IMPROVED PREDICTION OF NEARLY-PERIODIC SIGNALS

W. Bastiaan Kleijn[†] *and Jan Skoglund*[*]

[†]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
[*]Google, Inc., 1600 Amphitheatre Parkway, Mountain View, USA

ABSTRACT

We present methods that are relevant for the commonly used pitch predictors in speech/audio coders. We first discuss optimal pre- and post-filtering and derive a general result that post-filtering is more effective than pre-filtering. This, together with the perceived nature of the distortion, suggests the omission of the pre-filter if system delay is critical. We propose a practical paired-zero filter design for the low-rate regime. We extend this design to handle frequency-dependent periodicity levels. We also provide a general performance measure for a post-filter that only uses information available at the decoder. This criterion allows the optimization or selection of a post-filter without increasing the rate. Our experiments show that the resulting methods provide state-of-the-art performance both for objective measures and in terms of quality as perceived by test subjects.

1. INTRODUCTION

Rate-distortion optimal encoding of a stationary signal according to a squared-error criterion results, in general, in a stationary signal that has a power-spectral density that differs from that of the original signal. For the stationary Gaussian (SG) signal case, the phenomenon is well understood and referred to as *reverse waterfilling* (e.g., [1]).

We are interested in the coding of speech/audio signals, which are commonly encoded using *transform coding* [2]. For transform coding, reverse waterfilling does not need to be considered explicitly. However, a major disadvantage of transform coding is that it requires a significant delay. Particularly in applications such as webjamming and those where also a direct acoustic path exists (e.g., flight-control rooms, remote microphones for hearing-aids), this delay can be prohibitive. This motivates the use of predictive coding (sometimes only for pitch), which can operate at a much lower delay.

While predictive coding is an effective method for coding at a low delay, its rate-distortion performance at low rate was poorly understood until recently. It does not naturally provide reverse waterfilling. It has long been known that the squared-error performance of predictive coding is not optimal and can be enhanced by postfiltering [3]. The relation to Wiener filtering has been cited as a motivation for the squared-error performance improvement of the post-filter [4]. However, the Wiener filter is optimized for a clean signal contaminated with additive, statistically independent noise, but for optimal coding of a SG signal the error signal is independent of the *reconstructed* signal, but not of the original signal [1].

Chen and Gersho [4] suggest that a major motivation for post-filtering is perception. It is, however, beneficial to separate rate-distortion optimization and processing for perception. The signal can be transformed to a domain where the coding criterion is an accurate representation of perception. A simple filtering structure that was proposed by [5] is commonly used for this purpose. Thus, perception does not need to be considered in the improvement of predictive coding.

A rigorous approach to account for reverse waterfilling in the context of analysis-by-synthesis predictive coding was presented in [6]. The system was implemented for a first-order filter and the solution is approximate for low rates. Conventional post-filtering was interpreted as an approximation of the proposed method.

A solution to optimal coding of SG signals using prediction can be based on dithered quantization [7]. The solution is based on insight gained from the *optimum test channel*, which is a formal solution to the rate-distortion problem. In its *forward* form the test channel is a *pre*-filtering, a noise addition, and a *post*-filtering. A realizable structure that is asymptotically optimal is obtained if the noise addition operation is replaced by predictive dithered quantization [7], exploiting that the quantization noise in a dithered quantizer is additive. It can then be shown that rate-distortion optimal performance can be obtained if parallel sources are encoded with one vector quantizer [7]. In this case the post-filter is a Wiener filter that has the input of the quantizer as target signal.

Zamir [7] proved that a scalar predictive entropy-constrained dithered quantizer (ECDQ) scheme with pre- and post-filtering is rate-distortion optimal for SG signals except for a space-filling loss of 0.254 dB. This performance was earlier shown for a special case by [8] by means of numerical optimization of pre- and post-filtering (and noise shaping) using a conventional quantizer without dither. The performance for the stationary Gaussian case was also confirmed by [9]. The method also performs well when applied to practical audio signals; a first practical scheme was reported in [9].

Our main aim is to improve pitch predictors that are used to model spectral fine structure in speech/audio coders by post-filtering. We show that post-filtering is more effective than pre-filtering in section 2. In section 3 we introduce a new postfilter design and a distortion measure that facilitates creating the post-filter at the decoder.

2. CODING NEARLY-PERIODIC AUDIO SIGNALS

Voiced speech often exhibits a high level of periodicity, particularly at frequencies below 1500 Hz. The periodicity can start abruptly at a voicing onset. Music instruments can display similar behavior.

A so-called long-term predictor is commonly used to model the periodic behavior of speech in source coding. The prediction filter generally has a single tap, at the pitch period (delay), P. The single tap is often generalized to facilitate fractional delay. Our solutions discussed generalize to this case. This section states some new results relevant for pitch post-filtering. The results assume stationary Gaussian (SG) signals.

We begin with considering a process X_n that has a flat spectral envelope and is encoded using a generalized single-tap pitch predictor. The pitch predictor models the signal as an autoregressive (AR) process with power-spectral density

$$S_X(\mathrm{e}^{j\omega}) = \frac{\sigma^2}{|1-\alpha \mathrm{e}^{-j\omega P}|^2}, \qquad (1)$$

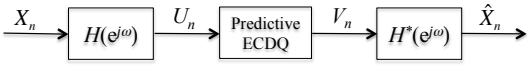

Fig. 1: Predictive entropy-constrained dithered quantization (predictive PECDQ) with pre- and post-filtering.

where $\alpha > 0$ is a real coefficient and σ^2 determines the signal power. The spectral density (1) is periodic with fundamental frequency $\frac{2\pi}{P}$.

Let us consider the optimal coding of the AR process of (1). Let $\lambda \geq 0$ represent the so-called water level that determines the coding rate and distortion. The distortion is

$$D(\mathrm{e}^{j\omega}) = \begin{cases} \lambda, & \lambda \leq S_X(\mathrm{e}^{j\omega}) \\ S_X(\mathrm{e}^{j\omega}), & \text{elsewhere} \end{cases} \tag{2}$$

If the condition $\lambda \leq S_X(\mathrm{e}^{j\omega})$ is true for for all ω (i.e., the system operates in the high-rate regime), then the power-spectral density $S_{\hat{X}}$ can be realized with an easily realizable rational filter [6]. Optimal performance can be obtained with the predictive coding structure described in [7], which uses ideal pre- and post-filters and ECDQ. Fig. 1 outlines the basic configuration. The magnitude response for the ideal pre- and post-filters is determined by

$$|H(\mathrm{e}^{j\omega})|^2 = 1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})}. \tag{3}$$

The phase response of the pre-filter is arbitrary but the response of the post-filter must be the complex conjugate of that of the pre-filter. For the one-tap predictor of (1) the response (3) becomes

$$|H(\mathrm{e}^{j\omega})|^2 = \begin{cases} 1 - \frac{\lambda}{\sigma^2}(1 + \alpha^2 - 2\alpha\cos(\omega P)), & \lambda \leq S_X(\mathrm{e}^{j\omega}) \\ 0, & \text{elsewhere} \end{cases} \tag{4}$$

The magnitude response as given by (4) has maxima at $\omega = \frac{n2\pi}{P}$, $n \in \mathbb{Z}$; the gain is near unity for $\alpha \approx 1$. For the high-rate regime $\lambda \leq S_X(\mathrm{e}^{j\omega}), \forall\omega \in [-\pi, \pi]$, the frequency response $H(\mathrm{e}^{j\omega})$ can be implemented exactly with an all-zero filter with its zeros at $\omega = \frac{\pi + n2\pi}{P}$ $n \in \mathbb{Z}$. For the low-rate regime, the response (4) does not have a practical analytic solution. In section 3 we propose an approximate solution.

As the pre- and post-filters introduce delay, we consider the effect of omitting either the pre- or post-filter. For mathematical expediency, we consider a SG process and a general predictive coder with infinite-order predictor. The pre- and post-filters are those optimized for the case that both exist.

Let us consider the coding operation including both pre- and post-filtering. From (3) it is seen that the pre-filtered signal as a power-spectral density

$$S_U(\mathrm{e}^{j\omega}) = (1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})})S_X(\mathrm{e}^{j\omega}). \tag{5}$$

Let us assume the filter to have zero phase. The signal distortion $X_n - U_n$ in U_n then has power-spectral density

$$S_{X-U}(\mathrm{e}^{j\omega}) = 2(1 - \sqrt{1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})}})S_X(\mathrm{e}^{j\omega}) - D(\mathrm{e}^{j\omega}). \tag{6}$$

The pre-filtered signal U_n is subjected to the predictive dithered quantizer, which adds white quantization noise W_n with a power spectrum λ, assuming the predictor is optimal for the noisy output of the dithered quantizer [7].[1] Under these conditions, the predictive ECDQ of Fig. 1 is equivalent to the forward test channel shown. As the quantization noise W_n is independent from the signal X_n, the output V_n of the dithered quantizer has an error power-spectral density

$$S_{X-V}(\mathrm{e}^{j\omega}) \geq 2(1 - \sqrt{1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})}})S_X(\mathrm{e}^{j\omega}). \tag{7}$$

Note that for small $\frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})}$ (7) converges to $D(\mathrm{e}^{j\omega})$.

The output V_n of the predictive dithered quantizer consists of two independent components: the signal component U_n with power-spectral density $S_U(\mathrm{e}^{j\omega})$ and the noise component W_n with power-spectral density λ. After post-filtering, the estimated signal $\hat{X}_n$ is obtained. It has a signal component that has power-spectral density $S_X(\mathrm{e}^{j\omega})(1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})})^2$ and a signal component distortion spectral density $S_X(\mathrm{e}^{j\omega})\frac{D(\mathrm{e}^{j\omega})^2}{S_X(\mathrm{e}^{j\omega})^2}$. The noise component is attenuated to have an output power-spectral density

$$\lambda(1 - \frac{D(\mathrm{e}^{j\omega})}{S_X(\mathrm{e}^{j\omega})}) = D(\mathrm{e}^{j\omega}) - \frac{D(\mathrm{e}^{j\omega})^2}{S_X(\mathrm{e}^{j\omega})}. \tag{8}$$

The sum of the signal distortion and the noise component in the output is, as expected, $S_{X-\hat{X}} = D(\mathrm{e}^{j\omega})$.

Using the same approach, we analyzed what happens if the pre- and post-filters are omitted. If the pre-filter is omitted we have to account for the resulting change in rate. The main result of our analysis is: *Consider the encoding and decoding of a stationary Gaussian process with an optimal predictive ECDQ quantizer that produces Gaussian quantization noise with variance λ. Let the pre- and post-filters be defined by (3) and have zero phase. Then the ratio of the rate increase and the distortion reduction of using only a post-filter instead of only a pre-filter is never more than $\frac{1}{2\lambda}$.*

A corollary of our result is that if the filters are restricted to be of the form (3) and have zero phase then post-filtering is more effective than pre-filtering. A perception-based perspective generally leads to a similar conclusion: the pre-filter pre-distorts the signal, reducing the rate required at a certain SNR, whereas the post-filter reduces quantization noise, cf. (8), which has a strong perceptual impact. More-over post-filters do not affect the operation of existing coders.

3. A PRACTICAL PITCH POST-FILTER DESIGN

The optimal response of pre- and post-filter given (4) can be implemented by an all-zero structure of the form

$$A_{\mathrm{ltpf}}(z, \beta_0, \beta_1) = \beta_0(1 + \beta_1 z^{-P}). \tag{9}$$

where P is the pitch delay in samples (as before, the logic generalizes to fractional-delay pitch).

The filter (9) has two significant drawbacks. First, it is not valid for the low-rate regime ($S_X(\mathrm{e}^{j\omega}) < \lambda$ for finite intervals of ω), which is the normal operating mode for pitch predictors. Second, most audio signals vary in periodicity level with frequency. Next we describe a post-filter that alleviates both disadvantages.

Let us consider the real filter coefficient β_1. Rotating this coefficient by $\mathrm{e}^{P\omega_0}$ results in

$$A_{\mathrm{ltpf}}(z, \beta_0, \mathrm{e}^{P\omega_0}\beta_1) = \beta_0(1 + \mathrm{e}^{P\omega_0}\beta_1 z^{-P}). \tag{10}$$

[1] To facilite analysis, including verification of optimality, the quantization noise must be Gaussian; this can be accomplished only by simultaneously quantizing multiple channels [7].

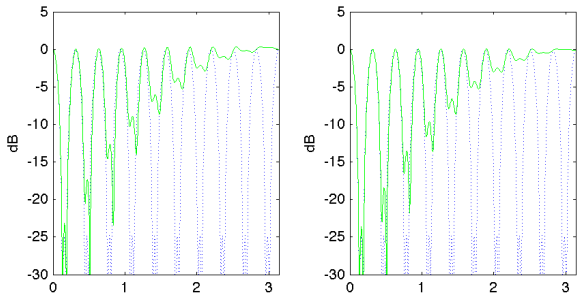

Fig. 2: Example of the filter responses obtained using (12). The effect of the paired zeros is clearly visible.

While the corresponding filter now results in complex output, it can be used as a building block for a filter with real output. Let us consider the concatenation of two filters: one where we rotate the zeros clockwise, and one where we rotate the zeros counterclockwise by the same amount. The filter

$$B_{\mathrm{ltpf}}(z,\beta_0,\mathrm{e}^{P\omega_0}\beta_1) = A_{\mathrm{ltpf}}(z,\sqrt{\beta_0},\mathrm{e}^{-P\omega_0}\beta_1)\ A_{\mathrm{ltpf}}(z,\sqrt{\beta_0},\mathrm{e}^{P\omega_0}\beta_1) \quad (11)$$

is real, has the same maximum gain as the filter $A_{\mathrm{ltpf}}(z,\beta_0,\mathrm{e}^{P\omega_0}\beta_1)$, but has broader valleys.

A filter with an appropriate frequency-dependent gain can be obtained by mixing the filter (11) and a unit-response filter with a gain of β_0 (in practice a delay is also required). Let $H_{\mathrm{lp}}(z,\mu)$ be a linear-phase low-pass filter with one adjustable parameter μ and a unity gain at $\omega = 0$. The complementary high-pass filter is then $1 - H_{\mathrm{lp}}(z,\mu)$. This enables us to create a long-term post-filter with frequency-varying periodicity by creating the following filter:

$$G(z) = B_{\mathrm{ltpf}}(z,\beta_0,\mathrm{e}^{P\omega_0}\beta_1)H_{\mathrm{lp}}(z,\mu)+\beta_0(1-H_{\mathrm{lp}}(z,\mu)). \quad (12)$$

Fig. 2 shows two examples of filters designed in this manner. An analytic solution to the simultanous optimization of the filter $H_{\mathrm{lp}}(z,\mu)$ and $B_{\mathrm{ltpf}}(z,\beta_0,\mathrm{e}^{P\omega_0}\beta_1)$ is cumbersome. In practice we use selection from a fixed set of pre-defined filters with the criterion that is described in the next section.

4. DECODER-BASED PARAMETER SELECTION

It is possible to select the parameter settings based directly on the output of the predictive ECDQ before the post-filter. In this section we assume the pre-filter is omitted and that the power-spectral density of the output of the predictive ECDQ, $S_{\tilde{Y}}(\mathrm{e}^{j\omega})$ (the $\tilde{}$ indicates omission of the pre-filter) as well as quantization noise variance λ are known. In practice this means that the post-filter parameters can be estimated at the decoder. It is straightforward to extend the method for quantization noise that is not spectrally flat. The criterion is general and applies to any type of post-filter.

Using the fact that a predictive ECDQ results in additive quantization noise, its output spectral density $S_{\tilde{Y}}(\mathrm{e}^{j\omega})$ can be split into a signal contribution $S_X(\mathrm{e}^{j\omega}) = S_{\tilde{Y}}(\mathrm{e}^{j\omega}) - \lambda$ and a noise contribution λ. (In existing coders, these contributions are considered of equal importance. This is not necessarily correct from a perceptual viewpoint.) Let the frequency response of the post-filter be $f(\mathrm{e}^{j\omega},\theta)$ with parameters θ. The filter typically satisfies $0 \le |f(\mathrm{e}^{j\omega},\theta)|^2 \le 1, \forall\omega \in [-\pi,\pi]$. To determine the optimal θ we minimize the total squared error:

$$\hat{\theta}' = \operatorname*{argmin}_{\theta} \frac{1}{2\pi}\int_{-\pi}^{\pi} |1-f(\mathrm{e}^{j\omega},\theta)|^{\xi}\left(\frac{S_{\tilde{Y}}(\mathrm{e}^{j\omega})}{\lambda}-1\right)d\omega - \frac{b}{2\pi}\int_{-\pi}^{\pi} 1-|f(\mathrm{e}^{j\omega},\theta)|^{\xi}d\omega, \quad (13)$$

where $\xi = 2$ and $b = 1$. In (13) the first term describes the distortion of the original signal introduced by the post-filter and the second term is a measure of noise removal by the post-filter.

Note that if f is real (as it would be for an optimal Wiener filter), then $|1-f|^2$ is concave and $|f|^2$ is convex. This implies that at low attenuation levels$f \sim 1$ the distortion term is relatively small, whereas the noise removal term is relatively large. As a result spectral regions without spectral structure may affect the filter selection process. This effect can be reduced with a heuristic power coefficient ξ suitably in the range $1 \le \xi \le 2$. This and the selection of b accounts for differences in perception between the two components.

An important property of (13) is that it favors post-filters that have a structure similar to the signal. For pitch prediction this implies that if the signal $S_{\tilde{Y}}(\mathrm{e}^{j\omega})$ does not display a harmonic structure in some region, then a post-filter with no periodicity enhancement is favored. Note also that if $S_{\tilde{Y}}(\mathrm{e}^{j\omega})$ is underestimated for any reason, then the criterion will tend toward favoring periodicity enhancement even if the signal is not periodic. This can be prevented by considering frequency bands separately and ensuring that the overall signal-to-noise ratio is reasonable in each band. From a computational perspective, it is advantageous to determine the pitch separately.

5. EXPERIMENTAL RESULTS

We performed experiments for both artificial data and for speech signals. For the artificial data we did two experiments with an AR process having a spectrum given by (1), using a forward test-channel simulating predictive entropy-constrained dithered quantization adding noise with different λ, and averaging multiple realizations of the process and filterings. The process parameters selected for this example were $\{P,\alpha,\sigma\} = \{80, 0.97, 5\}$. Delay compensation was utilized to calculate distortion.

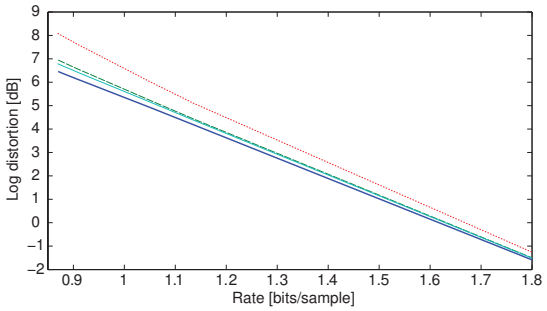

Fig. 3: High-rate performance rates using optimal pre- and post-filters from (4). No pre- or post-filter (dotted red), both pre- and post-filter (bold blue), only pre-filter (dashed green), and only post-filter (thin cyan).

In the first experiment we used all-zero filters given by (9) (optimal at high rates). Fig. 3 presents the log distortion of four systems: no filtering, both pre-and postfiltering, and only pre- or post-filtering. The plots start at the rate where $\lambda = S_X(\mathrm{e}^{j\omega})$ occurs, which in our example is 0.87 bits/sample. The bold blue curve is the optimal performance using both filters. The other curves confirm that using only

a post-filter is better than using only a pre-filter. The curves converge at high rate since the optimal filters converge to unity.

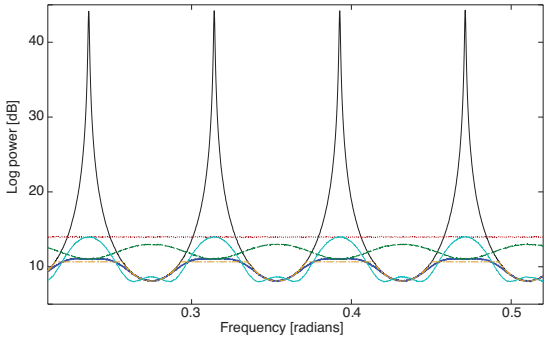

Fig. 4: Coding of the AR process at low rate (0.48 bits/sample) using pre- and post-filters from (11). Signal spectrum (black), error spectra: no pre- or post-filter (dotted red), both pre- and post-filter (bold blue), only pre-filter (dashed green), only post-filter (thin cyan), and RD-optimal (dashed-dotted orange) from (2).

In the second experiment we used new paired-zero filters with parameters $\{\beta_0, \beta_1, \omega_0\} = \{1, 0.99, 0.15\}$. Fig. 4 depicts signal and distortion spectra when coding the AR process at a low rate (0.48 bits/sample). Note that the spectra are only plotted for a part of the frequency range. The curve showing both pre- and post-filters closely approximates the optimal performance and again a post-filter only is better than a pre-filter only.

Fig. 5 depicts the performance of the paired-zero filters for rates corresponding to the high rate results in Fig. 3. We see that that at rates between 0.4 and 0.6 bits/sample a pre- and post-filter combination reaches a nearly optimal performance. Again, a post-filter only is performing better than a pre-filter only setup. When the rate increases the paired-zero filters are sub-optimal as their parameter settings are not adapted.

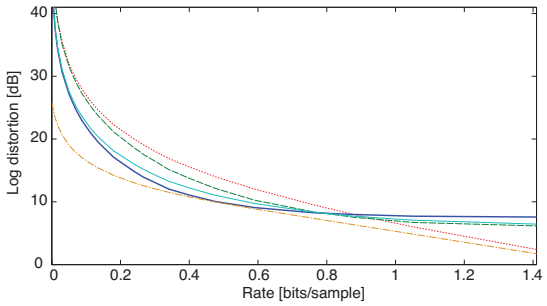

Fig. 5: Performance for low rates using pre- and post-filters from (11). No pre- or post-filter (dotted red), both pre- and post-filter (bold blue), only pre-filter (dashed green), only post-filter (thin cyan), and RD-optimal (dashed-dotted orange) from (2).

We then applied the paired-zero post-filtering concept to enhance coded speech. For each block of speech the pitch was estimated and a set of filters defined, each having the same pitch but with different cut-off frequencies for perodicity, c.f., Fig. 2. The filter yielding the lowest value of the criterion (with $\{\xi, \lambda, b\} = \{1.6, 0.3, 1.0\}$) in (13) was selected and utilized. Post-filtering was applied to speech coded with the G.722.1 codec at 16 kbps [10], the G.722.2 (AMR-WB) codec at 9 kbps and 16 kbps [11], and the iSAC codec at 16 kbps [12]. A small informal listening test was conducted where eight experienced listeners compared pairs of speech clips with and without post-filtering and indicated their preference. The speech material consisted of six female sentences and five male sentences. Results from the listening test and P (significance) values from t-tests are presented in Table 1, and it can clearly be seen that post-filtering improves the subjective quality.

Codec	Pref. w. PF	Pref. w/o. PF	P value
G.722.1 - 16 kbps	88%	12%	$4 \cdot 10^{-4}$
G.722.2 - 16 kbps	78%	22%	$3 \cdot 10^{-3}$
G.722.2 - 9 kbps	84%	16%	$4 \cdot 10^{-3}$
iSAC - 16 kbps	97%	3%	$7 \cdot 10^{-7}$

Table 1: Subjective performance of paired-zero post-filtering.

6. CONCLUSION

We introduced new refinements for pitch prediction in speech and audio coding. We found theoretically that post-filtering is more effective than pre-filtering. Our experiments confirm this result but show that the difference can be small. We proposed a methodology to select or design post-filters based on information available at the decoder only. We combined the selection method with a new paired-zero post-filter design for the low-rate regime. Our objective and subjective experiments confirmed that our post-filtering approach has significant practical benefits.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[2] J. Huang and P. Schultheiss, "Block quantization of correlated gaussian random variables," *IEEE Trans. Communication Systems*, vol. 11, no. 3, pp. 289–296, 1963.

[3] V. Ramamoorthy and N. Jayant, "Enhancement of ADPCM speech by adaptive postfiltering," *Bell Syst. Tech. J.*, vol. 63, no. 8, pp. 1465–1475, 1984.

[4] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 59–71, Jan. 1995.

[5] B. Edler and G. Schuller, "Audio coding using a psychoacoustic pre- and post-filter," in *Proc. IEEE Int Conf. Acoust. Speech Signal Process.*, vol. 2, 2000, pp. II881 –II884 vol.2.

[6] S. V. Andersen and W. B. Kleijn, "Reverse water-filling in predictive encoding of speech," in *IEEE Speech Coding Workshop, Porvoo*, 1999, pp. 105–107.

[7] R. Zamir, Y. Kochman, and U. Erex, "Achieving the Gaussian rate-distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, pp. 3354–3364, 2008.

[8] O. G. Guleryuz and M. T. Orchard, "On the DPCM compression of Gaussian autoregressive sequences," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 945–956, Mar. 2001.

[9] O. A. Moussa, M. Li, and W. B. Kleijn, "Predictive audio coding using rate-distortion-optimal pre- and post-filtering," in *Workshop Applications Signal Process Audio Acoustics (WASPAA)*, oct. 2011, pp. 213 –216.

[10] International Telecommunications Union, "ITU-T Recommendation g.722.1: Coding at 24 and 32 kbits/s for hands-free operation in systems with low frame loss," 1999.

[11] ——, "ITU-T Recommendation g.722.2: Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (AMR-WB)," 2002.

[12] "iSAC wide and superwideband codec." [Online]. Available: http://www.webrtc.org/