Bäckström, Tom; Bouafif, Mariem; Perez Zarazaga, Pablo; Ranjit, Meghna; Das, Sneha; Lachiri, Zied

PyAWNeS-Codec: Speech and audio codec for ad-hoc acoustic wireless sensor networks

Published in:
Proceedings of the European Signal Processing Conference 2021 (EUSIPCO)

Published: 01/09/2021

Document Version
Peer reviewed version

Please cite the original version:
Bäckström, T., Bouafif, M., Perez Zarazaga, P., Ranjit, M., Das, S., & Lachiri, Z. (2021). PyAWNeS-Codec: Speech and audio codec for ad-hoc acoustic wireless sensor networks. In *Proceedings of the European Signal Processing Conference 2021 (EUSIPCO)* (European Signal Processing Conference). IEEE.

PyAWNeS-Codec: Speech and audio codec for ad-hoc acoustic wireless sensor networks

1st Tom Bäckström
Dept of Signal Processing and Acoustics
Aalto University
Espoo, Finland
https://orcid.org/0000-0002-5590-2349

2nd Mariem Bouafif Mansali
SITI laboratory
National Engineering School of Tunis
El Manar University, Tunisia
https://orcid.org/0000-0002-3971-0697

3rd Pablo Pérez Zarazaga
Dept of Signal Processing and Acoustics
Aalto University
Espoo, Finland
https://orcid.org/0000-0002-6166-9061

4th Meghna Ranjit
Dept of Signal Processing and Acoustics
Aalto University
Espoo, Finland
mailto:meghna.ranjit@aalto.fi

5th Sneha Das
Dept of Signal Processing and Acoustics
Aalto University
Espoo, Finland
https://orcid.org/0000-0002-4017-1280

6th Zied Lachiri
SITI laboratory
National Engineering School of Tunis
El Manar University, Tunisia
https://orcid.org/0000-0002-1289-5089

***Abstract*—Existing hardware with microphones can potentially be used as sensor networks to capture speech and audio signals for the benefit of better signal quality than possible with a single microphone. A central pre-requisite for such ad-hoc acoustic wireless sensor networks (ASWNs) is an efficient communication protocol with which to transmit audio data between nodes. For that purpose, we present the world's-first speech and audio codec especially designed for ASWNs, which has competitive quality also in single-channel operation. To ensure quality in the single-channel scenario, it closely resembles conventional codecs of the TCX-type, but extended with features to facilitate multi-device operation, including dithered quantization, delay estimation and compensation, as well as multi-channel post-filtering. The codec is intended to become a baseline for future research and we therefore provide it as an open-access library. Our experiments confirm that performance is in the same range as recent commercial single-channel codecs and that added devices improve quality.**

***Index Terms*—speech and audio coding, ad-hoc acoustic sensor networks, time difference of arrival estimation, delay compensation, multi-channel post-filtering**

I. Introduction

Wireless acoustic sensor networks (WASNs) can be used to sample the spatial acoustic space to gain a better signal quality than a single sensor ever could [1], [2]. Moreover, ad-hoc WASNs, viz., collections of all arbitrary available devices with microphones, can work as WASNs with very low hardware costs. For example, typical offices, meeting rooms and living rooms often have many devices with microphones and network access. Using them all, for example, in a teleconferencing scenario, could improve sound quality without added hardware costs and can allow improving the user experience by making the interface user-centric. Similarly, WASNs could be used as acoustic front-ends for speech interfaces such for smart speakers.

Recent communication codecs such as 3GPP EVS and the ETSI LC3plus codec [3], [4] are however designed to be used with a single sensor device. On the other hand, distributed source coding techniques for WASNs have been widely studied, e.g. [5]–[11], but as far as we know, none of such contributions have resulted in a publicly available implementation of a codec, which would reach a competitive performance also in a single-channel mode. In particular, most such works have studied rate-distortion theory, but have not actually implemented a quantizer and codec. Still, based on our experience with standardization [12], we argue that a codec can be successful in the market only if, in addition to some novel benefits, its performance is *at least comparable* to prior standards in terms of perceptual quality, algorithmic delay and resource consumption. There is therefore demand for an implementation of a codec specifically designed for ASWNs.

The contributions of this paper are; 1) We present, as far as we know, the first, publicly available speech and audio codec for acoustic sensor networks, whose performance is comparable with conventional communication codecs in a single channel/device configuration. 2) The proposed codec combines elements from many of our recent works, including [13]–[21], but for the first time, allows their testing in a realistic environment. 3) The proposed codec also includes a TDoA estimator and a novel delay compensation method in the MDCT-domain, which is an important part of WASN codecs [22]. We must however emphasize that many of the individual components are not state-of-the-art, but we have rather opted to use simple methods to keep the complexity of development task reasonable. In particular, we have focused on building a baseline for future experiments.

Our particular focus in this paper is a typical home scenario, where a user has a mobile or wearable device near him, such as a smartphone or smartwatch, as well as another device further away, such as a smart speaker or -TV. The characteristic trait of this scenario is that the device near the user is mobile and therefore necessarily constrained in resources such as computation capacity and battery life. The faraway device, on the other hand, is stationary and connected to a power outlet,

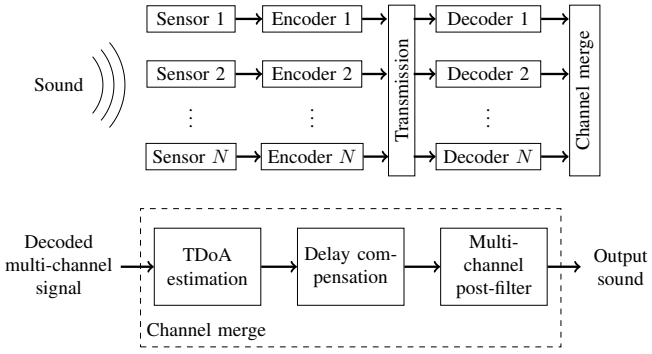

Fig. 1. Systems structure of proposed codec.

but since it is far away, its signal quality is lower. It is this dichotomy between resources and signal quality which we aim to leverage for our benefit.

II. Systems structure

The overall systems structured is illustrated in Fig. 1. For simplicity of systems design, each sensor node works independently without communication between other sensor nodes. In particular, we have not implemented authentication, on-line bitrate optimization nor other complicated interactions between sensor nodes. Each encoded bit-stream can thus also be independently decoded and the decoded quantization levels of all sensors are forwarded to the channel merge process; First, we calculate the time-difference of arrival (TDoA) between pairs of channels. We assume that the number of sensors is small such that we can calculate TDoA's for all pairs of sensors. Furthermore we assume that the signal contains only one dominant source. Second, channels are delayed to align with the most-delayed channel. Finally, channels are combined with a post-filter. In the following, we present each block in more detail.

The single-channel codec part follows the principal structure of the TCX-mode in the Enhanced Voice Services codec, standardized by 3GPP [3], which is based on the MDCT-transform with frame-length of 30 ms and -step of 20 ms [12], [23]. The windowing function is half-sine, but with a flat top to obtain low-delay overlaps between adjacent windows [12]. The spectral power envelope is modeled with a matrix transform A, such that for the power-spectrum x, we obtain the envelope parameters as $y = Ax$. The logarithms of the parameters are quantized to $\hat{y}$, exponentiated and converted to a power-envelope $\hat{s}$ with an inverse transform B as $\hat{s} = B\hat{y}$. We have chosen to use $M = 16$ envelope parameters at a sampling rate of 16 kHz. The transform matrices, A and B as well as the quantization accuracy of envelope parameters are numerically optimized to optimize bitrate [15], in difference to the traditional approach of matching average envelope quantization error [24]. Spectral whitening is implemented by dividing the spectrum by the square root of the quantized power envelope [20]. The envelope parameters are encoded with a variable bit-rate entropy coder using a multivariate Gaussian distribution following [19].

The perceptual model is a neural network approximating the perceptual model of 3GPP EVS [3]. We use an approximation because the perceptual model in EVS is defined by linear predictive analysis in the time-domain, which would introduce unnecessary computational complexity if used here. The network takes as input the quantized logarithmic envelope parameters and gives the logarithm of the relative target error magnitude e_k for each spectral component k. The network is a four-layer fully connected network where the layer sizes are $16, 50, 50, 50$ and 320, with ReLU activations for the hidden layers. It was trained with the 100,000 frames of audio files from LibriSpeech, using an Adam optimizer [25] with a learning rate of 10^{-4}, and batches of size 200.

The relative target bitrate of the kth spectral component is then $b_k = -\log_2 e_k + o_k$, where o_k is a frequency-specific bit-rate bias term. If the target absolute bitrate for the whole frame is B_{frame} and the number of bits used for the envelope is B_{envelope}, then the remaining bits for the spectral components is $B_{\text{spectrum}} = B_{\text{frame}} - B_{\text{envelope}}$. Consequently, we must offset the relative target bitrate b_k by γ such that $B_{\text{spectrum}} = \sum_k \mathrm{ReLU}(b_k + \gamma)$. Here, the bitrate of individual spectral components must be non-negative, such that we threshold $(b_k + \gamma)$'s at zero with the linear rectifying unit $\mathrm{ReLU}()$, and iteratively solve the largest possible γ such that the overall bitrate is optimally used. For the iteration we use the binomial search for a rapid and simple solution. We thus obtain the target bitrate for each spectral component.

The probability distributions of spectral components are modeled by logistic mixture distributions with 5 components, following [15]. The bit-rate bias terms o_k of all spectral components are further estimated numerically by determining the bitrate of fixed-accuracy quantization for the given logistic mixture distributions. The bias terms are trained off-line and stored in a look-up table.

As a last step of the encoder, the spectral components are quantized with uniform quantization, with a random offset (dithering) [14]. This offset is assumed to be known at the decoder and is not transmitted. In practice we can use, for example, the bitstream of the envelope coder as a seed value for a pseudo-random generator to determine the offsets. The benefit of dithering is that output energy of the codec remains non-zero also for low-bitrate components and that independent sensors have uncorrelated quantization errors [14]. The quantized spectral components are then encoded with arithmetic coding [12], [26].

The transmitted data thus consists of envelope parameters and spectral components. Observe that there is no separate gain or energy term for the spectrum, but for simplicity, it is taken to be part of the envelope model. The decoder reverses the steps of the encoder to obtain the quantization bins of spectral components. The quantized values are estimated as the expectations within respective quantization bins, where the expectations are calculated over the corresponding logistic mixture models. To obtain numerically stable expectations of the spectral components, we cannot use the analytical formula for the expectation. Therefore, for a quantization

bin $x \in [L, R]$, we approximate the expectation of spectral components by taking the mean of the cumulative probabilities of L, R, that is,

$$E\left[x \in [L,R]\right] \approx c^{-1}\left(\frac{1}{2}\left[c(L)+c(R)\right]\right), \qquad (1)$$

where $c()$ and $c^{-1}()$ are, respectively, the corresponding cumulative distribution function and its inverse.

After receiving the decoded multi-channel signal, we need to estimate the target signal by multi-channel filtering. Observe that conventional beamforming approaches are not directly applicable here since they rely on phase-rotations of the complex-valued spectrum [27], [28], whereas we have access only to a real-valued MDCT-spectrum. We therefore implement a simple delay-compensation scheme as follows.

We compute the cross-correlation function between whitened signal from multi-device signals and then estimate the TDoA as the maximum peak location with reference to the zero time lag [29]. We then compensate for the difference in delay across channels, by delaying channels to the delay of the latest-arriving signal. This delay is implemented by calculating the correlation in the basis functions of the MDCT of the target delay with two consecutive frames with actual delay. The delay is thus implemented as a mapping from two observed frames to the target frames of desired delay. Finally, the synchronized channels are merged with a classical multi-channel Wiener filter, using the average background noise plus quantization noise energies. The output signal is then obtained by an inverse MDCT [12], [23].

III. Implementation

We implemented the codec using the PyTorch machine-learning library. All optimizations were performed over the train-clean-100 set of the LibriSpeech corpus [30], at a sampling rate of 16 kHz. For optimization of the envelope transforms and quantization, as well as the logistic mixture distributions for spectral components, we used the Adam-algorithm with a single epoch and batches over single files. The mean and covariance of envelope parameters were estimated as an average over all frames in the whole training set. We have not yet incorporated an actual implementation of the arithmetic coder, but just estimated its bitrate from the cumulative probability distributions, since this is sufficient to get a qualified estimate of the sound output. For the current experiments, we implemented only a two-device version of the multi-channel decoder.

The codec is provided as a free-of-charge, open-access library[1]. Observe that we do *not* have the liberty or authority to claim that the codec would be free of intellectual property rights. As a consequence, we provide the codec only with an academic evaluation license.

[1]https://gitlab.com/speech-interaction-technology-aalto-university/pyawnes-codec

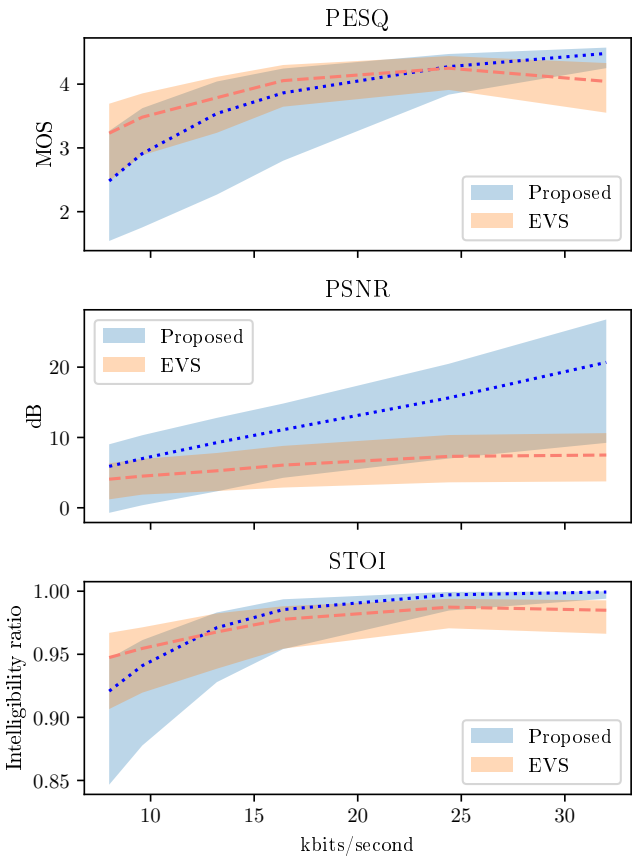

Fig. 2. Single channel quality across bitrates (8, 9.6, 13.2, 16.4, 24.4 and 32 kbit/s); dashed red and dotted blue lines express the median of EVS and proposed codecs, respectively, and the corresponding filled areas their 95 % quantiles.

IV. Experiments

To evaluate the quality of the proposed codec, we evaluate its quality in single and multi-channel scenarios. For single-channel evaluation, we compare quality to the 3GPP EVS codec [3] at bitrates 8, 9.6, 13.2, 16.4, 24.4, and 32 kbit/s. To get a fair evaluation of quality, the EVS codec was not constrained to the TCX mode, even if the proposed codec operates only in the MDCT-domain. Since EVS thus can use also its CELP-modes as well as advanced coding tools such as bandwidth extension, it has an advantage over the proposed codec [12].

As objective measures of quality, we calculated the SNR in a perceptually weighted domain (pSNR), PESQ and STOI over the test set of the LibriSpeech corpus [30]–[32]. The perceptual model used for pSNR calculations is the same as that used within the proposed codec.

Objective results of the single-channel experiment are illustrated in Fig. 2. The PESQ median scores at low bitrates are higher for EVS than the proposed codec, but at 32 kbit/s the roles are switched. The STOI measures give a similar trend, but such that the proposed codec is better already from 13.2 kbit/s and upwards. The proposed codec thus scales up

better with the bitrate indicating that the statistical model is more accurate in the proposed codec. However, the 95 % ranges are overlapping over the whole range of bitrates demonstrating that the difference in quality is relatively small. However, STOI scores are saturated, near its maximum value such that it is questionable whether this measure is meaningful.

In contrast, the pSNR median score is better for the proposed codec across all bitrates. This measure however favours the proposed codec since it uses the same perceptual model for which the codec is optimized. The linear increase in quality however again demonstrates that the proposed codec is stable and scales uniformly to varying bitrates. At low bitrates the difference in pSNR scores are again small, indicating that the quality between codecs are comparable. However, since the pSNR shows a difference between codecs so much larger than the PESQ and STOI scores, it leads to the conclusion that the perceptual model in the proposed codec should be better tuned in future work.

For multi-channel evaluation, our purpose was to evaluate a scenario where we have two microphones, one microphone close to the source (Mic 1) and second microphone further away (Mic 2). We can then assume that the wearable microphone, Mic 1, has a better input SNR than Mic 2, 30 dB and 10 dB, respectively. This corresponds to a living-room scenario where a user has a handheld smartphone or wearable microphone near him and a smart speaker further away. We further assume that the nearby microphone is resource-constrained and able to transmit only at 8 kbit/s, while the far away microphone can transmit at a higher rate of 32 kbit/s. In our comparison, we compare the two-microphone scenario with a single-microphone scenario, located at Mic 2. To make the single-microphone scenario fair, we use the same total bitrate as in the two microphone case, of 40 kbit/s.

We simulated the multi-channel recording as follows. As clean speech samples, we used the LibriSpeech test set and noise samples from the QUT-NOISE database living room scenario (LIVINGB-1) [33]. The clean speech signals were filtered with room-impulse-response (RIR) corresponding to the two different microphone locations in a room ($RT_{60} = 0.3$), following Fig. 3. To add the effect of the room acoustics to the LibriSpeech recordings, we used the Pyroomacoustics library [34]. Noise samples were selected from random locations from the entire samples. Since these noise samples already included room reverberation, they were added to the speech which already had been filtered with the room impulse response. The same noise signal was added to both channels such that the noise statistics would be similar, albeit with different SNRs as explained above. Observe that the noises will go out of synchronization, when the signals are delay-compensated.

To evaluate the results, we used the following procedure. First, the signals were time-aligned and scaled to maximize the correlation with the original source signal. The signal to distortion ratio (SDR) was then calculated for the whole signal (without windowing), where the distortions include room reverberation, background noise, quantization noise as well as estimation errors. For the perceptual SNR, similarly, we calculated the ratio of the clean input signal energy and the distortion energy, but with the same windowing, time-frequency transform and perceptual weighting as used in the codec. However, the perceptual model was calculated from the time-aligned original clean source signal, to get the most accurate weighting and fair comparison. The calculation of the SDRs and pSNRs thus follow the conventions of the source separation and speech coding communities, respectively. It should however be noted that the pSNRs thus obtained give unrealistically low values, since non-speech frames are dominated by background noise, even after the multi-channel filter, such that the frame-wise SNRs are very low. The absolute pSNR values might thus not be meaningful and we should focus only on the improvement in pSNR. Still, since perceptual weighting is central to coding applications and none of the other standard measures (SDR, SIR, SAR etc.) support such weighting, we chose to include pSNR. In calculation of PESQ and STOI scores, the only difference to above was that we scaled signals to match the original signal energy.

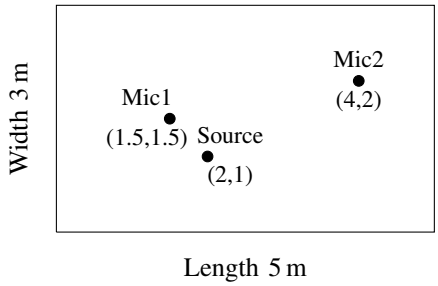

Fig. 3. Room configuration in multi-channel experiments. Coordinates of source and microphones in parenthesis.

The results of the multi-channel experiments are listed in Table I. We observe that the signal to distortion ratio (SDR) is on average better for the proposed joint estimate than the single-channel reference, by 1.74 dB. The standard deviations for the SDRs themselves are rather large indicating that the distributions are overlapping heavily. For their difference Δ, the standard deviation is much smaller, which suggests that though the SDRs have a large standard deviation, in their mutual ordering, the proposed method is on average clearly better. The same arguments apply for the perceptual SNR (pSNR), PESQ and STOI scores as well. The improvement from multi-channel processing is not particularly large, but this was to be expected as the proposed method only uses classic, rudimentary methods for merging the two channels.

V. Discussion

Distributed speech coding is attractive because users already have plenty of hardware available, which could be used to gain better audio quality and a more user-centric user-interface. To make a distributed codec commercially viable, we argue that it must give competitive quality in single-channel scenarios and provide improved quality when applied to multi-sensor scenarios. Since no such codec has been publicly available, in this work, we present an implementation of such a codec.

TABLE I

RESULTS OF THE MULTI-CHANNEL EXPERIMENT; *Joint* IS THE PROPOSED ESTIMATE OF THE SIGNAL USING BOTH CHANNELS (MIC 1 AND MIC 2), *Single* IS FROM MIC 2 WITH SAME TOTAL BITRATE, AND $\Delta = (\text{JOINT} - \text{SINGLE})$ IS THEIR DIFFERENCE. FOR EACH MEASURE, THE MEAN IS LISTED AS WELL AS THE STANDARD DEVIATION IN PARENTHESIS.

	Joint	Single	Δ
SDR (dB)	$\mathbf{-3.60\,(3.39)}$	$-5.34\,(3.63)$	$1.75\,(1.19)$
pSNR (dB)	$\mathbf{-4.95\,(3.24)}$	$-6.94\,(3.66)$	$1.99\,(1.37)$
PESQ	$\mathbf{1.29\,(0.11)}$	$1.20\,(0.08)$	$0.09\,(0.06)$
STOI	$\mathbf{0.66\,(0.07)}$	$0.58\,(0.05)$	$0.08\,(0.05)$

The structure of the codec is based on a similar structure as the TCX-mode in the 3GPP EVS codec [3], but improved with many of our recent contributions. In particular, we use dithering to make the quantization errors across devices uncorrelated and improve the entropy model with end-to-end optimization [14], [15]. Novelties include also the DNN-based approximation of the perceptual model as well as the bit-assignment algorithm for quantization based on the perceptual model [21].

The presented codec is intended to be a baseline codec for further experiments. As such and for brevity of this paper, we have not included several modules which are known to improve quality, including models of the fundamental frequency such as [35], 1-bit quantization [14], [36], temporal noise shaping (TNS) [37] and beamforming [28].

Our experiments demonstrate that quality of the proposed codec is comparable to 3GPP EVS in terms of objective criteria in a single-channel mode. In a multi-device scenario, our experiments show that the codec can improve quality in comparison to a typical single-channel scenario with the same total bitrate. Subjective evaluation as well as more extensive multi-device evaluation is left for further study.

REFERENCES

[1] A Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *Comm Vehi Tech Benelux (SCVT), IEEE Symp*, 2011.

[2] T Bäckström, "Speech coding, speech interfaces and IoT – opportunities and challenges," in *52nd Asilomar Conference on Signals, Systems and Computers*, 2018.

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

[4] *ETSI TS 103 634 – Digital Enhanced Cordless Telecommunications (DECT); Low Complexity Communication Codec plus (LC3plus)*, 2019.

[5] O Roy and M Vetterli, "Distributed compression in acoustic sensor networks using oversampled A/D conversion," in *Proc. ICASSP*. IEEE, 2006, vol. 4.

[6] O Roy and M Vetterli, "Rate-constrained collaborative noise reduction for wireless hearing aids," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 645–657, 2009.

[7] J Amini, R C Hendriks, R Heusdens, M Guo, and J Jensen, "Rate-constrained noise reduction in wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1–12, 2019.

[8] H Dong, J Lu, and Y Sun, "Distributed audio coding in wireless sensor networks," in *Comp Intel Sec, 2006 Int Conf*. IEEE, 2006, vol. 2, pp. 1695–1699.

[9] J Østergaard and M S Derpich, "Sequential remote source coding in wireless acoustic sensor networks," in *Proc. EUSIPCO*. IEEE, 2012, pp. 1269–1273.

[10] A Majumdar, K Ramchandran, and I Kozintsev, "Distributed coding for wireless audio sensors," in *Appl Sig Proc Audio Acou, IEEE Workshop*, 2003, pp. 209–212.

[11] Z Xiong, A D Liveris, and Y Yang, "Distributed source coding," *Handbook on Array Processing and Sensor Networks*, pp. 609–643, 2009.

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

[13] 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.

[14] T Bäckström, J Fischer, and S Das, "Dithered quantization for frequency-domain speech and audio coding," in *Proc. Interspeech*, 2018, pp. 3533–3537.

[15] T Bäckström, "End-to-end optimization of source models for speech and audio coding using a machine learning framework," in *Proc. Interspeech*, 2019.

[16] S Das and T Bäckström, "Postfiltering using log-magnitude spectrum for speech and audio coding," *Proc. Interspeech 2018*, pp. 3543–3547, 2018.

[17] S Das and T Bäckström, "Postfiltering with complex spectral correlations for speech and audio coding," *Proc. Interspeech 2018*, pp. 3538–3542, 2018.

[18] S Das and T Bäckström, "Enhancement by postfiltering for speech and audio coding in ad-hoc sensor networks," *JASA Express Letters*, 2021.

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

[20] 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.

[21] Meghna Ranjit, "Efficient application of perceptual models in speech and audio coding," M.S. thesis, Aalto University, Espoo, Finland, 2020.

[22] R Lienhart, I Kozintsev, S Wehr, and M Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *Proc. ICASSP*, 2003, vol. 4, pp. IV–840.

[23] M Bosi and R E Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publs., 2003.

[24] K K Paliwal and B S Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 1, pp. 3–14, 1993.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[27] J Benesty, M Sondhi, and Y Huang, *Springer Handbook of Speech Processing*, Springer, 2008.

[28] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, *Audio source separation and speech enhancement*, John Wiley & Sons, 2018.

[29] M Bouafif and Z Lachiri, "Src-num-tdoa: Multiple speech sources' number and their tdoa estimation from a stereo recorded mixture," *SoftwareX*, vol. 5, pp. 234–242, 2016.

[30] V Panayotov, G Chen, D Povey, and S Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[31] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[32] C H Taal, R C Hendriks, R Heusdens, and J Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19 (7), no. 7, pp. 2125–2136, 2011.

[33] D B Dean, S Sridharan, R J Vogt, and M W Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proc. Interspeech*, 2010.

[34] R Scheibler, E Bezzam, and I Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*. IEEE, 2018, pp. 351–355.

[35] Sneha Das, Tom Bäckström, and Guillaume Fuchs, "Fundamental frequency model for postfiltering at low bitrates in a transform-domain speech and audio codec," *Proc. Interspeech 2020*, pp. 2837–2841, 2020.

[36] J Fischer, *Contributions to speech and audio coding for single- and multi-device scenarios*, Ph.D. thesis, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany, 2020.

[37] J Herre and J D Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Proc AES Convention 101*, Nov. 8–11 1996.