Interspeech 2018
2-6 September 2018, Hyderabad

Dithered Quantization for Frequency-Domain Speech and Audio Coding

Tom Bäckström[1], *Johannes Fischer*[2] *and Sneha Das*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2]International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany

first.lastname@aalto.fi

Abstract

A common issue in coding speech and audio in the frequency domain, which appears with decreasing bitrate, is that quantization levels become increasingly sparse. With low accuracy, high-frequency components are typically quantized to zero, which leads to a muffled output signal and musical noise. Bandwidth extension and noise-filling methods attempt to treat the problem by inserting noise of similar energy as the original signal, at the cost of low signal to noise ratio. Dithering methods however provide an alternative approach, where both accuracy and energy are retained. We propose a hybrid coding approach where low-energy samples are quantized using dithering, instead of the conventional uniform quantizer. For dithering, we apply 1 bit quantization in a randomized sub-space. We further show that the output energy can be adjusted to the desired level using a scaling parameter. Objective measurements and listening tests demonstrate the advantages of the proposed methods.
Index Terms: speech and audio coding, dithering, noise filling, perceptual coding, 1 bit quantization.

1. Introduction

State-of-the art codecs, such as 3GPP Enhanced Voice Services (EVS) and MPEG Unified speech and audio coding (USAC) use frequency domain coding in their intermediate and high bitrate ranges, but revert to time-domain coding at lower bitrates [1–3]. The reason is that the scalability of frequency-domain codecs in terms of coding efficiency at low bitrates remains a bottleneck even if they provide several other advantages such as low algorithmic complexity. A symptom of the issue is that frequency-domain codecs tend to quantize low-energy areas to zero, which further reduces their energy. This leads to a muffled character in the quantized output, since high-frequency components often have low energy and are thus zeroed (see Fig. 1).

The problem of uniform quantization, in the conventional application, is that if the quantization bins are zero-centered, then the energy of the quantized signal decreases with decreasing accuracy. Alternatively, with off-center quantization we can retain the average energy, but are limited in bit-rate to above 1 bit/sample, since we have to transmit the sign. Moreover, at the extreme, at low bitrates, non-zero values can require so many bits to encode in the entropy coder, that we cannot ever afford to transmit them. Entropy coding with uniform quantization therefore does not scale well to bitrates below 1 bit/sample.

This problem has been addressed in prior works primarily with two approaches. Firstly, we can encode high-frequency regions with *bandwidth extension* methods, where the objective is to retain the spectral magnitude envelope of the original signal, but sacrifice phase-information and other fine-structure, such that bitrate is limited [3–6]. Sometimes such methods also copy spectral structures from lower frequencies (copy-up) since the fine-structures are generally similar. Secondly, with a method known as *noise filling*, we can insert noise in areas of the spectrum which have been quantized to zero such that absence of energy is avoided [7]. A recent improvement, known as *intelligent gap filling*, combines these methods by using both noise filling and copy-up [8]. All three approaches thus aim to retain energy at a similar level as the original signal, but they do not optimize signal-to-noise ratio.

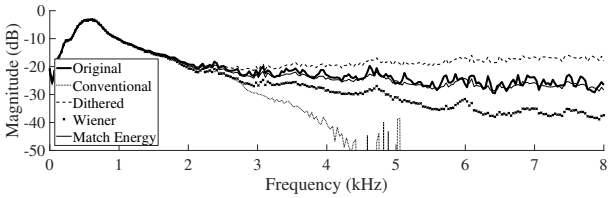

Figure 1: *Mean energy of perceptually weighted and normalized MDCT-spectra over the TIMIT database, for original signal (thick line), conventional quantization (dotted), dithered (dashed), as well as dithered in combination with Wiener filtering (crosses) and matching energy (thin line). Quantization was scaled to match a bitrate of 13.2 kbit/s.*

Classical dithering algorithms however also include methods which can retain the signal distribution without reduction in signal to noise ratio [9]. Common dithering methods such as Floyd-Steinberg, are based on error-diffusion or randomization of quantization levels, such that quantization errors can be diffused without loss in accuracy [10]. Alternatively, we can modify quantization bin locations to retain the probability distribution of the original signal even after quantization and coding [11] or use lattice quantization to pack quantization more densely [12]. These approaches however do not address the issue of very low bitrates, where we cannot afford to encode anything else than the most likely quantization bin. Algebraic coding can be used also at very low bitrates, but its output is also very sparse and it is not applicable on all bitrates [3, 13]. A further alternative would be vector coding, which provides optimal coding efficiency also at very low bitrates. However, vector coding approaches are not easily scalable across bitrates. Moreover, their computational complexity becomes prohibitively high at higher bitrates and if the vector length is high [3, 14]. Vector coding is thus also not a scalable approach.

In our recent works, we have presented an alternative method using dithered quantization, where the input signal is multiplied with a random rotation before quantization such that the quantization levels are obscured when the rotation is inverted for the output signal [15]. A similar approach is applied in the Opus codec [16], though only with Givens-rotations without permutations. We can thus apply simple quantization such as 1 bit quantization to obtain high performance at low complexity and very low bitrates [17]. The proposed randomized

10.21437/Interspeech.2018-46

quantization methods are unique in the way they allow quantization and coding of signals without a lower limit on bitrate, while simultaneously providing the best SNR per bit ratio. Concurrently, the proposed methods provide the benefits of vector coding by joint processing of multiple samples, without significant penalty on complexity.

In this paper we present an application of the proposed randomization for dithered quantization in frequency-domain coding of speech and audio, to allow coding at very low bitrates without excessive sparseness or low energy in the output. The central novelty is a hybrid structure where dithering is applied to low-energy samples and uniform quantization with arithmetic coding elsewhere. Perceptual listening tests demonstrate that the proposed dithered quantizer gives the best performance.

2. Quantization Methods

Our objective is to study the performance of dithered quantization methods in comparison to conventional uniform quantization, and in combination with entropy coding. In the TCX mode of EVS [1, 18], entropy coding and uniform quantization is implemented assuming that the sample distribution is Laplacian, and the sample variance is estimated using the linear predictive envelope. The quantization accuracy is determined in a rate-loop such that the bit-budget is used as effectively as possible. In a vector of samples, trailing zeros are truncated. The scaling of the signal is determined after quantization, such that the output signal-to-noise ratio is optimized. We will use this implementation of uniform quantization as our baseline system.

We have recently proposed an approach for dithering and encoding data at low bit-rates (less than 1 bit/sample), based on random rotations and which is defined as follows [15]. Suppose we have a vector $x \in \mathbb{R}^{N\times 1}$, which we want to encode with $B \leq N$ bits. Using a random orthonormal matrix A, known at both the encoder and decoder, we can then quantize

$$\hat{x} = A^T Q_B[Ax], \tag{1}$$

where $Q_B[\cdot]$ is a quantizer defined as

$$Q_B[y] := \gamma \begin{bmatrix} \text{sign}(y_0) \\ \text{sign}(y_1) \\ \vdots \\ \text{sign}(y_{B-1}) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2}$$

and γ is a scaling coefficient. In other words, it uses a 1 bit-quantizer, where the B first samples are quantized with the sign of the input sample, for a total bitrate of B. The quantized $\hat{x}$ then has an approximately normal distribution and the variance is $E[|\hat{x}|^2] = \gamma^2 \frac{B}{N}$. We have furthermore shown that the orthonormal matrix A can be approximated by a low-order rotation such that the algorithmic complexity is linear $\mathcal{O}(N)$ [15].

We can then readily show that γ can be chosen according to a number of criteria, for example:

1. $\gamma_{\text{MMSE}} = \sigma\sqrt{\frac{2}{\pi}}$ is the minimum mean square error (MMSE) scaling for normal input of variance σ^2. This thus corresponds to Wiener filtering the quantized signal.
2. $\gamma_{\sigma^2} = \sigma\sqrt{\frac{N}{B}}$ retains the variance σ^2 (i.e. energy matching) of the original signal. This corresponds to quantization on the surface of an N-dimensional hyper-sphere, which is normalized to original energy.

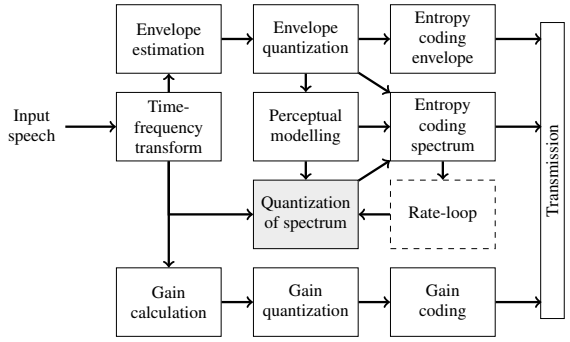

Figure 2: *Diagram of the speech and audio encoder. The gray box is modified in the current work.*

Clearly, γ can thus be tuned according to perceptual criteria, for a balance between accuracy and how well the quantizer retains the signal distribution and variance.

This approach to quantization provides error-diffusion similar to Floyd-Steinberg -type methods. However, in difference, the error is not diffused forward to following components, but instead, it is diffused among the samples of a vector. In signal processing terms, Floyd-Steinberg -type methods are thus similar to infinite impulse responses (IIR) filters, whereas the proposed method is more like a finite impulse response (FIR) operation.

3. Speech and Audio Coding Framework

To evaluate the error characters of different quantizers, we need to implement them in a speech and audio codec which allows a fair comparison. This task is less straightforward than one might expect. The main issue is that codecs regularly use *ad hoc* tricks to overcome saturation effects of the arithmetic coder [1, 3, 18]. Namely, for example, high-frequency samples quantized to zero are typically truncated from the spectrum above the last non-zero sample. By omitting the transmission of zero samples we can save a considerable amount of bits, which can instead be used for coding low-frequency components. The performance of the arithmetic coder, in isolation, does therefore not accurately reflect the performance of the overall codec.

To obtain a fair comparison, we will therefore implement a state-of-the-art baseline system following the simplified structure of the 3GPP Enhanced Voice Services (EVS) [1, 3, 18] (see Fig. 2). For frequency-domain coding, we use here the MDCT-transform with a window length of 30 ms, 50 % overlap, a half-sine window and pre-emphasis with a filter $P(z) = 1 - 0.68z^{-1}$. At a sampling rate of 16 kHz, the magnitude envelope is modeled with a linear predictive model of order $M = 20$, which we use as an estimate of the variance of each frequency component, and which is further fed into a conventional arithmetic coder with an assumption of a Laplacian distribution. We apply quantization in the perceptually weighted domain as in [3]. Note that we did not implement a deadzone-quantizer, even if it is known to improve signal-to-noise ratio [7], because it also amplifies the saturation effect at high frequencies. A deadzone-quantizer would therefore have unfairly penalized the baseline codec in terms low-rate performance.

Conventional codecs include noise-fill and bandwidth-extension methods to reduce the bitrate and to compensate for the energy-loss at high frequencies. To allow a straightforward and fair comparison between methods, we did not include

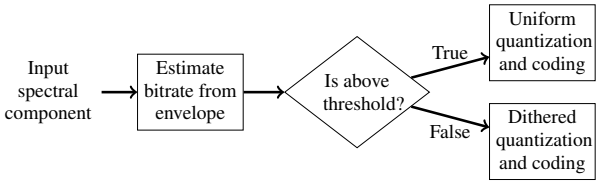

Figure 3: *Hybrid coding of spectral components.*

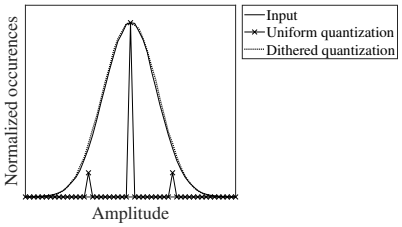

Figure 4: *Collated histograms of $K = 10000$ vectors of unit variance Gaussian input, quantized by uniform quantization and entropy coding as well as the proposed dithered coder ($N = 32$, γ_{σ^2}), with 1 bit/sample.*

bandwidth-extension in the codec. Our noise-fill algorithm is applied at frequencies above 1.6 kHz, on all spectral components which are quantized to zero, where we add noise with a random sign, and adjust the magnitude to match that obtained with the proposed dithering method with gain γ_{MMSE}. The noise-fill used in EVS uses advanced signal analysis to fine-tune noise-filling, but but we chose this simplified method to make the test easy to reproduce. All parameters should anyway be tuned to the particular configuration of the final codec and thus further perceptual tuning of parameters is not worthwhile for these experiments.

For the bitrate of the codec, we assume that spectral envelope, gain and other parameters are encoded with 2.6 kbits/s and thus the remaining bits can be used for encoding the spectrum. Further, for simplicity and reproducability, we did not quantize any other parameters of the signal. It should be noted, however, that bitrate calculations in this paper are provided only to assist the reader in getting a realistic impression of performance, as the bitrate of side-information can vary in particular implementations of codecs.

4. Proposed Hybrid Coder

The combination of uniform quantization and arithmetic coder saturates at low bitrates and hence we propose to replace the conventional approach by dithered coding for spectral samples whose bitrate is below 1 bit/sample. It is thus a hybrid entropy coder, which uses uniform quantization and arithmetic coding following [18] in high-energy areas of the spectrum and dithered coding at the low-energy areas.

The baseline entropy coder uses the linear predictive envelope to estimate the variance σ_k^2 of frequency components [18]. Note that this envelope has to be scaled such that the expected bitrate of a signal which follows that envelope, matches the target bitrate. Based on the variance σ_k^2 of the kth component, we can then estimate the expected bitrate of a sample as $b_k = \frac{1}{2}\log_2(4.1159\sigma_k^2)$ but limited to $b_k \geq 0$. For spectral components with $b_k > 1$ we use uniform quantization and arithmetic coding, for b_k we apply dithered coding (see Fig. 3). The bit-allocation between uniform and dithered quantization is derived directly from the expected bitrate b_k.

We thus collate all low-energy samples into a vector x and quantize them with Eq. 1. Implicitly, we thus assume that vector x follows the normal distribution with uniform variance. To improve accuracy, we could further subdivide x according to their variance, but such modifications are left for further study.

In [15], we have demonstrated that the randomization matrix A of sufficient quality can be readily generated with 4 iterations of $N/2$ random 2×2 rotations and length N permutations, when the bitrate is $B = N$. However, with $B \ll N$, a majority of samples are zeros, and therefore we increased the number of iterations to 8 such that the output distribution remains normal. Random rotations between the non-zero and zeroed values could be readily used to reduce the computational complexity without effect on the statistics of the output signal.

5. Experiments

To evaluate the performance the proposed hybrid codec, in comparison to uniform quantization, we performed three types of experiments. Firstly, we study performance of dithering in isolation with synthetic input. Secondly, we encode speech from the TIMIT corpus and evaluate performance by objective criteria. Finally, using samples from the TIMIT corpus, we performed a MUSHRA subjective listening test to determine perceptual preference among methods [19].

The output distribution of the proposed dithered quantization (Eq. 1) in comparison to uniform quantization is illustrated in Fig. 4. Here we encoded normally distributed $K = 10000$ vectors of length $N = 32$ with $B = 32$ bits and used the gain factor γ_{σ^2}. We can readily observe that uniform quantization is unable to retain the shape of the original distribution, whereas the distribution of the output of the proposed dithered codec exactly matches that of the input.

The performance of the proposed coder for a single frame of speech is illustrated in Fig. 5; The spectral magnitude envelope is estimated using linear predictive modelling in Fig. 5(a), the expected bit-rate for each frequency is estimated from the envelope using the method developed in [18] in Fig. 5(b) and a threshold is applied to determine the choice of quantizer. Finally, in Fig. 5(c), the quantized output of the conventional method is compared with the proposed method, where the gain factor was γ_{σ^2}. We can clearly see that whereas for the conventional approach, all frequencies above 2 kHz are quantized to zero, the proposed method retains the spectral shape also at the higher frequencies.

For objective evaluation of performance on real speech, we encoded the entire TIMIT database (training and evaluation) with different combinations of quantization and entropy coding [20]. Namely, we applied 1. uniform quantization with arithmetic coding following [18] (Conventional), 2. a dithering simulation by adding white noise to obtain same signal to noise ratio as the conventional approach (Dithering), 3. the proposed hybrid codec using γ_{MMSE} (1 bit MMSE) and 4. using γ_{σ^2} (1 bit EM). The mean output energy across frequencies for each method is illustrated in Fig. 1.

We can immediately see that all modifications of conventional arithmetic coding bring the amount of energy closer to the original energy envelope. The dithering simulation saturates at the perceptual noise floor near −20 dB, which is higher than the original energy envelope. Informal listening confirms that such dithering has a noisy character, where the conventional is

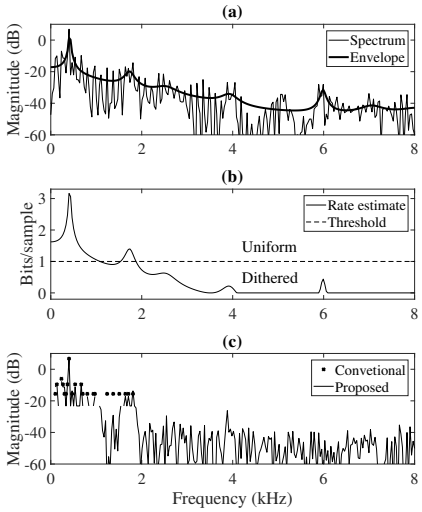

Figure 5: *Illustration of performance for a typical speech spectrum at 13.2 kbits/s. (a) Input signal spectrum and its envelope, (b) the bit-rate estimated from the envelope and the threshold where quantizers are switched and (c) the quantized spectra with conventional uniform quantization and entropy coding in comparison to the proposed, dithered coder.*

Table 1: *Mean signal to noise ratio in the perceptually weighted domain for the conventional and the two proposed methods.*

	1 bit MMSE	1 bit EM	Conventional
SNR (dB)	10.75	10.46	8.93

muffled. The two proposed 1 bit dithering methods are closer to the original energy envelope, such that the MMSE approach γ_{MMSE} is clearly below the original while the energy matching method γ_{σ^2} approximates nicely the desired energy envelope.

The average signal to noise ratios (SNR) in the perceptual domain for the conventional and proposed methods are listed in Table 1. Clearly the 1 bit MMSE approach reaches the highest SNR, since it was designed to optimize SNR. However, as the conventional method is also designed to optimize SNR, it is surprising that we obtained such a large improvement of 1.81 dB. The energy-matching approach γ_{σ^2} looses slightly in SNR to the MMSE approach, but the difference of 0.29 dB is not large. We need a subjective listening test to determine if it is more important to preserve envelope shape or optimize SNR.

Finally, to determine subjective preference among methods, we performed a MUSHRA listening test [19]. In the test, we included 6 samples (3 male and 3 female) randomly chosen from the TIMIT corpus [20]. In addition to the above methods, Conventional, Dithered, 1 bit MMSE and 1 bit EM, we included here also a case where the conventional uniform coder is enhanced by noise filling in post-processing. It was not included in the previous tests because it is a blind post-processing method in the sense that it adds noise without any transmitted information from the input signal. It thus reduces SNR even if it is designed to improve perceptual quality. In the listening test, we had 14 normal hearing subjects in the age-range 26 to 43 years. Fig. 6 illustrates the results.

We observe for all items, that the proposed dithered 1 bit quantizers have a higher mean than the other methods. Moreover, in the mean over all items (the "All" column), the proposed dithered 1 bit quantizers have a statistically significant

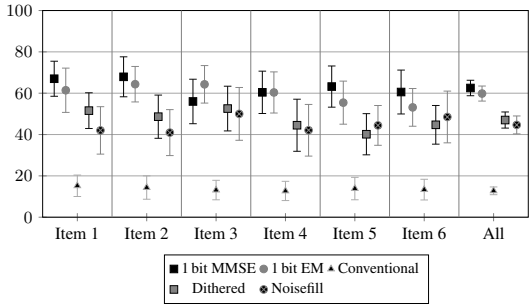

Figure 6: *Results of a subjective MUSHRA listening test, comparing the proposed 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor.*

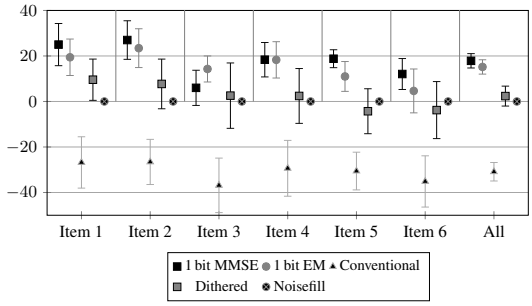

Figure 7: *Differential scores of a subjective MUSHRA listening test, comparing the proposed 1 bit dithered quantizers with conventional arithmetic coding, as well as a synthetic dithering serving as an anchor. Differences are calculated with the noisefill as reference.*

difference to the antecedent methods. Conventional arithmetic coding without noisefill also shows a statistically significant reduction in quality in comparison to all other methods. To further determine whether listeners have a preference among the two proposed dithered quantizers, we calculated the differential MUSHRA scores with noisefill as a reference (see Fig. 7). However, the differential scores revealed no additional details.

6. Discussion and Conclusions

Dithering is a classic method in signal processing and it is thus useful to investigate whether is applicable also in speech and audio coding. Our literature survey shows that conventional methods in coding, such as noisefill and bandwidth-extension, include features similar to dithering, but that they do not attempt to optimize SNR. In contrast, in the current work, we propose to use a recently developed method for dithering and coding which is applicable to very low bitrates [15]. The approach is based on a random rotation, sign-quantization in the randomized domain, and an inverse transform. We propose to apply it in combination with conventional uniform quantization and entropy coding, such that only frequency components where we can afford to use very little accuracy, are coded with the dithered quantizer.

By using dithering we can avoid the characteristic problem of conventional frequency-domain codecs, where higher frequencies are often quantized to zero such the output sounds muffled. In other words, the output is not unnaturally sparse. Our objective and subjective experiments demonstrate that the method gives a nice improvement in perceptual quality.

7. References

[1] 3GPP, *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 2014.

[2] ISO/IEC 23003–3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

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

[4] M Dietz, L Liljeryd, K Kjorling, and O Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*, 2002.

[5] V Atti, V Krishnan, D Dewasurendra, V Chebiyyam, S Subasingha, D J Sinder, V Rajendran, I Varga, J Gibbs, L Miao, V Grancharov, and H Pobloth, "Super-wideband bandwidth extension for speech in the 3GPP EVS codec," in *Proc. ICASSP*, 2015, pp. 5927–5931.

[6] P Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA'02)*, 2002.

[7] G Fuchs, C R Helmrich, G Marković, M Neusinger, E Ravelli, and T Moriya, "Low delay LPC and MDCT-based audio coding in the EVS codec," in *Proc. ICASSP*, 2015, pp. 5723–5727.

[8] S Disch, A Niedermeier, C R Helmrich, C Neukam, K Schmidt, R Geiger, J Lecomte, F Ghido, F Nagel, and B Edler, "Intelligent gap filling in perceptual transform coding of audio," in *Audio Engineering Society Convention 141*. Audio Engineering Society, 2016.

[9] J Vanderkooy and S P Lipshitz, "Dither in digital audio," *Journal of the Audio Engineering Society*, vol. 35, no. 12, pp. 966–975, 1987.

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

[11] M Li, J Klejsa, and W B Kleijn, "Distribution preserving quantization with dithering and transformation," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, 2010.

[12] J D Gibson and K Sayood, "Lattice quantization," *Advances in electronics and electron physics*, vol. 72, pp. 259–330, 1988.

[13] T Bäckström, "Enumerative algebraic coding for ACELP," in *Proc. Interspeech*, 2012.

[14] A Gersho and R M Gray, *Vector quantization and signal compression*, Springer, 1992.

[15] T Bäckström and J Fischer, "Fast randomization for distributed low-bitrate coding of speech and audio," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, January 2018.

[16] 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.

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

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

[19] Recommendation BS.1534, *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU-R, 2003.

[20] J S Garofolo, Linguistic Data Consortium, et al., *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.