparing Figures 3a-3b and 3c-3d, adjusting γ leads to a minimal change in R_1 but a reasonable decrease in R_2.

Our proposed approach works for close-looped LP coders that generate excitations based on noises that are perceptually weighted. It does not apply to coders that do not use this mechanism, such as FS MELP.

Table 1: γ_j^a generalized across voice streams and loss scenarios.

Coder	FS-1016 CELP				G723.1 ACELP				G723.1 MP-MLQ			
MDC	2-way		4-way		2-way		4-way		2-way		4-way	
Metric	R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2	R_1	R_2
γ_j^a	0.62	0.94	0.60	0.98	0.85	0.85	0.45	0.85	0.30	0.90	0.20	0.90
$\Delta R_j^{\min}$	0.02	0.04	0.03	0.03	0.06	0.07	0.04	0.04	0.07	0.05	0.02	0.04

Optimization of R_j. The optimal γ_j^* under i-way MDC is:

$$\gamma_j^* \mid_{\ell \in \mathcal{L}_i, v \in \mathcal{V}} = \text{argmin}_{\gamma \in \mathcal{G}} R_j(\ell, v, \gamma), \quad j = 1, 2. \quad (3)$$

It is obvious that γ_j^* depends on the voice stream v tested, the loss scenario ℓ, and the objective measure R_1 or R_2. A simple strategy to address the dependence on voice streams is to aggregate the voice streams into one. This is not sound because different streams are of different durations, and the optimization will be biased by the longest stream. For the same reason, we cannot combine multiple loss scenarios by defining a probability distribution of the scenarios, as such a distribution is connection-dependent.

To address the issue, we propose a strategy to choose γ_j^* in such a way that yields the smallest degradation from the optimal performance and that works across all loss scenarios and test streams. Given loss scenario $\ell \in \mathcal{L}_i$ in i-way MDC, voice stream $v \in \mathcal{V}$, and an objective measure R_j, we first define the set of γ_j's whose objective value deviates from the optimum by ΔR_j:

$$\Gamma(\Delta R_j) = \left\{\gamma_j | R_j(\ell, v, \gamma_j) - R_j(\ell, v, \gamma_j^*) \leq \Delta R_j\right\}. \quad (4)$$

The main idea of the optimization is to make ΔR_j in (4) so small that there is exactly one unique γ_j^a across all loss scenarios and test streams. That is,

$$\text{Select } \Delta R_j = \Delta R_j^{\min} \text{ such that } \bigcap_{\ell, v} \Gamma(\Delta R_j^{\min}) = \{\gamma_j^a\}. \quad (5)$$

The γ_j^a found is the common γ_j that would incur a maximum degradation of $\Delta R_j^{\min}$ in each test stream and loss scenario. Table 1 shows the γ_j^a selected for the three coders.

Figure 2 illustrates the improved noise shaping using a modified PWF with R_2 as the performance measure and $\gamma_2^* = 0.94$.

Optimizing post-filters. In most current LPC coders, the LP-decoded speech is further post-filtered to enhance its quality. Since a post-filter is similar to a PWF, our proposed method for tuning PWFs can be applied directly. We omit those details due to space limitations.

3. EXPERIMENTAL RESULTS

Prototype. To test our proposed scheme, we have built a trace-driven simulator on a Linux computer. The speech is coded in real time and sent to a receiver, while packets are delayed or dropped according to the traces collected. Based on the loss statistics fed back from receivers, the sender chooses an appropriate interleaving factor.

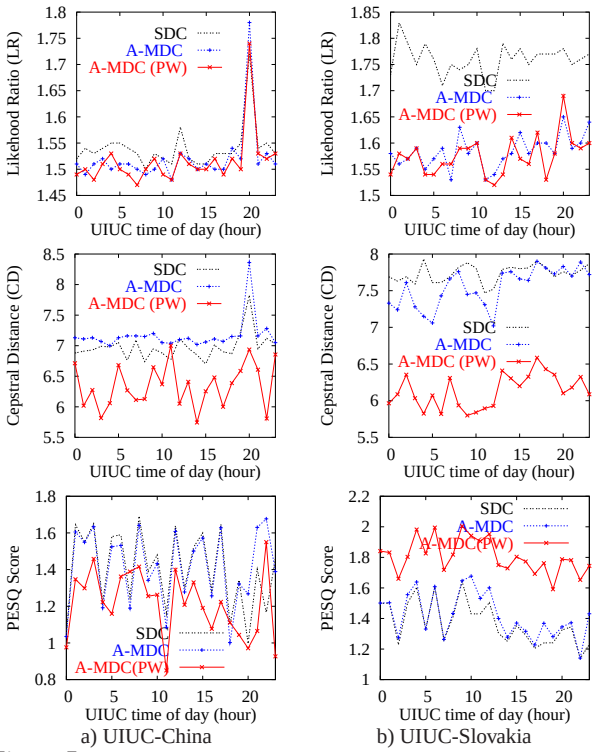


Figure 5: Comparisons of reconstruction quality using audio file 0 among SDC with no loss concealment, adaptive MDC (A-MDC), and adaptive MDC with improved perceptual weighting (A-MDC (PW)) for FS-1016 CELP. The experiments were conducted on received and reconstructed frames over a 24-hour period for the round-trip connections between UIUC and two remote hosts.

Results. We have evaluated the playback quality of speech transmitted using the traces collected. The quality in LR and CD is measured on a frame-by-frame basis and averaged over time, whereas PESQ is measured by treating the entire transmitted stream as a single audio file.

Figure 5 compares the results on simulating SDC with no loss concealment, adaptive MDC (A-MDC) with the original PWF, and adaptive MDC with the proposed PWF for FS-1016 CELP. The results on the other two coders are not shown due to space limitations. The LR of the three schemes are similar because LR reflects mostly the quality of the LPC vectors, which can be reconstructed accurately. The CD of our proposed MDC scheme is better because CD reflects the quality of the excitations, which can be reconstructed better with the modified PWF. The three schemes have similar PESQ because PESQ is not sensitive to changes in MOS within 0.5.

4. REFERENCES

[1] D. Lin and B. W. Wah, "LSP-based multiple-description coding for real-time low bit-rate voice over IP," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 167–178, Feb. 2005.