PITCH ENHANCEMENT MOTIVATED BY RATE-DISTORTION THEORY

Obada Alhaj Moussa[⋆] *Minyue Li*[†] *W. Bastiaan Kleijn*[‡⋆]

[⋆] Electrical Engineering, KTH - Royal Institute of Technology, Stockholm, Sweden
[†] Google Sweden AB, Kungsbron 2, Stockholm, Sweden
[‡] School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
alhaj@kth.se, minyue@google.com, bastiaan.kleijn@ecs.vuw.ac.nz

ABSTRACT

A pitch enhancement filter is designed with the objective to approach the optimal rate-distortion trade-off. The filter shows significant perceptual benefits, restating that information-theoretical and perceptual criteria are usually consistent. The filter is easy to implement and can be used as a complement to existing audio codecs. Our experiments show that it can improve the reconstruction quality of the AMR-WB standard.

Index Terms— pitch enhancement, rate-distortion theory, speech quality.

1. INTRODUCTION

Pitch enhancement techniques are widely used in speech communication as a means to achieve better speech quality [1, 2, 3, 4, 5, 6]. In general, pitch enhancement methods aim at increasing the periodicity of the speech signals or, equivalently, suppressing the gaps between the harmonics. In [7, 8, 9, 10], the Wiener or Kalman filter, are used to remove the noise between pitch harmonics. In [11, 2], a de-noising method based on estimating the noise spectrum from the so-called tunneling samples (spectral gaps between harmonics) is used. In [12, 13, 14], harmonic regeneration is used to preserve the speech harmonics, while suppressing the frequency bands with low energy.

Usually pitch enhancement methods are motivated from human perception. Some methods, e.g., [15], explicitly use masking properties of human auditory system to suppress noise. Others, e.g. [14, 12], ensure that the pitch structure is emphasized to make the speech more intelligible. However, perception-based design is not straightforward since it depends on subjective evaluation.

Our recent work [16] shows that some aspects of perceptual enhancement are consistent with information theoretical objectives. Specifically, it was concluded that, a post-filtering technique that is derived from rate-distortion theory has the same functionality but is more efficient than the conventional post-filtering techniques that are motivated from perception. The post-filtering technique was based on the spectral envelope of the signal. In this paper, we apply a post-filtering technique to the spectral fine structure of the signal. In [17], it is shown that a pitch enhancement filter can also be interpreted as a means to improve the coding efficiency and thus can be designed from an information theoretical viewpoint. [17] selects one pitch-filter from a pre-selected set of such filters according to a criterion. In this paper we compute the post-filter directly with a design that minimizes the impact of the filter adaptation. In contrast to [17] we test our results in a formal setting. The new methods provide good results in a listening test, and are easy to implement.

The remainder of this paper is organized as follows. Section 2 presents an information theoretical analysis of pitch enhancement. A new pitch enhancement filter based on rate-distortion theory is described in section 3. The evaluation of the proposed pitch enhancement filter is conducted in section 4. Finally, the conclusions are drawn in section 5.

2. INFORMATION THEORETICAL ANALYSIS OF PITCH ENHANCEMENT

Pitch enhancement is often achieved by filtering. Wiener filtering is a popular method, e.g. [7, 18]. For a stationary signal and an additive Gaussian noise, the minimum mean squared error (MSE) can be achieved using the Wiener filter

$$|H(\omega)| = 1 - \frac{N(\omega)}{S(\omega) + N(\omega)}, \tag{1}$$

where $S(\omega)$ and $N(\omega)$ are the power spectral density (PSD) of the source signal and the noise, respectively. It can be seen from (1) that the Wiener filtering suggests that pitch valleys should be attenuated. Specifically, when $S(\omega)$ is significantly smaller than $N(\omega)$, which happens in the harmonic valleys, the filter suppresses the corresponding frequency bands since $|H(\omega)| \approx 0$.

Naturally, if a signal is coded optimally then filtering cannot optimize performance further. The Wiener-filter assumption that the distortion is an additive noise is incorrect in that case. However, in practice filtering does have a role in quantization. Pre- and post-filtering can make dithered quantization rate-distortion optimal [19]. Dithered quantization leads to additive quantization noise and this has two major advantages. Firstly, the approach corresponds to the so-called forward test channel in rate-distortion theory, which generally leads to relatively straightforward analysis for Gaussian signals. This was exploited, for example, in an approach to rate-distortion optimal predictive coding that uses pre- and post-filtering and dithered quantization [20]. The second advantage is that the additive nature of the quantization noise reduces perceptable artifacts in a number of coding contexts.

If the quantization noise can be considered as additive, then the optimal pre- and post-filter satisfy [19]:

$$|H_i(\omega)| = \left(1 - \frac{\min\{S(\omega), N(\omega)\}}{S(\omega)}\right)^{\frac{1}{2}}, \; i \in 1, 2, \tag{2}$$

where $H_1(\omega)$ and $H_2(\omega)$ represent the frequency response of the pre- and post-filter, respectively. To achieve optimal performance for the pre- and post-filter (for Gaussian signals) the system must be completed with entropy coded dithered lattice quantization (ECDQ).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

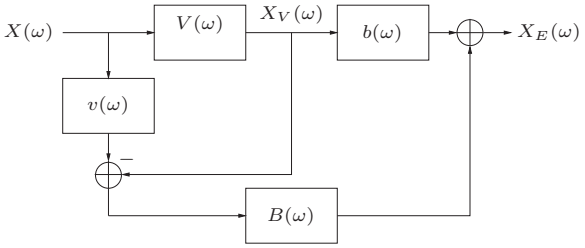

Fig. 2. Diagram of proposed pitch enhancement filter.

By observing (2), we notice that this filtering, like Wiener filtering but more explicitly, tends to remove the harmonic valleys. Whenever $S(\omega)$ is less that $N(\omega)$, both the pre- and post-filter completely remove the corresponding frequency bands. This happens to the spectral valleys of a periodic signal.

The pre- and post-filter play slightly different roles. The pre-filter reduces valleys so that they will be assigned a low or zero rate by the quantizer. Thus, its function is to optimize the rate distribution across frequency. In contrast, the post-filter removes the quantization noise that has filled in the same valleys after the signal has passed through the quantizer. Since the valleys comprise a small portion of the total power of the source, the pre-filter causes only mild change to the source signal and the corresponding rate distribution. It may, therefore, be removed without significantly effecting the rate-distortion performance [17]. This explains the fact that pitch enhancement as a pre-processing technique is not common in practical audio coding systems. We should also note that if the pre-filter is omitted, then the rate-distortion optimal post-filter is, in fact, the Wiener filter. This is of little consequence for practical pitch postfiltering as we must approximate the post-filter to facilitate implementation.

Most of the discussion in this section assumes that a dithered quantizer is used and that the signal is Gaussian. For systems that use conventional quantizers (which correspond to the backward test channel), such as existing audio-coding standards, our analysis does not formally apply. However, many systems effectively do have a noise floor and the same methods can be used to increase performance.

3. PITCH ENHANCEMENT FILTER

From the above analysis. we know that the main functionality of a pitch enhancement filter is to remove the valleys in spectral fine structure, except for those bands that have high energy. It is generally difficult to design a filter that differentiates spectral valleys at different frequencies. Most pitch enhancement filters treat the valleys equally, see, e.g., [21]. The filter proposed in [17] solves the problem by separating frequency bands. In this paper, we provide an alternative design of the pitch enhancement filter. The frequency-selective adaptation operates only on a low-power signal consisting of the spectral valleys. This design minimizes the audibility of rapid adaptation of this filter.

The basic idea is to remove all the valleys, and then add the necessary valleys back, which is illustrated in Fig. 2. $V(\omega)$ is a filter that suppresses the harmonic valleys. That is, $X_V(\omega) = V(\omega)X(\omega)$ can be interpreted as a valley-removed version of the input. $B(z)$ is a multi-band-pass filter that passes through the bands where the valleys are to be added back. Note that this filter operates on the low-power signal $X(\omega) - X_V(\omega)$ that consists of valleys only. Thus, the impact of its adaptation is minimized. The two remaining filters, $v(\omega)$ and $b(\omega)$, are introduced to compensate for the delay of $V(\omega)$ and $B(\omega)$, respectively. Specifically, they satisfy $v(\omega) = \mathrm{e}^{\mathrm{j}\angle V(\omega)}$ and $b(\omega) = \mathrm{e}^{\mathrm{j}\angle B(\omega)}$, which can be implemented as simple delay lines when $V(\omega)$ and $B(\omega)$ are linear-phase filters.

The final output of the pitch enhancement filter is

$$\begin{aligned} X_E(\omega) =& X_V(\omega)b(\omega) + B(\omega)\left(X(\omega)v(\omega) - X_V(\omega)\right) \quad (3)\\ =& \left(|V(\omega)| + |B(\omega)| - |V(\omega)||B(\omega)|\right) \times \\ & X(\omega)\mathrm{e}^{\mathrm{j}(\angle V(\omega) + \angle B(\omega))}. \quad (4)\end{aligned}$$

We see that the proposed pitch enhancement filter has the desired behavior:

- It has a unit gain on the pitch peaks, i.e., $|X_E(\omega)| = |X(\omega)|$ for ω that satisfies $|V(\omega)| = 1$;
- It maintains the spectrum of the input signal at the pass-bands of $B(\omega)$, i.e., $|X_E(\omega)| = |X(\omega)|$ for ω that satisfies $|B(\omega)| = 1$;
- The valley removing filter is active at the stop-bands of $B(\omega)$, i.e., $|X_E(\omega)| = |V(\omega)X(\omega)|$ for ω that satisfies $|B(\omega)| = 0$;
- It has a linear phase response, if $V(\omega)$ and $B(\omega)$ are linear-phase filters.

For the valley-removing filter, we choose the following linear-phase filter:

$$V(\omega) = \frac{1}{4}\left(1 + \mathrm{e}^{-T\mathrm{j}\omega}\right)^2, \quad (5)$$

where the factor of $\frac{1}{4}$ is to achieve a unity gain on the pitch peaks and T is the pitch period. Other filters like those suggested in [21, 17] can also be used for this purpose. Those filters use more parameters to control the shape of the magnitude response more accurately, but require adjustments on the parameters. We choose this simple valley-suppression filter to avoid the requirement of parameter tuning.

Next we have to determine the multi-band-pass filter $B(\omega)$. We select this filter so that the filter of (5) suppresses the valleys where the noise power exceeds the signal amplitude. In practice we determine $B(\omega)$ by determining where the *lower contour* of the power spectrum of the input signal is below the noise spectrum. Specifically, the ideal amplitude response of $B(\omega)$ is

$$|B(\omega)| = \begin{cases} 1 & S_L(\omega) \geq N(\omega) \\ 0 & S_L(\omega) < N(\omega) \end{cases}, \quad (6)$$

where $S_L(\omega)$ denotes the lower contour of the source power spectrum. Fig. 3 illustrates the relationship among a power spectrum, its lower contour, a noise spectrum, and the ideal magnitude response of the corresponding multi-band-pass filter. The estimation of $N(\omega)$ in a practical system will be described in section 4.

4. EVALUATION

We used two approaches to assess the performance of the proposed pitch enhancement filter. On the one hand we designed a complete testbed that uses closed-loop prediction and dithered quantization. We tested this system on both synthetic data and real audio data. For the testbed design the presented post-filter design approached optimality. On the other hand we applied our post-filter to a standardized predictive coder that uses a conventional quantizer and tested this system on real audio data.

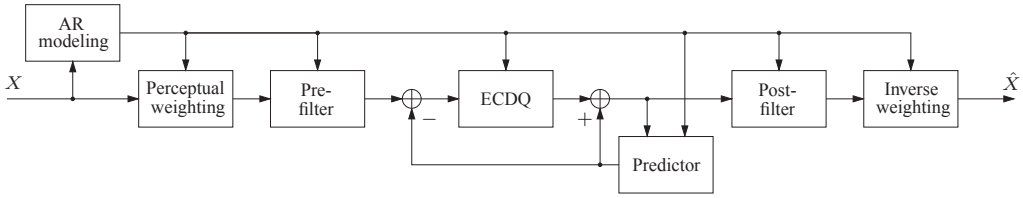

Fig. 1. Audio coder based on pre-/post-filtered DPCM.

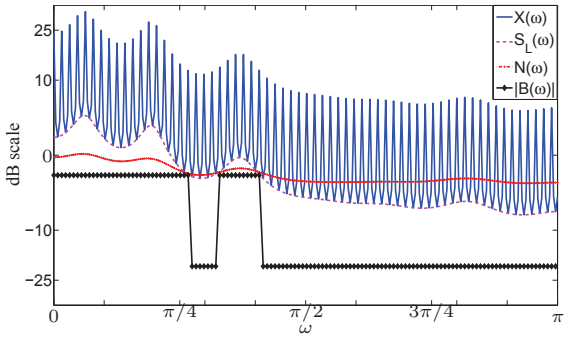

Fig. 3. An example of the desired multi-band-pass filter (The values of $|B(\omega)|$ are not read from the y-axis; They are either 0 or 1).

4.1. Testbed Design

We designed a complete speech/audio coder based on the pre-/post-filtered DPCM proposed in [20]. In [20], the scheme is presented as coding structure that achieves the RDF. To build a practical system, we added a modeling block and a perceptual weighting block to the basic design. The coder is illustrated in Fig. 1, where the proposed pitch enhancer is used in the pre- and post-filtering blocks. In addition to the pitch enhancement filter, the pre- and post-filter also include the aforementioned spectral envelope filter [16], which realizes the varying gain for the harmonic peaks.

The modeling block obtains an autoregressive (AR) model of the input signal:

$$\frac{1}{A(z)} = \underbrace{\frac{\alpha}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}}}_{\frac{1}{A_S(z)}} \times \underbrace{\frac{1}{1 - \beta z^{-T}}}_{\frac{1}{A_L(z)}}, \tag{7}$$

where p is the model order, $\{a_i\}_{i=1}^{p}$ is a set of AR coefficients, α is a gain factor, and β is the pitch gain. A short-term linear prediction coefficient (LPC) analysis and pitch estimation are used to obtain these parameters.

The model in (7) provides an approximate power spectrum of the signal. It also defines the lower contour of the power spectrum, which can be used for deriving the multi-band-pass filter $B(\omega)$ in our proposed pitch enhancer. In particular, the spectral lower contour is

$$S_L(\omega) = \frac{1}{(1+\beta)^2 |A_S(\omega)|^2}. \tag{8}$$

By complementing the DPCM scheme of [20] with perceptual weighting it becomes optimal for a weighted MSE, thus accounting for the relevant properties of the human auditory system. Perceptual weighting filters are generally derived from auditory modeling that fits data from auditory masking experiments. Here, we only consider spectral masking. Specifically, the perceptual weighting filter in our system follows a widely used setup, known as the γ_1, γ_2 model [1]:

$$W(z) = \frac{A_S(z/\gamma_1)}{A_S(z/\gamma_2)}. \tag{9}$$

The perceptual weighting filter defines the noise spectrum.

$$N(\omega) = \theta^2 |W(\omega)|^2 \tag{10}$$

where θ^2 is the quantization noise variance, which can be evaluated given the bit-rate. In our testbed, assuming a uniform scalar quantizer, the noise level θ is approximated by the mean distortion measure $\frac{\triangle^2}{12}$, where $\triangle$ is the quantization step size.

We note that the spectral envelope filter, which is defined by an envelope of the signal spectrum in [16], should be derived from the upper contour of the signal spectrum, in the presence of pitch. Based on the AR model (7), the upper contour can be obtained by

$$S_U(\omega) = \frac{1}{(1-\beta)^2 |A_S(\omega)|^2}. \tag{11}$$

We use this test-bed in both an objective and a subjective test. The objective test is to evaluate the efficiency of the proposed filter in achieving the rate-distortion optimality, and the subjective test is to show its perceptual benefits. We discuss these two tests in the following subsections.

4.2. Objective evaluation

In this section we perform an experiment to show that the proposed pitch enhancement method facilitates a coder to approach the theoretical rate-distortion optimality, and to show that it is sufficiently effective when applied only as a post-filter.

To clarify the roles of the pre- and post-filter experimentally, we evaluate four different versions of the coder in Fig. 1. In the first scenario, we deactivate both the pre- and post-filter. In the second scenario, we deactivate the pre-filter only. In the third scenario, we use the complete coder and in the fourth scenario, we use the complete coder without the pitch enhancement filter, leaving only the spectral envelope filter in the pre- and post-filter. The input signal is a stationary Gaussian process with the PSD depicted as the signal spectrum in Fig. 3. In all four systems, the AR modeling is performed once over the whole data. The trade-off between bit-rate and the perceptual MSE is shown in Fig. 4, which also includes the RDF of the source process.

The results show that the proposed pitch filter is efficient in achieving the RDF. There is a gap of about 0.254 bits/sample from the RDF at high bit-rates, which is due to the use of scalar quantization. The gap vanishes as the bit-rate approaches zero. By comparing the curves from the first scenario to the curves from the second and third scenarios, we see that the role played by the pre- and post-filter in achieving the RDF increases with decreasing bit-rate. This

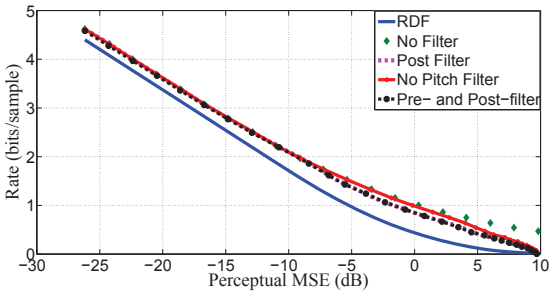

Fig. 4. Rate-distortion performance on the generated Gaussian process.

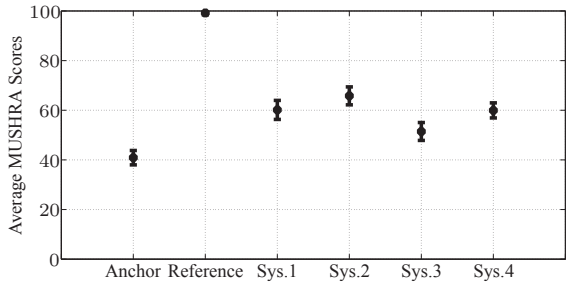

Fig. 5. Mean and 95% confidential interval of the MUSHRA test. Sys 1, 2, 3 and 4 represent the AMR-WB, the extended AMR-WB, the "pitch model free" R-D optimal speech/audio coder, and the proposed R-D optimal audio coder, respectively.

is expected as the quantization noise decreases and becomes small compared to the signal and the transfer function (2) of the pre- and post-filter converges to being flat. It can also be seen that using only a post-filter is almost as efficient as using both the pre- and post-filter. By comparing the third and fourth scenarios, we can see that the pitch enhancement filter alone does yield a coding gain, especially at medium bit-rates.

4.3. Subjective evaluation

In this section, we describe the subjective experiment we performed to show that the proposed pitch enhancement filter, designed to improve coding efficiency, provides perceptual benefits in a practical setting. We assessed the perceptual performance of our complete RD-optimal speech/audio testbed which is depicted in Fig. 1. We also performed experiments that show that the proposed method can be used as a complement to different speech/audio codecs. For this purpose, we extended a state-of-the-art coder, the AMR-WB [22] with the same post-filter as in the testbed.

In the testbed-based coder, the AR model was updated every 16 ms using the same windowing and interpolation as in the AMR-WB codec [22]. The audio frame was divided into 4 subframes of 4 ms each. The short-term model parameters were transmitted every 16 ms, while the pitch parameters were transmitted every 4 ms. In a practical audio coder, the pre- and post-filter can be sensitive to model variations [16]. When the power of a band is near the quantization noise level, its signal power may be removed in one frame and not be removed in the next frame, creating audible artifacts. This can happen, and is particularly audible, when the original signal is relatively steady. Therefore, in almost-steady-state situations, the pre- and post-filter should not significantly change from one signal block to the next. To ensure that, we applied a smoothing technique described in [16].

In the system based on AMR-WB, to obtain the multi-band-pass filter $B(\omega)$, we use the result of the AR modeling and the perceptual weigthing of the AMR-WB. The noise level θ is approximated using the Shannon lower bound, i.e, the minimum achievable distortion is determined by the RDF given the source PSD and the bit-rate.

A formal MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test was conducted. The test included four coders: AMR-WB coder (Sys.1), AMR-WB extended with the proposed post-filter (Sys.2), the R-D optimal testbed shown in Fig. 1 (Sys.4), and version of the testbed without the pitch post-filter (note that this includes the envelope post-filter), which is similar to the coder described in our previous work [16] (Sys.3). The test database consists of the 12 audio excerpts from the widely used MPEG database. It contains four speech excerpts, three music excerpts, three speech and music excerpts and two speech and noise excerpts. Each excerpt was down-sampled to 16 kHz and coded at an average bit-rate of 24 kbps. The anchors were 3.5 KHz low-passed version of the reference signals. Eleven listeners participated in the listening test.

The results depicted in Fig. 5 show that the post-filter (including the pitch post-filter), which was designed to improve coding efficiency, also provides a better perceptual quality. This can be seen by comparing the average MUSHRA scores of the AMR-WB coder and its extended version (Sys.1 vs Sys.2), which confirms that the proposed method can be used as a complement to different audio codecs. It can also be seen, by comparing the average MUSHRA scores of Sys.3 and Sys.4, that the proposed pitch post-filter provides a significant perceptual improvement. We note that, among the four different types of signals, the music excerpts have the most subjective improvements. It is worth mentioning that our testbed speech/audio codec described in Fig. 1 provides a comparable performance to a state-of-the-art coder, the AMR-WB.

5. CONCLUSIONS

We have shown that pitch post-filtering can often significantly increase the quality of audio coders. While post-filters commonly are seen as a tool to enhance perceived performance, we showed that the improvement is seen both in terms of objective rate-distortion performance and subjective listening experiments. In a more general sense this reconfirms earlier results [23] that adaptive coding schemes that can handle various bit-rates can be based on simple objective criteria. Although the structure of our post-filter was motivated by the analysis of systems that use dithered quantizers, the observed subjective enhancement is similar when it was applied to the AMR-WB coder, which uses a predictive coding structure with a conventional quantizer.

6. REFERENCES

[1] 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.

[2] E. Douglas, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. EUROSPEECH*, 2001, vol. 1, pp. 437–450.

[3] M. Cordovilla, J. A. Ning Ma, V. Sanchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2011, pp. 4808–4811.

[4] B. Luis, J. Droppo, and A. Acero, "Speech enhancement using a pitch predictive model," in *IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2008, pp. 4885–4888.

[5] Q. Yan, S. Vaseghi, E. Zavarehei, and B. Milner, "Formant-tracking linear prediction model for speech processing in noisy environment," in *Proc. EUROSPEECH*, 2005.

[6] Q. Yan, S. Vaseghi, E. Zavarehei, and B. Milner, "Kalman filter with linear predictor and harmonic noise models for noisy speech enhancement," in *Proc. European Signal Process. Conf.*, 2006.

[7] V. Sunnydayal and T. Kishore Kumar, "Speech enhancement using sub-band Wiener filter with pitch synchronous analysis," in *Int. Conf. Adv. Comput., Comm., Inform.*, 2013, pp. 20–25.

[8] M. A. Ramalho and R. J. Mammone, "New speech enhancement techniques using the pitch mode modulation model," in *Midwest Symp. Circuits Systems*, 1993, vol. 2, pp. 1531–1534.

[9] C. Ruofei, F. Cheung-Fat, and S. C. Hing, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Trans. Audio, Speech, lang. process.*, vol. 20, no. 4, pp. 1324–1336, 2012.

[10] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," in *IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2001, pp. 125–128.

[11] L. Buera, J. Droppo, and A. Acero, "Speech enhancement using a pitch predictive model," in *IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2008, pp. 4885–4888.

[12] C. Palpous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *IEEE Int. Conf. Acoustic, Speech, Signal Process.*, 2005, pp. 157–160.

[13] X. Hou and X. Zhu, "Speech enhancement using harmonic regeneration," in *IEEE Int. Conf. Comp. Science, Autom. Eng.*, 2011, pp. 150–152.

[14] T. W. Shen, D.P. Lun, and C. Hsung T, "Speech enhancement using harmonic regeneration with improved wavelet based a-priori signal to noise ratio estimator," in *Int. Symp. Intel. Signal Process. Comm. Systems*, 2010, pp. 1–4.

[15] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources.," *Speech Communication*, vol. 49, pp. 874–891, 2007.

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

[17] W. B. Kleijn and J. Skoglund, "Improved prediction of nearly-periodic signals," in *Int. Workshop Acoustic Signal Enhancement*, 2012, pp. 1–4.

[18] A.R. Verma, R.K. Singh, A. Kumar, and K. Ranjeet, "An improved method for speech enhancement based on 2d-dwt using hybrid weiner filtering," in *IEEE Int. Conf. on Computational Intelligence and Computing Research*, 2012, pp. 1–6.

[19] R. Zamir and M. Feder, "Information rates of pre/post-filtered dithered quantizers," *Information Theory, IEEE Transactions on*, vol. 42, no. 5, pp. 1340 –1353, sep 1996.

[20] 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.

[21] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, 2001.

[22] I. Varga, R. D. D. Lacovo, and P. Usai, "Standardization of the AMR wideband speech codec in 3GPP and ITU-T," *IEEE Communication Magazine*, vol. 44, pp. 66–73, 2006.

[23] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *Asilomar Conf. Signals, Systems, Computers*, 2007, pp. 535–539.