Das, Sneha; Bäckström, Tom

Postfiltering Using Log-Magnitude Spectrum for Speech and Audio Coding

Published in:
Interspeech

DOI:
10.21437/Interspeech.2018-1027

Published: 01/09/2018

Document Version
Publisher's PDF, also known as Version of record

Interspeech 2018
2-6 September 2018, Hyderabad

Postfiltering Using Log-Magnitude Spectrum 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

Advanced coding algorithms yield high quality signals with good coding efficiency within their target bit-rate ranges, but their performance suffer outside the target range. At lower bitrates, the degradation in performance is because the decoded signals are sparse, which gives a perceptually muffled and distorted characteristic to the signal. Standard codecs reduce such distortions by applying noise filling and post-filtering methods. In this paper, we propose a post-processing method based on modeling the inherent time-frequency correlation in the log-magnitude spectrum. The goal is to improve the perceptual SNR of the decoded signals and, to reduce the distortions caused by signal sparsity. Objective measures show an average improvement of 1.5 dB for input perceptual SNR in range 4 to 18 dB. The improvement is especially prominent in components which had been quantized to zero.

Index Terms: Quantization noise, Speech modelling, postfiltering, noise filling, Time-Frequency correlation

1. Introduction

Speech and audio codecs are integral parts of most audio processing applications and recently we have seen rapid development in coding standards, such as MPEG USAC [1, 2], and 3GPP EVS [3]. These standards have moved towards unifying audio and speech coding, enabled the coding of super wide band and full band speech signals as well as added support of voice over IP. The core coding algorithms withing these codecs, ACELP and TCX, yield perceptually transparent quality at moderate to high bitrates within their target bitrate ranges. However, the performance degrades when the codecs operate outside this range. Specifically, for low-bitrate coding in the frequency-domain, the decline in performance is because fewer bits are at disposal for encoding, whereby areas with lower energy are quantized to zero. Such spectral holes in the decoded signal renders a perceptually distorted and muffled characteristic to the signal, which can be annoying for the listener.

To obtain satisfactory performance outside target bitrate ranges, standard codecs like CELP employ pre- and post-processing methods, which are largely based on heuristics. In particular, to reduce the distortion caused by quantization-noise at low bitrates, codecs implement methods either in the coding process or strictly as a post-filter at the decoder. Formant enhancement and bass post-filters are common methods [4] which modify the decoded signal based on the knowledge of how and where quantization noise perceptually distorts the signal. Formant enhancement shapes the codebook to intrinsically have less energy in areas prone to noise and is applied both at the encoder and decoder. In contrast, bass post-filter removes the noise like component between harmonic lines and is implemented only in the decoder.

Another commonly used method is noise filling, where pseudo-random noise is added to the signal [2], since accurate encoding of noise-like components is not essential for perception. In addition, the approach aids in reducing the perceptual effect of distortions caused by sparsity on the signal. The quality of noise-filling can be improved by parameterizing the noise-like signal, for example, by its gain, at the encoder and transmitting the gain to the decoder.

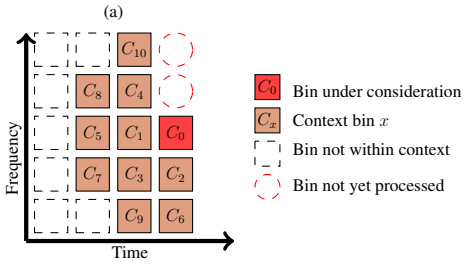

Figure 1: *Context neighborhood of size $C = 10$. The previous estimated bins are chosen and ordered based on the distance from the current sample.*

The advantage of post-filtering methods over the other methods is that they are only implemented in the decoder, whereby they do not require any modifications to the encoder-decoder structure, nor do they need any side information to be transmitted. However, most of these methods focus on solving the effect of the problem, rather than address the cause.

In this paper, we propose a post-processing method to improve signal quality at low bitrates, by modeling the inherent time-frequency correlation in speech magnitude spectrum and, investigating the potential of using this information to reduce quantization noise. The advantages of this approach are that it does not require the transmission of any side information and operates using solely the quantized signal as the observation and the speech models trained offline; Since it is applied at the decoder after the decoding process, it does not require any changes to the core structure of the codec; The approach addresses the signal distortions by *estimating* the information lost during the coding process using a source model. The novelties of this work lies in (i) incorporating the formant information in speech signals using log-magnitude modeling, (ii) representing the inherent contextual information in the spectral magnitude of speech in the log-domain as a multivariate Gaussian distribution (iii) finding the optimum, for the estimation of true speech, as the expected likelihood of a truncated Gaussian distribution.

2. Speech Magnitude Spectrum Models

Formants are the fundamental indicator of linguistic content in speech and are manifested by the spectral magnitude envelope of speech, therefore the magnitude spectrum is an important part of source modeling [5, 6]. Prior research has shown that frequency coefficients of speech are best represented by a Laplacian or Gamma distribution [7, 8, 9, 10]. Hence, the magnitude-spectrum of speech is an exponential distribution, as

10.21437/Interspeech.2018-1027

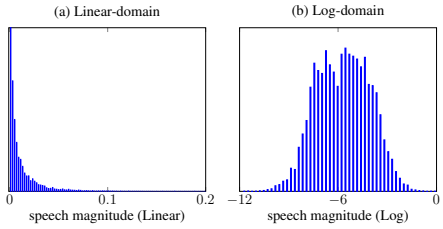

Figure 2: *Histograms of speech magnitude in (a) Linear domain (b) Log domain, in an arbitrary frequency bin.*

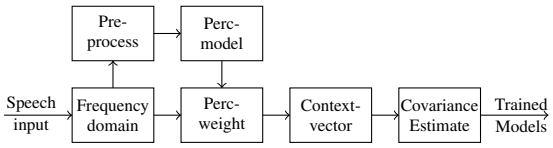

Figure 3: *Training of speech models*

shown in Fig. 2 a. The figure demonstrates that the distribution is concentrated at low magnitude values. This is difficult to use as a model because of numerical accuracy issues. Furthermore, it is hard to ensure the estimates are positive just by using generic mathematical operations. We address this problem by transforming the spectrum to the log-magnitude domain. Since the logarithm is non-linear, it redistributes the magnitude-axis such that the distribution of a exponentially distributed magnitude resembles the normal distribution in the logarithmic representation (Fig. 2 b). This enables us to approximate the distribution of the log-magnitude spectrum using a Gaussian probability density function (PDF).

In recent years, contextual information in speech has attracted a growing interest [11]. The inter-frame and inter-frequency correlation information have been explored previously in acoustic signal processing, for noise reduction [11, 12, 13]. The MVDR and Wiener filtering techniques employ the previous time- or frequency-frames to obtain an estimate of the signal in the current time-frequency bin. The results indicate a significant improvement in the quality of the output signal. In this work, we use similar contextual information to model speech. Specifically, we explore the plausibility of using the log-magnitude to model the context and, representing it using multivariate Gaussian distributions. The context neighborhood is chosen based on the distance of the context bin to the bin under consideration. Fig. 1 illustrates a context neighborhood of size 10 and indicates the order in which the previous estimates are assimilated into the context vectors.

The overview of the modeling process is presented in Fig. 3. The input speech signal is transformed to the frequency domain by windowing and then applying the short-time Fourier transform (STFT). The frequency domain signal is then pre-processed and perceptually weighted using the computed perceptual envelope, similar to CELP [14, 4]. Finally, the context vectors are extracted for each sample frequency-bin, and then the covariance matrix for each frequency band is estimated, thus providing the required speech models. We explored context sizes upto 40, which includes approximately four previous time frames, lower and upper frequency bins, each. Note that we operate with STFT instead of MDCT which is used in standard codecs, in order to keep this work extensible to enhancement applications. Expansion of this work to MDCT is ongoing and informal tests provide insights similar to this paper.

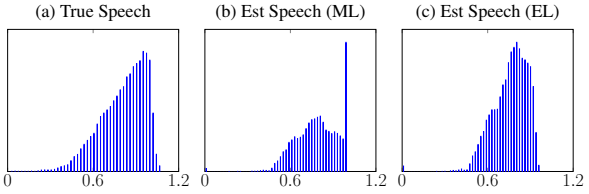

Figure 4: *Histograms of Speech distribution (a) True (b) Estimated: ML (c) Estimated: EL.*

3. Problem Formulation

Our objective is to estimate the clean speech signal from the observation of the noisy decoded signal using the statistical priors. To this end, we formulate the problem as the maximum likelihood (ML) of the current sample given the observation and the previous estimates. Assume a sample x has been quantized to a quantization level $Q \in [l, u]$. We can then express our optimization problem as:

$$\hat{x} = \arg\max_{x} \quad P(X|\mathbf{X}_c = \hat{\mathbf{x}}_c) \quad \text{subject to}, \; l \leq X \leq u \tag{1}$$

where $\hat{x}$ is the estimate of the current sample, l and u are the lower and upper limits of the current quantization bins, respectively, and, $P(a_1|a_2)$ is the conditional probability of a_1, given a_2. $\hat{\mathbf{x}}_c$ is the estimated context vector. Fig. 1 illustrates the construction of a context vector of size $C = 10$, wherein the numbers represent the order in which the frequency bins are incorporated. We obtain the quantization levels from the decoded signal and from our knowledge of the quantization method used in the codec, we can define the quantization limits; the lower and upper limits of a specific quantization level is defined midway between previous and subsequent levels, respectively.

To illustrate the performance of Eq. 1, we solved it using generic numerical methods. Fig. 4 illustrates the results through distributions of the true speech (a) and estimated speech (b), in bins quantized to zero. We scale the bins such that the varying l and u are fixed to $0, 1$, respectively, in order to analyze and compare the relative distribution of the estimates within a quantization bin. In (b) we observe a high data density around 1, which implies that the estimates are biased towards the upper limits. We shall refer to this as the edge-problem.

To mitigate this problem, we define the speech estimate as the expected likelihood (EL) [15, 16], as follows:

$$\hat{x} = \quad E[P(X|\mathbf{X}_c = \hat{\mathbf{x}}_c)] \quad \text{subject to.} \; l \leq X \leq u \tag{2}$$

The resulting speech distribution using EL is demonstrated in Fig. 4 c, indicating a relatively better match between the estimated-speech and the true-speech distributions. Finally, to obtain an analytical solution, we incorporate the constraint condition into the modeling itself, whereby we model the distribution as a truncated Gaussian PDF [17]. In appendices A & B, we demonstrate how the solution can be obtained as a truncated Gaussian. Algorithm. 1 presents an overview of the estimation method.

4. Experiments and Results

Our objective is to evaluate the advantage of modeling the log-magnitude spectrum. Since envelope models are the main method for modeling the magnitude spectrum in conventional codecs, we evaluate the effect of statistical priors both in terms of the whole spectrum as well as only for the envelope. Therefore, besides evaluating the proposed method for the estimation

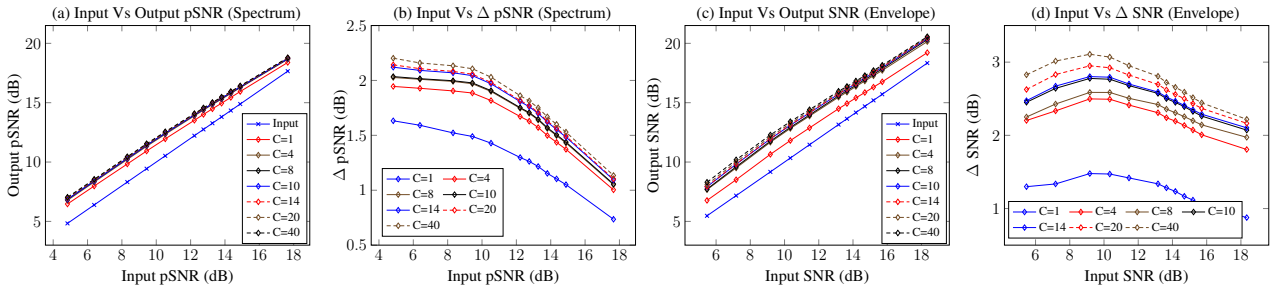

Figure 5: *Plots representing the improvement in SNR using the proposed method for different context sizes.*

Algorithm 1 Estimation of signal from quantized observation

Require: Quantized signal Y, prior-models C
  function ESTIMATION(Y, C)
    for $frame = 1 : N$ **do**
      for $b = 1 : Length(Y(frame))$ **do**
        $\mu_{up}, \sigma_{up} \leftarrow UpdateStatistics(C, \hat{X}_{prev})$
        $pdf \leftarrow TruncateGaussian(\mu_{up}, \sigma_{up}, l(b), u(b))$
        $\hat{X} \leftarrow Expectation(pdf)$

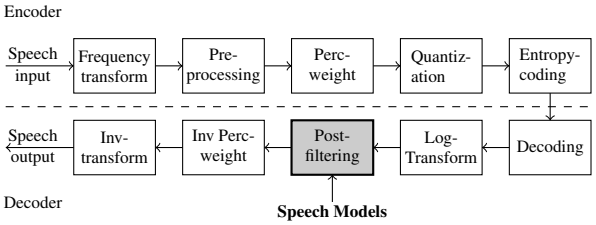

Figure 6: *Systems overview.*

of speech from the noisy magnitude spectrum of speech, we also test it for the estimation of the spectral envelope from an observation of the noisy envelope. To obtain the spectral envelope, after transforming the signal to the frequency domain, we compute the Cepstrum and retain the 20 lower coefficients and transform it back to the frequency domain. The next steps of envelope modeling are the same as spectral magnitude modeling presented in Sec. 2 and Fig. 3, i.e. obtaining the context vector and covariance estimation.

4.1. System overview

A general block diagram of the system is presented in Fig. 6. At the encoder, the signals are divided into frames of 20 ms with 50% overlap and Sine windowing. The speech input is then transformed to the frequency domain using the STFT. After pre-processing and perceptually weighting the signal by the spectral envelope, the magnitude spectrum is quantized and entropy coded using arithmetic coding [18]. At the decoder, the reverse process is implemented to decode the signal. The decoded signal is thus corrupted by quantization noise and our purpose is to use the proposed post-processing method to improve output quality. Note that we apply the method in the perceptually weighted domain. After post-processing, the estimated speech is transformed back to the temporal domain by applying the inverse perceptual weights and the inverse frequency transform. Since the focus of this paper is to study spectral magnitude modeling, we use true phase to reconstruct the signal back to temporal domain.

4.2. Experimental setup

For training we use 250 speech samples from the training set of the TIMIT database [19]. The block diagram of the training process is presented in Fig. 3. For testing, 10 speech samples were randomly chosen from the test set of the database. The codec is based on the EVS codec [20] in TCX mode and we chose the codec parameters such that the perceptual SNR (pSNR) [20, 4] is in the range typical to codecs. Therefore, we simulated coding at 12 different bitrates between 9.6 to 128 kbps, which gives pSNR values in the approximate range of 4 and 18 dB. Note that the TCX mode of EVS does not incorporate post-filtering. For each test case, we apply the post-filter to the decoded signal with context sizes $\in \{1, 4, 8, 10, 14, 20, 40\}$. The context vectors are obtained as per the description in Sec. 2 and illustration in Fig. 1. For tests using the magnitude spectrum, the pSNR of the post-processed signal is compared against the pSNR of the noisy quantized signal. For spectral envelope based tests, the signal-to-Noise Ratio (SNR) between the true and the estimated envelope is used as the quantitative measure.

4.3. Results and analysis

The average of the qualitative measures over the 10 speech samples are plotted in Fig. 5. Plots (a) and (b) represent the evaluation results using the magnitude spectrum and, plots (c) and (d) correspond to the spectral envelope tests. For both, the spectrum and the envelope, incorporation of contextual information shows a consistent improvement in the SNR. The degree of improvement is illustrated in plots (b) and (d). For magnitude spectrum, the improvement ranges between 1.5 and 2.2 dB over all the context at low input pSNR, and from 0.2 to 1.2 dB higher input pSNR. For spectral envelopes, the trend is similar; the improvement over context is between 1.25 to 2.75 dB at lower input SNR, and from 0.5 to 2.25 at higher input SNR. At around 10dB input SNR, the improvement peaks for all context sizes.

For the magnitude spectrum, the improvement in quality between context size 1 and 4 is significantly large, approximately 0.5 dB over all input pSNRs. By increasing the context size we can further improve the pSNR, but the rate of improvement is relatively lower for sizes from 4 to 40. Also, the improvement is considerably lower at higher input pSNRs. We conclude that a context size around 10 samples is a good compromise between accuracy and complexity. However, the choice of context size can also depend on the target device for processing. For instance, if the device has computational resources at disposal, a high context size can be employed for maximum improvement.

Performance of the proposed method is further illustrated in Figs. 7- 8, with an input pSNR of 8.2 dB. A prominent observation from all plots in Fig. 7 is that, particularly in bins

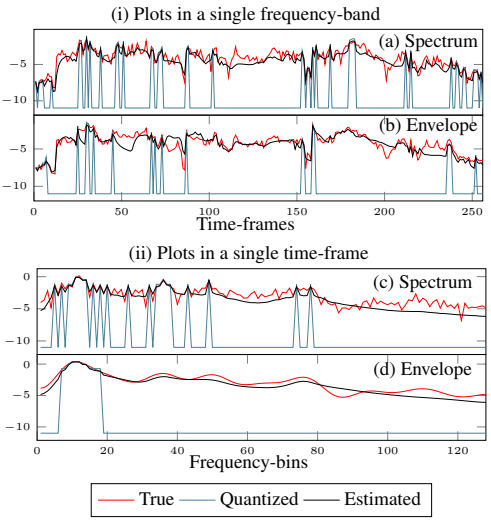

Figure 7: *Sample plots depicting the true, quantized and the estimated speech signal (i) in a fixed frequency band over all time frames (ii) in a fixed time frame over all frequency bands.*

quantized to zero the proposed method is able to estimate magnitude which is close to the true magnitude. Additionally from Fig. 7 (ii), the estimates seem to follow the spectral envelope, whereby we can conclude that Gaussian distributions pre-dominantly incorporate spectral envelope information and not so much of pitch information. Hence, additional modeling methods for the pitch will be addressed in future work.

The scatter plots in Fig. 8 represent the correlation between the true, estimated and quantized speech magnitude in zero-quantized bins for $C = 1$ and $C = 40$. These plots further demonstrate that context is useful in estimating speech in bins where no information exists. Thus this method can be beneficial in estimating spectral magnitudes in noise-filling algorithms. In the scatter plots, the quantized, true and estimated speech magnitude spectrum are represented by red, black and blue points, respectively; We observe that while the correlation is positive for both sizes, the correlation is significantly higher and more defined for $C = 40$.

5. Discussion and Conclusion

In this work, we investigated the use of contextual information inherent in speech for the reduction of quantization noise. We propose a post-processing method with focus on estimating speech samples at the decoder, from the quantized signal using statistical priors. Results indicate that including speech correlation not only improves the pSNR, but also provide spectral magnitude estimates for noise filling algorithms. While the focus of this paper was modeling the spectral magnitude, a joint magnitude-phase modeling method, based on current insights and the results from an accompanying paper [21], is the natural next step.

This work also begins to tread on spectral envelope restoration from highly quantized noisy envelopes by incorporating information for the context neighborhood. In future work we should explore the interaction between our method and quantization of conventional envelope parameterizations. Additionally, since this method has shown a capacity to restore information from areas where no information exists at all, it will be interesting to see the application of this method to packet-loss concealment.

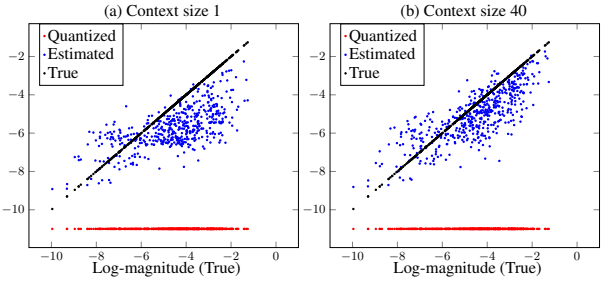

Figure 8: *Scatter plots of the true, quantized and estimated speech in zero-quantized bins for (a) $C = 1$, (b) $C = 40$. The plots demonstrate the correlation between the estimated and true speech.*

6. Acknowledgements

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

A. Truncated Gaussian PDF

Let us define $f_1(a) = e^{-\frac{(a-\mu)^2}{2\sigma^2}}$ and $f_2(a) = \text{erf}(\frac{a-\mu}{\sigma\sqrt{2}})$, where μ, σ are the statistical parameters of the distribution and erf is the error function. Then, expectation of a univariate Gaussian random variable X is computed as:

$$[E(X)]_{-\infty}^{\infty} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x f_1(x) dx, \tag{3}$$

Conventionally, when $X \in [-\infty, \infty]$, solving Eq. 3 results in $E(X) = \mu$. However, for a truncated Gaussian random variable, with $l < X < u$, the relation is

$$E(X|l < X < u) = \frac{[E(X)]_l^u}{\int_l^u P(x)dx} = \frac{\int_l^u x f_1(x)dx}{\int_l^u f_1(x)dx}, \tag{4}$$

which yields the following equation to compute the expectation of a truncated univariate Gaussian random variable:

$$E(X|l < X < u) = \mu - \sigma\sqrt{\frac{2}{\pi}} \left[\frac{f_1(u) - f_1(l)}{f_2(u) - f_2(l)}\right] \tag{5}$$

B. Conditional Gaussian parameters

Let the context vector be defined as $\mathbf{x} = [x_1, \mathbf{x}_2]^T$, wherein $x_1 \in \mathbb{R}^{1X1}$ represents the current bin under consideration, and $\mathbf{x}_2 \in \mathbb{R}^{CX1}$ is the context. Then, $\mathbf{x} \in \mathbb{R}^{(C+1)X1}$, where C is the context size. The statistical models are represented by the mean vector $\boldsymbol{\mu} \in \mathbb{R}^{(C+1)X1}$, and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{(C+1)X(C+1)}$, such that $\boldsymbol{\mu} = [\mu_1, \boldsymbol{\mu}_2]^T$ with dimensions same as x_1 and $\mathbf{x}_2$, and the covariance as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{6}$$

$\boldsymbol{\Sigma}_{ij}$ are partitions of $\boldsymbol{\Sigma}$ with dimensions $\Sigma_{11} \in \mathbb{R}^{1X1}$, $\boldsymbol{\Sigma}_{\mathbf{22}} \in \mathbb{R}^{CXC}$, $\boldsymbol{\Sigma}_{\mathbf{12}} \in \mathbb{R}^{1XC}$ and $\boldsymbol{\Sigma}_{\mathbf{21}} \in \mathbb{R}^{CX1}$. Thus, the updated statistics of the distribution of the current bin based on the estimated context is [22]:

$$\mu_{up} = \mu_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\hat{\mathbf{x}}_c - \boldsymbol{\mu}_2) \tag{7}$$

$$\sigma_{up} = \Sigma_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{8}$$

7. References

[1] S. Quackenbush, "MPEG unified speech and audio coding," *IEEE MultiMedia*, vol. 20, no. 2, pp. 72–78, 2013.

[2] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach *et al.*, "A novel scheme for low bitrate unified speech and audio coding–MPEG RM0," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[3] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the EVS codec architecture," in *ICASSP*. IEEE, 2015, pp. 5698–5702.

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

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

[6] T. Barker, "Non-negative factorisation techniques for sound source separation," Ph.D. dissertation, Tampere University of Technology, 2017.

[7] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *ICASSP*, vol. 9, Mar 1984, pp. 53–56.

[8] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *ICASSP*, vol. 1, May 2002, pp. I–253–I–256.

[9] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with superGaussian priors," in *ICASSP*, vol. 1, April 2003, pp. I–896–I–899 vol.1.

[10] T. H. Dat, K. Takeda, and F. Itakura, "Generalized gamma modeling of speech and its online estimation for speech enhancement," in *ICASSP*, vol. 4, March 2005, pp. iv/181–iv/184 Vol. 4.

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

[12] 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.

[13] 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.

[14] T. Bäckström and C. R. Helmrich, "Arithmetic coding of speech and audio spectra using TCX based on linear predictive spectral envelopes," in *ICASSP*, April 2015, pp. 5127–5131.

[15] E. T. Northardt, I. Bilik, and Y. I. Abramovich, "Spatial compressive sensing for direction-of-arrival estimation with bias mitigation via expected likelihood," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1183–1195, 2013.

[16] Y. I. Abramovich and O. Besson, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach part 1: The over-sampled case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5807–5818, 2013.

[17] N. Chopin, "Fast simulation of truncated Gaussian distributions," *Statistics and Computing*, vol. 21, no. 2, pp. 275–288, 2011.

[18] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM Journal of research and development*, vol. 23, no. 2, pp. 149–162, 1979.

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

[21] S. Das and T. Bäckström, "Postfiltering with complex spectral correlations for speech and audio coding," in *Interspeech*, 2018.

[22] S. Korse, G. Fuchs, and T. Bäckström, "GMM-based iterative entropy coding for spectral envelopes of speech and audio," in *ICASSP*. IEEE, 2018.