Interspeech 2018
2-6 September 2018, Hyderabad

Postfiltering with Complex Spectral Correlations for Speech and Audio Coding

Sneha Das, Tom Bäckström

Department of Signal Processing and Acoustics, Aalto University, Finland
sneha.das@aalto.fi, tom.backstrom@aalto.fi

Abstract

State-of-the-art speech codecs achieve a good compromise between quality, bitrate and complexity. However, retaining performance outside the target bitrate range remains challenging. To improve performance, many codecs use pre- and post-filtering techniques to reduce the perceptual effect of quantization-noise. In this paper, we propose a postfiltering method to attenuate quantization noise which uses the complex spectral correlations of speech signals. Since conventional speech codecs cannot transmit information with temporal dependencies as transmission errors could result in severe error propagation, we model the correlation offline and employ them at the decoder, hence removing the need to transmit any side information. Objective evaluation indicates an average 4 dB improvement in the perceptual SNR of signals using the context-based post-filter, with respect to the noisy signal, and an average 2 dB improvement relative to the conventional Wiener filter. These results are confirmed by an improvement of up to 30 MUSHRA points in a subjective listening test.

Index Terms: speech and audio coding, noise reduction, temporal correlation, post-filtering

1. Introduction

Speech coding, the process of compressing speech signals for efficient transmission and storage, is an essential component in speech processing technologies. It is employed in almost all devices involved in the transmission, storage or rendering of speech signals. While standard speech codecs achieve transparent performance around target bitrates, the performance of codecs suffer in terms of efficiency and complexity outside the target bitrate range [1].

Specifically at lower bitrates the degradation in performance is because large parts of the signal are quantized to zero, yielding a sparse signal which frequently toggles between zero and non-zero. This gives a distorted quality to the signal, which is perceptually characterized as musical noise. Modern codecs like EVS, USAC [2, 3] reduce the effect of quantization noise by implementing postprocessing methods [1, 4]. Many of these methods have to be implemented both at the encoder and decoder, hence requiring changes to the core structure of the codec, and sometimes also the transmission of additional side information. Moreover, most of these methods focus on alleviating the effect of distortions rather than the cause for distortions.

The noise reduction techniques widely adopted in speech processing are often employed as pre-filters to reduce background noise in speech coding. However, application of these methods for the attenuation of quantization noise have not been fully explored yet. The reasons for this are (i) information from zero-quantized bins cannot be restored by using conventional filtering techniques alone, and (ii) quantization noise is highly correlated to speech at low bitrates, thus discriminating between speech and quantization-noise distributions for noise reduction is difficult; these are further discussed in Sec. 2.

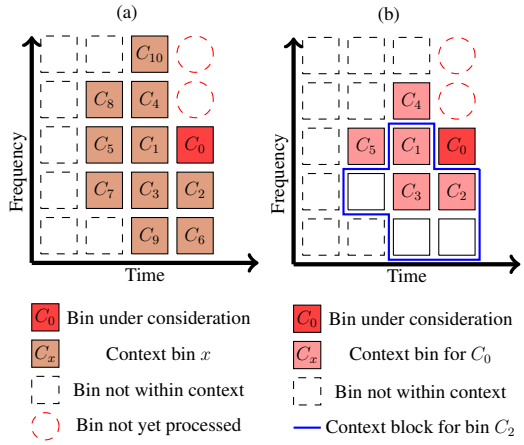

Figure 1: *(a) Context block of size, $L = 10$ (b) Recurrent context-block of the context bin C_2.*

Fundamentally, speech is a slowly varying signal, whereby it has a high temporal correlation [5]. Recently, MVDR and Wiener filters using the intrinsic temporal and frequency correlation in speech were proposed and showed significant noise reduction potential [6, 5, 7]. However, speech codecs refrain from transmitting information with such temporal dependency to avoid error propagation as a consequence of information loss. Therefore, application of speech correlation for speech coding or the attenuation of quantization noise has not been sufficiently studied, until recently; an accompanying paper [8] presents the advantages of incorporating the correlations in the speech magnitude spectrum for quantization noise reduction.

The contributions of this work are as follows: (i) modeling the complex speech spectrum to incorporate the contextual information intrinsic in speech, (ii) formulating the problem such that the models are independent of the large fluctuations in speech signals and the correlation recurrence between samples enables us to incorporate much larger contextual information, (iii) obtaining an analytical solution such that the filter is optimal in minimum mean square error sense. We begin by examining the possibility of applying conventional noise reduction techniques for the attenuation of quantization noise, and then model the complex speech spectrum and use it at the decoder to estimate speech from an observation of the corrupted signal. This approach removes the need for the transmission of any additional side information.

2. Modeling and Methodology

At low bitrates conventional entropy coding methods yield a sparse signal, which often causes a perceptual artifact known as musical noise. Information from such spectral holes cannot be recovered by conventional approaches like Wiener filter-

10.21437/Interspeech.2018-1026

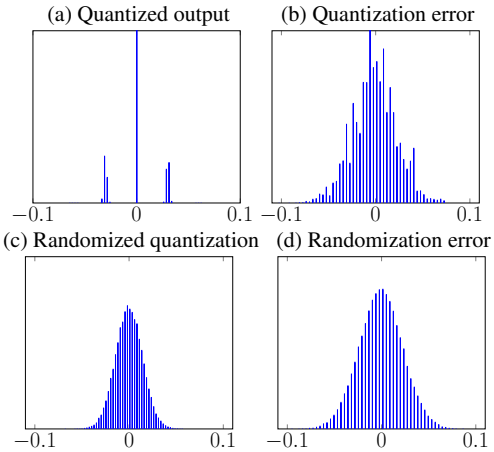

Figure 2: *Histograms of (a) Conventional quantized output (b) Quantization error (c) Quantized output using randomization (d) Quantization error using randomization. The input was a an uncorrelated Gaussian distributed signal.*

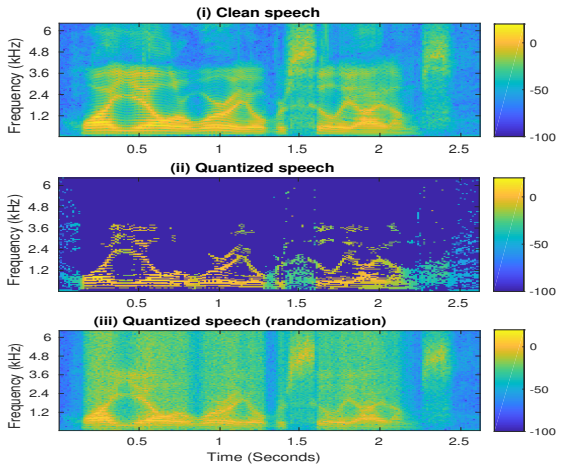

Figure 3: *Spectrograms of (i) true speech (ii) quantized speech and, (iii) speech quantized after randomization.*

ing, because they mostly modify the gain. Moreover, common noise reduction techniques used in speech processing model the speech and noise characteristics and perform reduction by discriminating between them. However, at low bitrates quantization noise is highly correlated with the underlying speech signal, hence making it difficult to discriminate between them. Figs. 2 - 3 illustrate these problems; Fig. 2 (a) shows the distribution of the decoded signal, which is extremely sparse, and (b) shows the distribution of the quantization noise, for a white Gaussian input sequence. Fig. 3 a & b depict the spectrogram of the true speech and the decoded speech simulated at a low bitrate, respectively.

To mitigate these problems, we can apply randomization before encoding the signal [9, 10, 11]. Randomization is a type of dithering [12] which has been previously used in speech codecs [13] to improve perceptual signal quality, and recent works [14, 11] enable us to apply randomization without increase in bitrate. The effect of applying randomization in coding is demonstrated in Fig. 2 c & d and Fig. 3 c; the illustrations clearly show that randomization preserves the decoded speech distribution and prevents signal sparsity. Additionally, it also lends the quantization noise a more uncorrelated characteristic, thus enabling the application of common noise reduction techniques from speech processing literature [15].

Due to dithering, we can assume that the quantization noise is an additive and uncorrelated normally distributed process,

$$Y_{k,t} = X_{k,t} + V_{k,t}, \tag{1}$$

where Y, X and V are the complex-valued short-time frequency domain values of the noisy, clean-speech and noise signals, respectively. k denotes the frequency bin in the time-frame t. In addition, we assume that X and V are zero-mean Gaussian random variables. Our objective is to estimate $X_{k,t}$ from an observation $Y_{k,t}$ as well as using previously estimated samples of $\hat{\mathbf{x}}_c$. We call $\hat{\mathbf{x}}_c$ the *context* of $X_{k,t}$

The estimate of the clean speech signal, $\hat{\mathbf{x}}$, known as the Wiener filter [15], is defined as:

$$\hat{\mathbf{x}} = \mathbf{\Lambda_X}(\mathbf{\Lambda_X} + \mathbf{\Lambda_N})^{-1}\mathbf{y}, \tag{2}$$

where $\mathbf{\Lambda_X}, \mathbf{\Lambda_N} \in \mathbb{C}^{(c+1)\times(c+1)}$ are the speech and noise covariance matrices, respectively, and $\mathbf{y} \in \mathbb{C}^{c+1}$ is the noisy observation vector with $c+1$ dimensions, c being the context length. The covariances in Eq. 2 represent the correlation between time-frequency bins, which we call the context neighborhood. The covariance matrices are trained off-line from a database of speech signals. Information regarding the noise characteristics is also incorporated in the process, by modeling the target noise-type (quantization noise), similar to the speech signals. Since we know the design of the encoder, we know exactly the quantization characteristics, hence it is a straightforward task to construct the noise covariance $\mathbf{\Lambda_N}$.

Context neighborhood: An example of the context neighborhood of size 10 is presented in Fig 1 a. In the figure, the block C_0 represents the frequency bin under consideration. Blocks C_i, $i \in \{1, 2, .., 10\}$ are the frequency bins considered in the immediate neighborhood. In this particular example, the context bins span the current time-frame and two previous time-frames, and two lower and upper frequency-bins. The context neighborhood includes only those frequency bins in which the clean speech has already been estimated. The structuring of the context neighborhood here is similar to the coding application, wherein contextual information is used to improve the efficiency of entropy coding [16]. In addition to incorporating information from the immediate context neighborhood, the context neighborhood of the bins in the context block are also integrated in the filtering process, resulting in the utilization of a larger context information, similar to IIR filtering. This is depicted in Fig 1 b, where the blue line depicts the context block of the context bin C_2. The mathematical formulation of the neighborhood is elaborated in the following section.

Normalized covariance and gain modeling: Speech signals have large fluctuations in gain and spectral envelope structure. To model the spectral fine structure efficiently [17], we use normalization to remove the effect of this fluctuation. The gain is computed during noise attenuation from the Wiener gain in the current bin and the estimates in the previous frequency bins. The normalized covariance and the estimated gain are employed together to obtain the estimate of the current frequency sample. This step is important as it enables us to use the actual speech statistics for noise reduction despite the large fluctuations.

Define the context vector as $\mathbf{u}_{k,t} = \begin{bmatrix} X_{k,t} & X_{k-1,t-1} & \cdots\cdots & X_{k-c,t-c} \end{bmatrix}$, thus the normalized context vector is $\mathbf{z}_{k,t} = \mathbf{u}_{k,t}/\|\mathbf{u}_{k,t}\|$. The speech

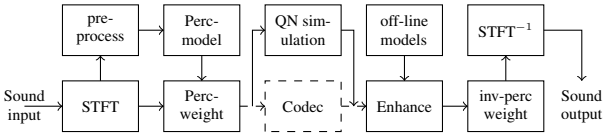

Figure 4: *Block diagram of the proposed system including simulation of the codec for testing purposes.*

covariance is defined as $\hat{\mathbf{\Lambda}}_{\mathbf{X}} = \gamma \mathbf{\Lambda}_{\mathbf{X}}$, where $\mathbf{\Lambda}_{\mathbf{X}}$ is the normalized covariance and γ represents the gain. The gain is computed as $\gamma = \hat{\mathbf{z}}_{k,t}\hat{\mathbf{z}}_{k,t}^H$ and the normalized covariances are calculated from the speech dataset as follows:

$$\mathbf{\Lambda}_{\mathbf{X}} = E\{\mathbf{Z}\mathbf{Z}^H\} = E\left\{ \begin{bmatrix} \mathbf{z}_{k,t} \\ \mathbf{z}_{k-1,t-1} \\ \cdots \\ \mathbf{z}_{k-c,t-c} \end{bmatrix} \begin{bmatrix} \mathbf{z}_{k,t} \\ \mathbf{z}_{k-1,t-1} \\ \cdots \\ \mathbf{z}_{k-c,t-c} \end{bmatrix}^H \right\}, \quad (3)$$

From Eq. 3, we observe that this approach enables us to incorporate correlation from a neighborhood much larger than the context size and more information, consequently saving computational resources. The noise statistics is computed as follows:

$$\mathbf{\Lambda}_{\mathbf{N}} = E\{\mathbf{W}\mathbf{W}^H\},$$
$$\mathbf{W} = \begin{bmatrix} N_{k,t} & \cdots & N_{k-c,t-c} \\ N_{k-1,t-1} & \cdots & N_{k-1-c,t-1-c} \\ \cdots & \cdots & \cdots \\ N_{k-c,t-c} & \cdots & N_{k-2c,t-2c} \end{bmatrix}. \quad (4)$$

Note that, in Eq. 4, normalization is not necessary for the noise models. Finally, the equation for the estimated clean speech signal is:

$$\hat{\mathbf{x}} = \gamma \mathbf{\Lambda}_{\mathbf{X}}[(\gamma \mathbf{\Lambda}_{\mathbf{X}}) + \mathbf{\Lambda}_{\mathbf{N}}]^{-1}\mathbf{y} \quad (5)$$

Owing to the formulation, the complexity of the method is linearly proportional to the context size. The proposed method differs from the 2D Wiener filtering in [18], in that it operates using the complex magnitude spectrum, whereby there is no need to use the noisy phase to reconstruct the signal unlike conventional methods. Additionally, in contrast to 1D and 2D Wiener filters which apply a scaler gain to the noisy magnitude spectrum, the proposed filter incorporates information from the previous estimates to compute the vector gain. Therefore, with respect to previous work the novelty of this method lies in the way the contextual information is incorporated in the filter, thus making the system adaptive to the variations in speech signal.

3. Experiments and Results

The proposed method was evaluated using both objective and subjective tests. We used the perceptual SNR (pSNR) [2, 1] as the objective measure, because it approximates human perception and it is already available in a typical speech codec. For subjective evaluation, we conducted a MUSHRA listening test.

3.1. System overview

The system structure is illustrated in Fig. 4 and is similar to the TCX mode in 3GPP EVS [2]. First, we apply STFT to the incoming signal to transform it to the frequency domain. We use here the STFT instead of the standard MDCT to make sure that our results are readily transferable to speech enhancement applications. Informal experiments verify that the choice of transform does not introduce any unexpected problems in the results [15, 1].

To ensure that the coding noise has least perceptual effect, the frequency domain signal is perceptually weighted. We compute the perceptual model, which is used in the EVS codec [2], based on the linear prediction coefficients (LPC). After weighting the signal with the perceptual envelope, it is normalized and entropy coded. For straightforward reproducibility, we simulated quantization noise by perceptually weighted Gaussian noise, following the discussion in Sec. 2. Thus, the output of the codec/quantization noise (QN) simulation block, in Fig. 4, is the corrupted decoded signal. The proposed filtering method is applied at this stage. The enhancement block acquires the off-line trained speech and noise models. Following the noise reduction process, the signal is weighted by the inverse perceptual envelope and then transformed back to the time domain to obtain the enhanced, decoded speech signal.

3.2. Objective evaluation

Experimental setup: The process is divided into training and testing phases. In the training phase, we estimate the static normalized speech covariances for context sizes $L \in \{1, 2..14\}$ from the speech data. For training, we chose 50 random samples from the training set of the TIMIT database [19]. All signals are resampled to 12.8 kHz, and a sine window is applied on frames of size 20 ms with 50% overlap. The windowed signals are then transformed to the frequency domain. Since the enhancement is applied in the perceptual domain, we also model the speech in the perceptual domain. For each bin sample in the perceptual domain, the context neighborhoods are composed into matrices, as described in section 2, and the covariances are computed. We similarly obtain the noise models using perceptually weighted Gaussian noise.

For testing, 105 speech samples are randomly selected from the database. The noisy samples are generated as the additive sum of the speech and the simulated noise. The levels of speech and noise are controlled such that we test the method for pSNR ranging from 0-20 dB with 5 samples for each pSNR level, to conform to the typical operating range of codecs. For each sample, 14 context sizes were tested. For reference, the noisy samples were enhanced using an oracle filter, wherein the conventional Wiener filter employs the true noise as the noise estimate, i.e., the optimal Wiener gain is known.

Evaluation results: The results are depicted in Fig. 5. The output pSNR of the conventional Wiener filter, the oracle filter, and noise attenuation using filters of context length $L = \{1, 14\}$ are illustrated in Fig. 5 a. In Fig. 5 b, the differential output pSNR, which is the improvement in the output pSNR with respect to the pSNR of the signal corrupted by quantization noise, is plotted over a range of input pSNR for the different filtering approaches. These plots demonstrate that the conventional Wiener filter significantly improves the noisy signal, with 3 dB improvement at lower pSNRs and 1 dB improvement at higher pSNRs. Additionally, the contextual filter $L = 14$ shows 6 dB improvement at higher pSNRs and around 2 dB improvement at a lower pSNR.

Fig. 5 c demonstrates the effect of context size at different input pSNRs. It can be observed that at lower pSNRs the context size has significant impact on noise attenuation; the improvement in pSNR increases with increase in context size. However, the rate of improvement with respect to context size decreases as the context size increases, and tends towards saturation for $L > 10$. At higher input pSNRs, the improvement reaches saturation at relatively smaller context size.

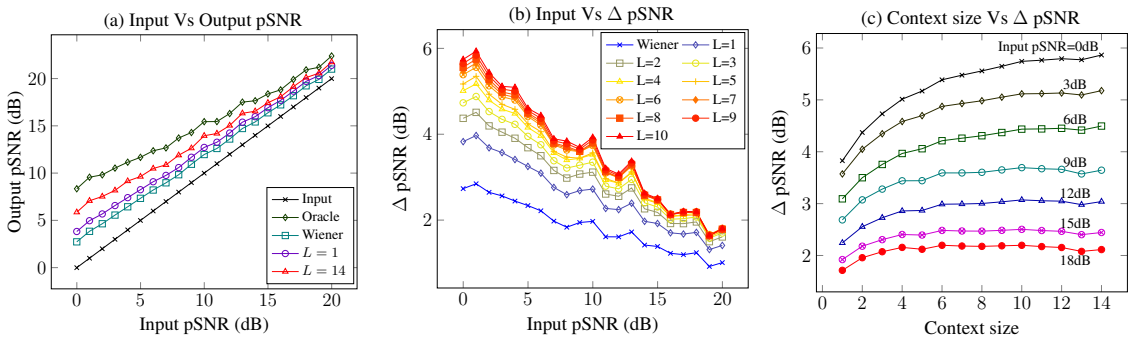

Figure 5: *Plots showing (a) the pSNR and (b) pSNR improvement after postfiltering, and (c) pSNR improvement for different contexts.*

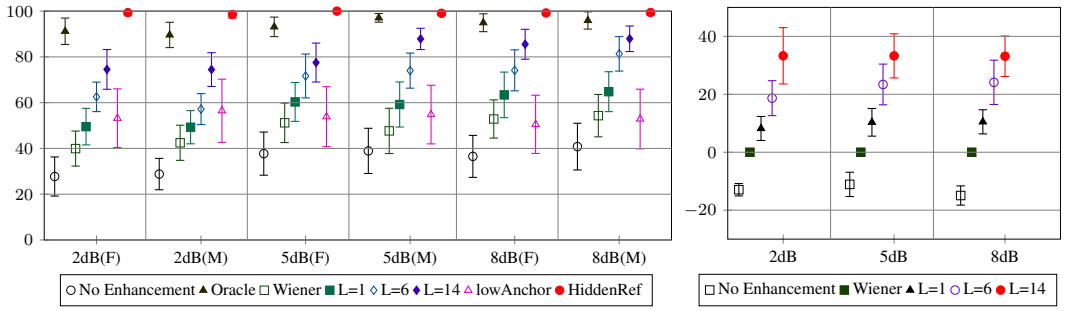

Figure 6: *MUSHRA listening test results a) Scores for all items over all the conditions b) Difference scores for each input pSNR condition averaged over male and female. Oracle, lower anchor and hidden reference scores have been omitted for clarity.*

3.3. Subjective evaluation

We evaluated the quality of the proposed method with a subjective MUSHRA listening test [20]. The test comprised of six items and each item consisted of 8 test conditions. Listeners, both experts and non-experts, between the age 20 to 43 participated. However, only the ratings of those participants who scored the hidden reference greater than 90 MUSHRA points were selected, resulting in 15 listeners whose scores were included for this evaluation.

Six sentences were randomly chosen from the TIMIT database to generate the test items. The items were generated by adding perceptual noise, to simulate coding noise, such that the resulting signals' pSNR were fixed at 2, 5 and 8 dB. For each pSNR, one male and one female item was generated. Each item consisted of 8 conditions: Noisy (no enhancement), ideal enhancement with the noise known (oracle), conventional Wiener filter, samples from the proposed method with context sizes one (L=1), six (L=6), fourteen (L=14), in addition to the 3.5kHz low-pass signal as the lower anchor and the hidden reference, as per the MUSHRA standard.

The results are presented in Fig. 6. From Fig. 6 a, we observe that the proposed method, even with the smallest context of $L = 1$, consistently shows an improvement over the the corrupted signal, in most cases with no overlap between the confidence intervals. Between the conventional Wiener filter and the proposed method, mean of the condition $L = 1$ is rated around 10 points higher on average. Similarly, $L = 14$ is rated around 30 MUSHRA points higher than the Wiener filter. For all the items, the scores of $L = 14$ do not overlap with the Wiener filter scores, and is close to the ideal condition, especially at higher pSNRs. These observations are further supported in the difference plot, illustrated in Fig. 6 b. The scores for each pSNR were averaged over the male and female items. The difference scores were obtained by keeping the scores of the Wiener condition as reference and obtaining the difference between the three context-size conditions and the no enhancement condition. From these results we can conclude that, in addition to dithering, which can improve the perceptual quality of the decoded signal [12], applying noise reduction at the decoder using conventional techniques and further, employing models incorporating correlation inherent in the complex speech spectrum can improve pSNR significantly.

4. Conclusion and Future work

We propose a time-frequency based filtering method for the attenuation of quantization noise in speech and audio coding, wherein the correlation is statistically modeled and used at the decoder. Therefore, the method does not require the transmission of any additional temporal information, thus eliminating chances of error propagation due to transmission loss. By incorporating the contextual information, we observe pSNR improvement of 6 dB in the best case and 2 dB in a typical application; subjectively, an improvement of 10 to 30 MUSHRA points is observed.

In this work, we fixed the choice of the context neighborhood for a certain context size. While this provides a baseline for the expected improvement based on context size, it is interesting to examine the impact of choosing an optimal context neighborhood. Additionally, since the MVDR filter showed significant improvement in background noise reduction, a comparison between MVDR and the proposed MMSE method should be considered for this application.

In summary, we have shown that the proposed method improves both subjective and objective quality, and it can be used to improve the quality of any speech and audio codecs.

5. Acknowledgements

This project was supported by the Academic of Finland research project 312490.

6. References

[1] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*. Springer, 2017.

[2] "EVS codec detailed algorithmic description; 3GPP technical specification," http://www.3gpp.org/DynaReport/26445.htm.

[3] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach *et al.*, "Unified speech and audio coding scheme for high quality at low bitrates," in *ICASSP*. IEEE, 2009, pp. 1–4.

[4] ——, "A novel scheme for low bitrate unified speech and audio coding–MPEG RM0," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[5] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *ICASSP*. IEEE, 2011, pp. 273–276.

[6] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, 2012.

[7] H. Huang, L. Zhao, J. Chen, and J. Benesty, "A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction," *Digital Signal Processing*, vol. 33, pp. 169–179, 2014.

[8] S. Das and T. Bäckström, "Postfiltering using log-magnitude spectrum for speech and audio coding," in *Interspeech*, 2018.

[9] T. Bäckström, F. Ghido, and J. Fischer, "Blind recovery of perceptual models in distributed speech and audio coding," in *Interspeech*. ISCA, 2016, pp. 2483–2487.

[10] T. Bäckström and J. Fischer, "Coding of parametric models with randomized quantization in a distributed speech and audio codec," in *Proceedings of the 12. ITG Symposium on Speech Communication*. VDE, 2016, pp. 1–5.

[11] T. Bäckström and J. Fischer, "Fast randomization for distributed low-bitrate coding of speech and audio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2017.

[12] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial gray-scale," in *Proc. Soc. Inf. Disp.*, vol. 17, 1976, pp. 75–77.

[13] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the OPUS codec," in *Audio Engineering Society Convention 135*. Audio Engineering Society, 2013.

[14] T. Bäckström, J. Fischer, and S. Das, "Dithered quantization for frequency-domain speech and audio coding," in *Interspeech*, 2018.

[15] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2007.

[16] G. Fuchs, V. Subbaraman, and M. Multrus, "Efficient context adaptive entropy coding for real-time applications," in *ICASSP*. IEEE, 2011, pp. 493–496.

[17] T. Bäckström, "Estimation of the probability distribution of spectral fine structure in the speech source," in *Interspeech*, 2017.

[18] Y. Soon and S. N. Koh, "Speech enhancement using 2-D Fourier transform," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 717–724, 2003.

[19] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[20] M. Schoeffler, F. R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: a case study assessing basic audio quality following the ITU-R recommendation BS. 1534 (MUSHRA)," in *1st Web Audio Conference*. Citeseer, 2015.