Microphone Array Coding Preserving Spatial Information for Cloud-based Multichannel Speech Recognition

Daniel T. Jones[1], Dushyant Sharma[2], Stanislav Yu. Kruchinin[3], and Patrick A. Naylor[1]

[1]Imperial College London, UK
[2]Nuance Communications Inc., USA
[3]Nuance Communications GmbH, Austria
Email: d.jones20@imperial.ac.uk

***Abstract*—An efficient method of coding multichannel signals from a microphone array is presented. This is advantageous for cloud-based audio processing, such as Direction-of-Arrival (DOA) and Automatic Speech Recognition (ASR). The method operates by encoding separately the signal information - using a reference signal - and the spatial information - using Relative Transfer Functions (RTFs). Results for ASR and DOA performance are presented for the proposed codec in comparison to a baseline multichannel implementation of the Opus codec. Both stationary and time-varying acoustic scenarios have been included in the tests. The proposed RTF-based codec is shown in our experiments to preserve spatial information in the array signals whereas the baseline codec does not. The proposed codec is also shown to outperform the baseline on the ASR task at low bit rates in the region of 6 kbits per second per channel.**

***Index Terms*—Audio encoding, relative transfer functions, multichannel audio, microphone arrays, adaptive filters.**

I. Introduction

Microphone arrays combined with suitable processing algorithms are highly effective at noise reduction and dereverberation. Because of this they are a crucial component within the front-end of multichannel ASR systems, which are often sensitive to the high variability which can be found in noise sources and reverberation. A commonly used technique for noise reduction that is paired with microphone arrays is acoustic beamforming [1], where the channels of a multichannel signal are combined with a filter-and-weighted sum to give an enhanced single channel signal. In real-world high performance ASR systems and services, it is often the case that the ASR is performed in the cloud rather than on the device, requiring the transmission of the speech signals. One option is that beamforming is applied on the array before the enhanced single channel signal is transmitted to a cloud-based ASR. This is effective and has the advantage that only a single channel of audio needs to be transmitted, however, it is restrictive as it removes spatial information which can be used for further processing tasks such as DOA estimation [2] and diarization [3], [4]. The option furthermore makes it impossible to apply the latest multichannel end-to-end ASR paradigms in the cloud, since only the (albeit enhanced) signal information is transmitted whereas the spatial information embodied in the multichannel signals never reaches the cloud-based processing.

Conventional audio codecs such as Opus [5], [6] have been shown to be effective for use with ASR systems [7]. However, such codecs have not so far been designed to fully leverage the high correlations found in microphone array signals, either failing to use them at all or destructively encoding the phase/spatial information in an unrecoverable way. Opus is a hybrid codec which combines a modified version of the linear prediction based SILK codec [8], which is optimised for speech, with the Modified Discrete Cosine Transform (MDCT) based CELT codec [9], an all-purpose codec. This class of perceptual codec is not specifically designed for use in ASR applications since the spatial information present in multichannel signals is often distorted or removed completely in favor of preserving perceptual quality. Recently however, a method to optimise the Opus codec for microphone array coding with ASR in mind was proposed in [10]. Settings within the standard Opus codec were modified to optimise performance for beamforming and ASR, such that the signals could still be decoded by the standard Opus decoder.

The first step taken in [10] towards overcoming this problem was to disable a number of features of the SILK and CELT parts of Opus that were detrimental to ASR/beamforming performance. To this end, the CELT intensity stereo and folding features were disabled and set to their lowest allowable amount respectively. Furthermore, the SILK codec conventionally encodes the signal into mid and side channels, with the option to discard the side channel to optimise for perceptual quality if there is a low bitrate allocation. Later, in [10] a method for waveform matching is proposed where the split is optimised in order to preserve the phase response. These three changes together represent one step in modifying Opus for multichannel audio coding.

The second step proposed in [10] is to pre-process the input signals to the Opus codec using a spatial Discrete Fourier

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

Transform (DFT). Due to the high correlation between neighboring channels, [10] states that any orthogonal transformation could decorrelate the signals thus reducing the amount of redundant information. The DFT was chosen as it is not signal or array geometry dependent, giving greater flexibility in implementation. [10] showed that the spatial DFT was decorrelating the signals by showing a table of the power of the raw channels in comparison to the DFT channels.

When these two steps were combined together the new Opus codec consistently outperformed the alternative of independently encoding all audio channels with their own Opus encoder, with a maximum Word Error Rate Reduction (WERR) 6.4% when the two Opus codecs were compared at 16 kbps using 7-channel microphone array signals.

An alternative method for Microphone Array Coding (MAC), known as Relative Transfer Function Microphone Array Coding (RTF-MAC), was proposed in [11] which allows for microphone array data to be transmitted with a minimal increase in bitrate. This was achieved by expressing the multichannel signal from M microphones in terms of a reference channel, compressed with a conventional single-channel audio codec, $M-1$ Relative Transfer Functions RTFs, and (optionally) $M-1$ RTF-coding residual signals. The RTFs, which in [11] are estimated using the Improved Proportionate Normalized Least Mean Squares (IPNLMS) adaptive filtering algorithm [12], [13], encode all the non-reference channels, which under certain conditions are a compact representation, particularly when the acoustic scenario is close to being stationary such that neither the sound source signals nor the microphone array move. A block diagram for the system can be seen in Fig 1.

The previous system was evaluated in [11] in terms of time domain Mean Square Error (MSE), frequency domain error, a perceptual metric Perceptual Evaluation of Speech Quality (PESQ) [14], [15] and by evaluating the Word Error Rate (WER) of a pre-trained ESPnet2 [16], [17] ASR system with an IPNLMS based RTF-MAC [11] and a BeamformIt beamformer [18] front-end. The codec was previously compared to a multichannel Opus codec where each channel was encoded independently, which was prior to the publication of [10]. The system showed improved performance, outperforming Opus [5], [6] particularly at low bitrates where the codec extended the lower bound of bitrates that the ASR could successfully operate at. For example, when paired with the same ASR back-end, the RTF-MAC achieved approximately 10% WER for bitrates around 3−5 kbps/ch while Opus showed a WER of around 26% for the bitrates around 8 kbps/ch.

The present paper progresses the previous work carried out in [11] in a number of ways. First of all, the codecs were again used as front-ends to ASR systems, but in this research we trained multichannel Self Attention Channel Combinator (SACC) ASR systems [19], [20], allowing us to evaluate how different codecs preserve the spatial information used by multichannel ASR. Furthermore, new configurations of the RTF-MAC have been tested which periodically update the RTFs every 2 s. This allows the RTF-MAC to adapt to time-varying acoustic scenarios in a manner that can be matched to the expected level of non-stationarity in the application of interest. In addition, DOA estimation was performed on the decoded signals using GCC-PHAT [21] as an example of a commonly deployed 'workhorse' baseline method, and the error between the estimated and the ground truth DOA was calculated. This was done to investigate the viability of the multichannel signal for further processing in the cloud such as spatial processing and scene analysis.

II. Methodology and Experimental Setup

A. Periodically Updated RTF-MAC

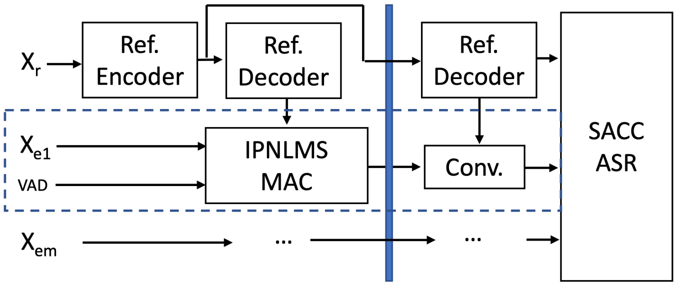

Fig. 1: Block diagram of RTF-MAC . The dashed box indicates the parts that need to be repeated for the M^{th} non-reference channel X_{em}

Figure 1 shows a block diagram of how the proposed RTF-MAC was implemented with a SACC-based multichannel ASR back-end. The RTF-MAC properties that were constant for all configurations were determined empirically and were set as follows: The IPNLMS leaning rate, denoted by μ, was set to 0.1. This parameter controls the trade-off between quick convergence and converged accuracy. The α parameter of IPNLMS which controls the sparsity of the estimated RTFs [12], [13] was set to $\alpha = 0.5$. An $M = 8$ channel microphone array was employed in our tests and the reference channel chosen in all cases was channel 4, as this is located more centrally in the array, allowing for shorter RTF filters to be used. In the particular configurations evaluated in this work, all RTF-encoding residuals were ignored, as is appropriate when targeting low bitrate configurations, since it was shown in [11] that the addition of residuals did not significantly improve ASR WER. Finally, the system uses the output of a Voice Activity Detector (VAD) in order to freeze the adaptation of the IPNLMS during non-speech periods. The VAD used in these test was [22] and was calculated offline by calculating the VAD of clean reference signals and subsequently time aligning to account for the propagation delay found in the recorded signals.

Here, we present two RTF-MAC configurations. C2 uses the highest performing combination of parameters found in the grid search with a 256-tap RTF being updated every 2 s and a reference channel encoded at 32 kbps. C100 matches a configuration found in [11], with a 16 kbps reference channel and a 512-tap RTF, with the only difference being that the RTF was also updated every 2 s. This is crucial as it allows for processing of non-stationary acoustic scenarios commonly found in real-world applications, such as conversational speech. The C2 and C100 have total 8-channel bitrates of 46.9 and 44 kbps, respectively. These configurations were found by running a sweep search with differing IPNLMS algorithm values using a ESPnet2 model pre-trained on the LibriSpeech data [23]. The configurations with the lowest WER within the desired bitrate region were then chosen for further investigation. More details of the system used for the parameter sweep can be seen in [11].

B. Multichannel Opus

In both cases, we use the SILK part of the OPUS codec due to its better performance for speech at low bitrates [8] and a constant bitrate of 6 kbps/ch giving the total bitrate of 48 kbps. In the first test, each channel was encoded separately (MC OPUS SEP) and in the second test we enabled stream coupling for all channels (MC OPUS CUP) [24].

C. ASR

The ASR experiments are based on a multichannel SACC front-end [19] to an attention-based, ContextNet [25] encoder and a single layer LSTM decoder. The SACC front-end combines the magnitudes of signals from M microphones in the Short Time Fourier Transform (STFT) domain and is trained jointly with the ASR system. For all experiments, the ASR systems were trained for 90 epochs and a matched condition trained model was used to evaluate the performance potential of each codec.

III. Data

The ASR systems were trained using 460 hours of simulated multichannel data based on clean speech from LibriSpeech [23] and the English partition of Mozilla Common Voice [26], selected using the NISA [27] method as described in [20]. The multichannel simulation was performed by convolving the clean speech utterances with 8 channel Room Impulse Responses (RIRs) simulating Uniform Linear Array (ULA) with 33 mm inter-microphone spacing and a number of directional sources placed at random positions in a large set of rooms (T60 in range 0.3 – 0.8 s) using the Image method [28]. Following convolution with the RIRs, ambient noise was added in an SNR range of 5 to 25 dB, plus 45 dB SNR white noise to simulate microphone self-noise. The gain of each channel was also augmented in the 0.1 to 2 dB range for each utterance and finally, the overall gain was adjusted, randomly in the −1 to −15 dBFS range.

The evaluation was performed with two test sets. The first one was a playback recording of a subset of 500 utterances from the clean-test partition of the LibriSpeech corpus in a typical office. The utterances were recorded with an 8-channel ULA mounted on a wall and played back from an artificial mouth simulator placed in four positions (denoted as P1–P4). The ground truth DOAs for each position can be seen in Table II. In addition to this playback data, we also created a test set by convolving the 500 clean Libri utterances with simulated RIRs representing 20 rooms and 25 source receiver positions (8 channel ULA) covering a T60 range of 0.2 – 0.4 s. Subsequently, 30 dB of ambient noise was added to the data. This simulated test set thus represents clean and mildly reverberant data with a large coverage of azimuth (from the centre of the ULA to each source, covering 20 to 160°). We use this simulated test set primarily for the DOA estimation accuracy experiments.

IV. Evaluation

The codecs were evaluated using two main metrics: the WER of the codec-ASR system and the DOA estimation error for different positions. We include an explicit evaluation of the DOA estimation because, like many ASR systems, the SACC front-end uses only magnitude spectral information, albeit multichannel. In contrast, many speech processing applications require use of the phase information such as for source localization.

A. DOA evaluation

The DOA of the signal was estimated using the previously identified GCC-PHAT baseline [21] which was applied to the uncompressed signals and then compared to the decoded signals. Following this, the magnitude of the differences between the DOAs estimated using the compressed and uncompressed signals were calculated. Estimates generated and evaluated using the uncompressed signals are denoted in the following by UC. The DOA was calculated using GGC-PHAT operating between channels 1 and 8. GCC-PHAT was calculated in the STFT domain with a window size of 2048 samples, a hop size of 512 samples, zero padding to 16384 samples for Inverse Short Time Fourier Transform (ISTFT), a smoothing factor of 0.32, and with a bandwidth limited to 0.4−7.8 kHz. The speech data was sampled at a rate of 16 kHz. The DOA estimations were filtered by energy-based VAD with an energy threshold of −75 dB.

V. Results and Discussion

Table I shows the ASR results for different codec-ASR systems when tested using the Libri playback test set in terms of WERs and WERRs.

Method	Bitrate (kbps)	WER (%)	WERR (%)
MC OPUS SEP	48	14.9	0
MC OPUS CUP	48	12.9	13.4
RTF-MAC C100	44	12.0	19.5
RTF-MAC C2	46.9	10.6	28.9

TABLE I: WER and WERR of different codec-ASR systems evaluated on the Libri playback test set

RTF-MAC clearly achieves improved performance in comparison to both configurations of the Opus codec with RTF-MAC C2 achieving a WERR of up to 28.9% relative to the MC OPUS SEP, and OPUS MC CUP achieving a 13.4% WERR. This indicates that the RTF-MAC-ASR system is able to retain significantly more information which is useful to the multichannel SACC ASR system. [10] showed a maximum WERR of 6.7% for Opus with waveform matching and the spatial DFT in comparison to Opus encoding each channel independently. Although this is significantly smaller than the 28.9% WERR achieved here by the RTF-MAC system and 13.4% WERR achieve by MC OPUS CUP, it should be noted that the differences in bitrate and ASR implementations here make a direct comparison difficult. For this reason we intend to implement the encoder in [10] for comparison in the near future.

Table II shows the ground truth DOA and the error in the DOAs that were estimated using the GCC-PHAT algorithm operating on the uncompressed signals (UC Error). Additionally it shows the magnitude of the estimation errors in degrees for all the DOAs estimated from the decoded signals, relative to the DOA estimated from the unencoded signals. As expected, DOAs estimated from the unencoded signals have the smallest estimation errors when averaged over all positions. When comparing the Opus codec and the RTF-MAC, it is clear to see that the RTF-MAC offers improvements. In the case of RTF-MAC, an increase in the DOA error by 7.4° relative to the unencoded signal on average. The Opus was shown to increase the error by 19.8° on average. However, this average is partly misleading as the DOAs estimated by Opus in this configuration were always between 89.6 and 90.7°, indicating that almost all phase information was removed in this configuration. This is confirmed by random fluctuation of angle around 90° in the repeated calculations. Therefore, instances where the Opus DOA errors appear to be lower, such as with P1–P3 are actually instances where the ground truth angle is closer to the 90° which is the angle always estimated for the Opus codec. This shows that Opus does not encode the spatial information but that with RTF-MAC it is possible to extract spatial features. Another notable point is that for positions P2 and P4 using the RTF-MAC codec matches or reduces the magnitude of the DOA estimation error compared to using uncompressed signals. This is due to the RTFs being a desirably ineffective codec for noise signals.

Position Method	P1	P2	P3	P4	Mean over P1-P4
Ground Truth	92	69	115	29	-
UC Error	1.8	1.3	7.0	7.1	4.3
MC OPUS SEP vs UC Error	3.9	20.1	18.4	54.6	24.3
MC OPUS CUP vs UC Error	0.1	21.6	11.8	47.5	20.3
RTF-MAC C2 vs UC Error	4.6	**1.3**	18.3	**5.2**	7.4

TABLE II: Mean absolute errors of DOA estimation in degrees for different codec-ASR systems and positions with the Libri playback test set

Figure 2 shows the box plots of DOA errors for the considered configurations of codecs. It is clear that the RTF-MAC configurations again give the most successful DOA estimates with errors comparable to the uncompressed signal, whereas the DOA estimation errors from Opus encoded signals spread over a broader range of angles showing that Opus does not preserve the spatial information. Notably, the RTF-MAC again slightly improves the DOA estimation performance with our simulated data set. The bias and standard-deviation of the data displayed in the box plots can be seen in Table III. This confirms that using the RTF-MAC codec reduced the bias and standard deviation of the DOA estimation. This is most likely due to the tendency of the RTF estimator to remove the noise that is not spatially correlated. This was verified by calculating the noise levels of the signal using NISA as shown in Table IV.

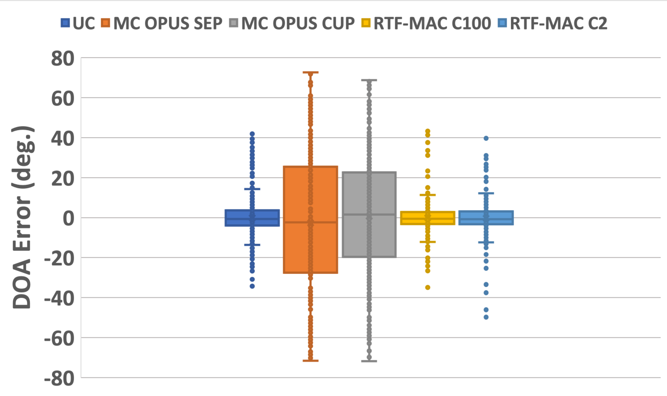

Fig. 2: DOA error for different codecs and their configurations tested on the simulated test set

Method	DOA Bias	DOA σ
UC	-0.13	10.4
MC OPUS SEP	-2.66	35.27
MC OPUS CUP	0.67	29.37
RTF-MAC C100	**-0.10**	**8.23**
RTF-MAC C2	-0.33	**8.25**

TABLE III: Bias and standard deviation σ of the DOA estimates for the simulated test set processed with each codec

Method	C50 (dB)	SNR (dB)	PESQ
UC	21.5	27.5	3.2
MC OPUS SEP	18.7	24.8	2.9
MC OPUS CUP	17.1	26.6	3.0
RTF-MAC C100	**23.4**	26.2	3.2
RTF-MAC C2	**23.3**	26.8	3.3

TABLE IV: Signal parameters of the decoded test set estimated using NISA

VI. Conclusions

We developed a codec for microphone arrays that supports multichannel spatial infromation suitable for acoustic signal processing and and cloud-based speech recognition. The codec explicitly codes the spatial information using RTFs. The proposed codec combines the RTF-MAC with multichannel ASR and has been shown to recude the WER in comparison to a system where Opus is employed as the codec in the front-end. Additionally, we have shown that the decoded signals obtained from the RTF-MAC are useful in the task of DOA estimation, with errors very similar to uncompressed data. Furthermore, the codec performs periodically updating the RTFs which makes it applicable to real-world scenarios, such as conversational speech.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, "DOA estimation of audio sources in reverberant environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2016, pp. 176–180.

[3] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[4] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A Review of Speaker Diarization: Recent Advances with Deep Learning," *arXiv:2101.09624 [cs, eess]*, Nov. 2021.

[5] J. M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec. RFC 6716," https://www.ietf.org/rfc/rfc6716.txt.

[6] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the Opus codec," *arXiv:1602.04845 [cs]*, Feb. 2016.

[7] A. Khare, S. Sundaram, and M. Wu, "Multi-channel Acoustic Modeling using Mixed Bitrate OPUS Compression," *arXiv:2002.00122 [cs, eess]*, Jan. 2020.

[8] K. Vos, S. S. Jensen, and K. V. Soerensen, "SILK Speech Codec," Internet Engineering Task Force, Internet Draft draft-vos-silk-02, Sept. 2010.

[9] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10-ms delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 58–67, Jan. 2010.

[10] L. Drude, J. Heymann, A. Schwarz, and J. M. Valin, "Multi-channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 1669–1673.

[11] D. Jones, D. Sharma, S. Y. Kruchinin, and P. A. Naylor, "Spatial Coding for Microphone Arrays Using Ipnlms-Based RTF Estimation," *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, pp. 76–80, 2021.

[12] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice Hall, 1996.

[13] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 508–518, Sept. 2000.

[14] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Int. Telecommun. Union (ITU-T), Recommendation P.862, Nov. 2003.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, 2001, pp. 749–752 vol.2.

[16] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," *arXiv:1804.00015 [cs]*, Mar. 2018.

[17] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, S. Karita, C. Li, J. Shi, A. S. Subramanian, and W. Zhang, "The 2020 ESPnet update: New features, broadened applications, performance improvements, and future plans," *arXiv:2012.13006 [cs, eess]*, Dec. 2020.

[18] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 7, pp. 2011–2022, Sept. 2007.

[19] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Lainez, and L. Milanovic, "Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-field Speech Recognition," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 3840–3844.

[20] D. Sharma, R. Gong, J. Frosburgh, S. Y. Kruchinin, P. A. Naylor, and L. Milanovic, "Spatial Processing Front-End for Distant ASR Exploiting Self-Attention Channel Combinator," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022 (to appear).

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[22] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[24] "Opus: Opus Multistream API," https://www.opus-codec.org/docs/opus_api-1.3.1/group__opus__multistream.html.

[25] W. Han, Z. Zhang, Y. Zhang, J. Yu, C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Oct. 2020.

[26] "Mozilla Common Voice," https://commonvoice.mozilla.org/.

[27] D. Sharma, L. Berger, C. Quillen, and P. A. Naylor, "Non-intrusive estimation of speech signal parameters using a frame-based machine learning approach," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 446–450.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.