Multi-channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget

Lukas Drude, Jahn Heymann, Andreas Schwarz, Jean-Marc Valin

Amazon.com

{drude, jahheyma, asw, jmvalin}@amazon.com

Abstract

Automatic speech recognition (ASR) in the cloud allows the use of larger models and more powerful multi-channel signal processing front-ends compared to on-device processing. However, it also adds an inherent latency due to the transmission of the audio signal, especially when transmitting multiple channels of a microphone array. One way to reduce the network bandwidth requirements is client-side compression with a lossy codec such as Opus. However, this compression can have a detrimental effect especially on multi-channel ASR front-ends, due to the distortion and loss of spatial information introduced by the codec. In this publication, we propose an improved approach for the compression of microphone array signals based on Opus, using a modified joint channel coding approach and additionally introducing a multi-channel spatial decorrelating transform to reduce redundancy in the transmission. We illustrate the effect of the proposed approach on the spatial information retained in multi-channel signals after compression, and evaluate the performance on far-field ASR with a multi-channel beamforming front-end. We demonstrate that our approach can lead to a 37.5 % bitrate reduction or a 5.1 % relative word error rate (WER) reduction for a fixed bitrate budget in a seven channel setup.

Index Terms: compression, beamforming, far-field ASR

1. Introduction

Far-field ASR is machine recognition of speech spoken at a distance from a microphone or microphone array [1]. This research field has gained popularity due to its widespread application in digital home assistants [2, 3, 4] and research challenges, e.g. [5]. A common and highly effective approach to improve far-field ASR is beamforming [6], which requires multiple channels and exploits inter-channel phase and level differences to spatially focus on the desired speaker.

To further improve recognition performance, applications often exploit the computational capabilities of the cloud to run large ASR models. In order to leverage multiple channels in these scenarios, there are two options and both come with different trade-offs:

1. The recording device itself is responsible for the signal processing and combines the channels into an enhanced signal which is then transmitted to the cloud for recognition.

2. The device just records the signal and sends all or selected channels to the cloud for further processing.

The first option minimizes the transmitted amount of data but limits the possible signal processing to the devices compute budget and creates a boundary between the signal processing and acoustic model. In contrast, the second option does not have this boundary, can combine both stages [7, 8], and is more flexible in terms of compute budget and methods. However, transmitting multiple signals requires more bandwidth which is often very limited for a large number of users. In order to save bandwidth, the device can compress the signals before transmission. The challenge in the compression of multi-channel signals is that the data rate needs to be sufficiently reduced to prevent impact on the latency perceived by the user, while ensuring that spatial information is preserved in order to allow cloud-side signal enhancement to be effective.

We here focus on Opus compression [9] due to its widespread use in voice over IP (VoIP) and ASR applications [10]. While Opus has been found to be effective for single- and multi-channel ASR applications [11], it is in some sense mismatched to the task at hand. First, like all lossy speech and audio compression methods, Opus has been optimized for subjective quality, not for ASR or subsequent spatial filtering. Second, Opus does not have dedicated support for efficient compression of highly correlated multi-channel signals.

Quite a few studies assess Opus compression in terms of human perception, e.g., in [12] the authors perform perceptual listening tests when Opus is applied in a VoIP setting but do not address the far-field ASR setup. A perceptual quality assessment in terms of speech quality and localization accuracy was performed in [13], where ambisonic representations were compressed with Opus where the authors mainly experimented with different bitrates but did not analyze coding features more targeted to spatial fidelity.

With focus on ASR applications, [14] analyzes the impact of different compression algorithms on a single-channel acoustic model (AM) and concludes that using a compression algorithms as a data augmentation step in the training phase of an AM can lead to a more robust model. Closely related to the document at hand, Khare et al. analyzed the impact of Opus compression on multi-channel AMs [11] with the main findings that multi-channel data is preferable over single-channel transmission with a higher bitrate. They further report gains when retraining the AM, which we do not repeat here. However, the authors leave it open how compression impacts a beamforming front-end and provide little insight into the spatial properties of the compressed signals. Their investigated ASR system combines the multi-channel processing and senone classification into one model capable of incorporating compression artifacts during training as their experiments in Section 5 show. In contrast, analytic beamforming, as analyzed here, relies on unaltered phase information and, therefore, requires different treatment.

This contribution analyzes the impact of Opus compression on a combined beamforming and ASR system, provides guidelines on optimizing compression parameters for far-field ASR, and proposes a microphone-independent multi-channel coding scheme.
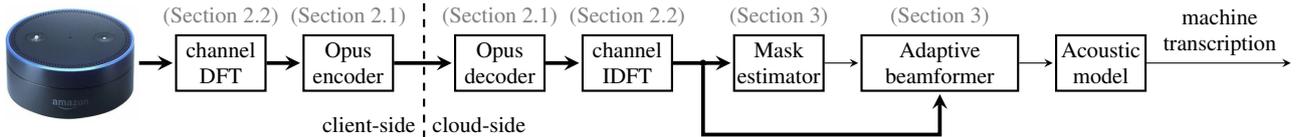
Figure 1: *Overview of a beamforming-based far-field ASR system. Compression is applied to the multi-channel signal to enable all other signal processing in the cloud. The visualization contains the proposed DFT across channels as a transform coding scheme.*

2. System description

The overall far-field ASR pipeline including a neural mask-based beamforming front-end is illustrated in Figure 1: A microphone array (here consisting of 7 channels) records far-field audio impaired by reverberation, background noise, and potentially non-device-directed speech. All channels are then compressed as detailed in Section 2.1. Section 2.2 details the proposed discrete Fourier transform (DFT)-based coding scheme. A neural mask estimator or a neural voice activity detection (VAD) is used to control an analytic beamformer (see Section 3). Finally, transcriptions are obtained with a recurrent neural network transducer (RNN-T) AM (see Section 3) [15].

2.1. Opus compression and proposed optimization

The Opus codec aims at real-time speech or music applications [9, 16]. It originates from Skype's SILK codec [17], which relies on linear predictive coding [18] and Xiph.Org's CELT codec [19], which uses a modified discrete cosine transform. The Opus codec allows to encode with SILK, with CELT, or with a hybrid of both within which lower frequencies dominated by speech are using SILK while higher frequencies fall back to CELT with specifics depending on selected parameters such as bitrate, the number of channels, or the audio content [16]. The encoder can then dynamically decide between both modes. Depending on externally provided media type flags one can bias these decisions towards either SILK or CELT.

While Opus has been found to be effective for single- and multi-channel ASR applications [11], it has been optimized for subjective quality, not for ASR accuracy. Moreover, its channel coupling has been optimized for human spatial perception, where the inter-aural phase difference (IPD) [20] is only noticeable at lower frequencies. This is why the CELT coding mode may make use of intensity stereo past a certain bitrate-dependent threshold frequency. Frequency bands coded with intensity stereo share the same phase and spectral details and only differ between channels by the energy (inter-aural intensity difference). Thus, intensity stereo can severely degrade the performance of beamforming algorithms, which rely on accurate phases. Similarly, when not enough bits are available, Opus relies on spectral folding to *fill* spectral bands with perceptually acceptable content. Again, while it is a good perceptual trade-off (allowing for more bits at lower frequency), the feature impairs beamforming algorithms.

Opus does not have dedicated multi-channel support. However, it supports joint compression of two channels [9]: In the SILK coding algorithm – optimized specifically for voice – coupled channels are split into a mid (average) and a side (difference) signal to decorrelate the channels. While the decomposition preserves the phase, the encoder will sometimes choose to discard the spatial information (side signal) in favor of reducing perceptible coding noise in the mid signal. This destroys useful phase information for a beamformer.

For these reasons, we propose to disable CELT intensity stereo and reduce the use of folding to the minimum permit-

ted by the Opus bitstream. For SILK, the bit allocation split between mid and side is set to optimize phase encoding rather than perceptual quality. In the following this is denoted as *waveform matching*[1]. Since this option only changes the encoder, the result can still be decoded with any compliant decoder.

2.2. Proposed transform coding scheme

Motivated by the high correlation between the channels in a closely spaced microphone array, one may look for an orthogonal decomposition to avoid coding redundant information. For spherical or hemispherical microphone arrays, orthogonal decompositions are known [21, 22]. However, given that all other signal components in this work are independent of the microphone array geometry, such additional restrictions on array geometry are undesired. Another option is spatial whitening [23] which can be achieved by a Karhunen-Loève Transform (KLT) [24]. However, this transform is signal-dependent and requires the estimation of a spatial correlation matrix and potentially updating it online due to the nonstationary nature of the spatial signal characteristics.

To avoid the aforementioned drawbacks while yielding some decorrelation, we propose to apply a D-point DFT across channels (note that this is not a DFT along time):

$$X_k(t) = \sum_{d=0}^{D-1} x_d(t) \cdot e^{-j2\pi dk/D}, \qquad (1)$$

where t is the time index, d is the channel index, and k is the DFT bin index. In contrast to spherical decomposition and spatial whitening this approach does not make strong assumptions on the array geometry and does not depend on the signal statistics themselves, yet still achieves a significant amount of decorrelation (compare Table 1 for a power distribution over channels). Since the DFT of a real-valued signal is even symmetric, i.e., $X_k(t) = X^*_{-k \bmod D}(t)$, we effectively obtain a real-valued DC component $X_0(t)$ and, when D is odd, $(D-1)/2$ complex-valued non-redundant channels. This can be seen as a generalization of mid/side stereo to an arbitrary number of channels. Further, it can be interpreted as an approximation of the KLT coding scheme [24]. Finally, after Opus decoding, the original channels can be reconstructed using the corresponding inverse DFT over channels corresponding to Equation 1.

In the experimental section we further demonstrate that the joint compression of the real and imaginary parts of one DFT bin, i.e., $\mathrm{Re}\{X_d\}$ and $\mathrm{Im}\{X_d\}$, is beneficial.

Table 1: *Signal power distribution across physical and DFT channels. All values are averaged over 1000 examples and listed in %.*

Index	0	1	2	3	4	5	6
Physical channels	13.5	14.1	13.6	15.6	16.4	14.6	12.1
DFT channels	87.1	3.9	1.4	1.1	1.1	1.4	3.9

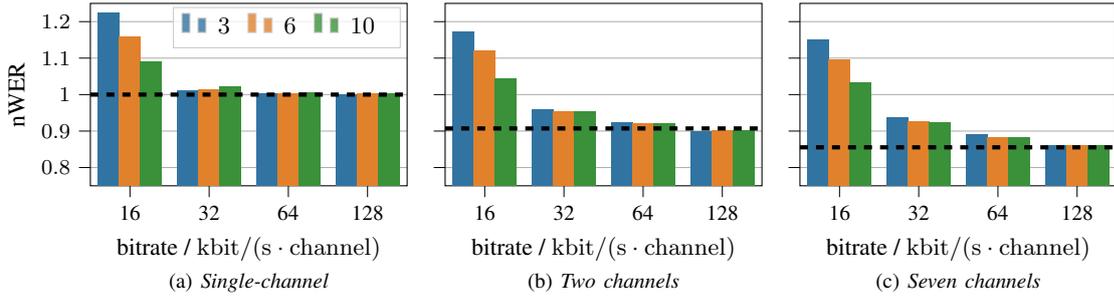[1] https://gitlab.xiph.org/xiph/opus/-/commit/cece1fd6

Figure 2: *Comparison of different compression complexities (in colors), different bitrates (groups in each plot) for single-channel (first plot), two channels (second plot), and seven channels (third plot) in terms of nWER. All results are shown with automatic compression mode selection and channel-independent compression. The corresponding uncompressed baseline is shown as a black dashed line.*

3. Experimental setup

All experiments are performed on $16\,\text{kHz}$ de-identified audio with up to seven channels. Typical microphone arrays for ASR consist of sensors spaced only few centimeters apart. The recorded signals are therefore highly correlated and differ primarily in phase. Beamforming [6] uses this phase difference to spatially filter the signal in order to suppress undesired components while enhancing the desired signal resulting in fewer recognition errors [25]. Here, we consider neural mask-based beamforming where the information about the dominant source is estimated on a time-frequency bin (tf-bin) level using a combination of a neural VAD and a neural mask estimator [26, 27]. The mask estimator is trained on separate synthetic data and is supported by a neural VAD [28] trained using production data and a forced alignment as reference. The beamforming vector is obtained using a Minimum Variance Distortionless Response (MVDR) formulation as in [29] with a fixed reference channel.

We use a RNN-T single-channel acoustic model [15]. It consists of an eight long short-term memory (LSTM) layer encoder and a two LSTM layer decoder. Each LSTM layer has 1024 units. The encoder and decoder outputs are projected to a 512 dimensional vector which is combined by the joint network with a single layer with 1024 units to provide a distribution over $4\,\text{k}$ wordpieces. It is trained with SpecAugment [30] on $\approx 60\,000\,\text{h}$ of de-identified production data, which is the single-channel output of an on-device beamformer. The RNN-T training data has been Opus-compressed with default settings and bitrates between 32 and $64\,\text{kbit/(s} \cdot \text{channel)}$.

For evaluation we use a de-identified dataset consisting of $126\,\text{h}$ of realistic production-like speech recorded by Amazon Echo-style microphone arrays. The dataset is available as uncompressed audio, which allows to conduct decoding experiments with different lossy compression setups against an uncompressed baseline. All experimental results are presented in terms of normalized word error rate (nWER) such that the single-channel baseline (single digit in percent) without any encoding results in a nWER of 1.0.

Table 2: *Comparison of channel-independent vs. pairwise compression. All compression results are with CELT compression.*

Bitrate / kbit/(s·channel)	Channels N	nWER		Delta / %
		indep.	pairwise	
uncompressed	2	0.907		
16	2	1.018	1.010	−0.8
32	2	0.949	0.933	−1.7
64	2	0.923	0.910	−1.4
128	2	0.901	0.899	−0.1

4. Experimental Results

Figure 2 visualizes the impact of Opus compression on a single-channel AM (2a), a two-channel beamforming front-end with a single-channel AM (2b) and a seven-channel beamforming front-end with a single-channel AM (2c). The compression mode is selected automatically (auto). The main observation is, that the nWER already saturates for about $32\,\text{kbit/(s} \cdot \text{channel)}$ for the single-channel AM, while all beamforming variants still heavily degrade with bitrates below $128\,\text{kbit/(s} \cdot \text{channel)}$. Further, we can deduce that especially for low bitrates a high compression complexity (here shown in colors) improves WERs significantly.

Table 2 compares pairwise compression using the proposed waveform matching with the channel-independent compression for different bitrates. It turns out that for all bitrates pairwise compression with waveform matching performs better with increased gains of up to $1.7\,\%$ relative for realistic bitrates.

Table 3 puts the compression mode, waveform matching and pairwise compression into relation. For channel-independent compression, CELT compression performs best independently of the proposed waveform matching. However, when channels are compressed in pairs the proposed waveform matching tends to reduce WERs.

To understand the impact of Opus compression on spatial properties we took a single $7.4\,\text{s}$ speech recording and delayed a copy by $\Delta = 5$ samples. Then, we calculated a direction-of-arrival spectrum using a fixed delay-and-sum beamformer for each look direction. Figure 3a shows the direction-of-arrival spectrum for the uncompressed two-channel signal. For all frequencies, the signals positively interfere at around 90°. U-shaped regions where signals cancel out are clearly visible. For pairwise compression with a bitrate of $32\,\text{kbit/(s} \cdot \text{channel)}$, Figure 3b shows that the curves of signal cancellation are much more distorted. Most prominently, the phase information is en-

Table 3: *Ablation study for a given bitrate of $32\,\text{kbit/(s} \cdot \text{channel)}$ in a two-channel beamforming system in terms of nWER.*

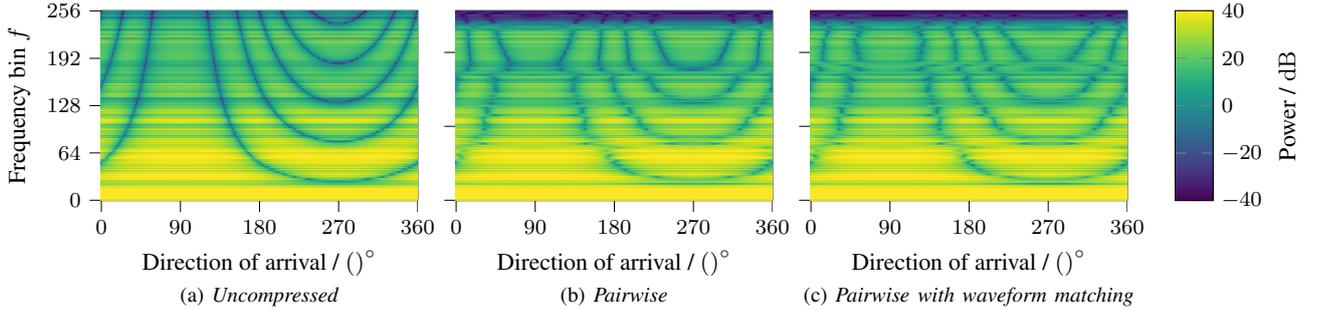| Compression mode	Channels N	Waveform matching	nWER	
			indep.	pairwise
uncompressed	2		0.907	
SILK	2	✗	0.960	0.933
SILK	2	✔	0.962	0.933
auto	2	✗	0.952	0.959
auto	2	✔	0.957	0.933
CELT	2	✗	0.949	0.959
CELT	2	✔	0.949	0.933

Figure 3: *Visualization of the impact of pairwise compression on the spatial characteristics of an artificially delayed recording in terms of a direction-of-arrival spectrum. A fixed steering vector for a delay-and-sum beamformer is calculated for each look direction. The response power for each direction is then a visualization of the spatial properties of the signal.*

tirely lost for frequency bin indices $f > 180$. Figure 3c shows results with the proposed waveform matching, demonstrating that the loss of phase information is avoided. Given the limited bitrate this likely resulted in reduced perceptual quality. However, with a beamforming application in mind, correct spatial properties are preferred over perceptual fidelity.

Finally, we evaluate the aforementioned improvements in combination with our proposed DFT-based coding scheme on a seven-channel microphone array signal. Table 4 shows the nWERs and the corresponding relative WER changes (delta). The "indep." column shows nWERs for independent coding of channels. The "DFT" column shows nWERs for signals coded with the proposed decorrelation using a DFT across channels. After Opus decoding the signal is reconstructed using an inverse DFT over channels. Transform coding improves WER for all bitrates with relative gains up to 6.7%, with the largest gain observed for the lowest bitrates, highlighting its importance for cohorts with limited internet connectivity. Alternatively, instead of optimizing for WER we can fix the WER at the value obtained for $32\,\mathrm{kbit/(s \cdot channel)}$ channel-independent compression. Using the proposed channel decorrelation we can reduce the bitrate down to $20\,\mathrm{kbit/(s \cdot channel)}$ without compromising on WER. This corresponds to a bitrate reduction of 37.5%. These findings encourage further research towards multi-channel extensions for lossy compression.

After applying a DFT across seven channels one obtains a real-valued channel (sometimes called the DC component) and three distinct complex-valued channels. Under the assumption that transformed channels and their respective quantization noise are mutually uncorrelated, transform coding principles suggest that bitrate should be allocated relative to signal power of each channel (compare Table 1 for a power distribution across channels). However, Table 5, in which we experimented with increasing the bitrate for the DC channel which has the highest power while keeping the average bitrate con-

Table 4: *Comparison of channel-independent CELT compression and transform coding using a channel-wise DFT before CELT compression and a channel-wise inverse DFT after decompression. The delta values in this table are w.r.t. the channel-independent column.*

Bitrate / $\mathrm{kbit/(s \cdot channel)}$	Channels N	nWER indep.	nWER DFT	Delta / %
16	7	1.015	0.947	−6.7
32	7	0.925	0.888	−4.0
64	7	0.882	0.870	−1.4
128	7	0.861	0.853	−0.9

stant, shows, that uniform bitrate assignment is better in our case. Additional experiments with signal power-based bitrate assignments resulted in worse WER and are omitted here for brevity reasons.

Given that the real and imaginary part of one DFT channel show similar dynamics, one may conclude that pairwise compression may be beneficial to allow the Opus codec to better exploit residual redundancies between those parts. Indeed, as indicated by the "DFT+pairwise" column in Table 5, larger WER changes can be achieved with additional pairwise compression. Here, we also use the proposed waveform matching for the complex pairs.

Table 5: *Relative WER change of the proposed transform coding over channel-independent compression including all prior optimizations. The "DFT" column corresponds to the DFT across channels while the "DFT+pairwise" column corresponds to additional pairwise compression of the real and imaginary part of each AC component, i.e., $\mathrm{Re}\{X_d\}$ and $\mathrm{Im}\{X_d\}$ for $d \in 1, \dots D-1$.*

Bitrate / $\mathrm{kbit/(s \cdot channel)}$	Delta / % indep.	DFT	DFT+pairwise
All channels 32	±0.0	−4.0	**−5.1**
DC channel 38, others 31		−3.5	−4.8
DC channel 44, others 30		−3.7	−4.5
DC channel 50, others 29		−3.3	−4.9
DC channel 56, others 28		−3.4	−4.4
DC channel 62, others 27		−3.3	−3.9

5. Conclusions

We analyzed why multi-channel far-field ASR degrades under Opus compression with limited bitrates and proposed modifications to the codec in order to reduce WER while maintaining compatibility with existing decoders. In a two-channel beamforming application this leads to up to 2% relative WER reduction without any increase in bitrate. We furthermore propose a new spatial transform coding scheme-based on a DFT decomposition across channels. While it is a simple operation it maintains compatibility to existing Opus encoder/decoder infrastructure, and is geometry- and signal-independent. For a seven-channel beamforming application it allows to either reduce the bitrate by 37.5% at no loss in accuracy or alternatively reduce WER by 5.1% relative with a fixed bitrate budget. Overall, our proposed multi-channel coding scheme achieves a relative WER reduction of 6.4% for a seven-channel microphone array compared to a naive channel-independent Opus compression.

6. References

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, 2020.

[2] Audio Software Engineering and Siri Speech Team, "Optimizing Siri on HomePod in far-field settings," https://machinelearning.apple.com/research/optimizing-siri-on-homepod-in-far-field-settings, 2018.

[3] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. K. Chin *et al.*, "Acoustic modeling for Google Home," in *Interspeech*, 2017.

[4] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal processing magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[7] K. Kumatani, W. Minhua, S. Sundaram, N. Ström, and B. Hoffmeister, "Multi-geometry spatial acoustic modeling for distant speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[8] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, 2017.

[9] J.-M. Valin, K. Vos, and T. B. Terriberry, "Definition of the Opus audio codec," https://tools.ietf.org/html/rfc6716, pp. 1–326, September 2012.

[10] J.-M. Valin and C. Bran, "WebRTC audio codec and processing requirements," https://tools.ietf.org/html/rfc7874, 2016.

[11] A. Khare, S. Sundaram, and M. Wu, "Multi-channel acoustic modeling using mixed bitrate Opus compression," *arXiv preprint arXiv:2002.00122*, 2020.

[12] P. Orosz, T. Skopkó, Z. Nagy, and T. Lukovics, "Performance analysis of the Opus codec in VoIP environment using QoE evaluation," in *International Conference on Systems and Networks Communications (ICSNC)*, 2013.

[13] M. Narbutt, S. O'Leary, A. Allen, J. Skoglund, and A. Hines, "Streaming VR for immersion: Quality aspects of compressed spatial audio," in *International Conference on Virtual System & Multimedia (VSMM)*. IEEE, 2017.

[14] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

[15] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[16] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the Opus codec," *Journal of the Audio Engineering Society*, October 2013.

[17] K. Vos, K. V. Sørensen, S. S. Jensen, and J.-M. Valin, "Voice coding with Opus," in *Audio Engineering Society Convention*. Audio Engineering Society, 2013.

[18] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[19] J.-M. Valin, T. B. Terriberry, and G. Maxwell, "A full-bandwidth audio codec with low complexity and very low delay," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2009.

[20] B. C. J. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[21] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[22] Z. Li and R. Ruraiswami, "Hemispherical microphone arrays for sound capture and beamforming," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2005.

[23] S. Venkatesan, L. Mailaender, and J. Salz, "An iterative algorithm for computing a spatial whitening filter," in *Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2004.

[24] D. Yang, H. Ai, C. Kyriakakis, and C.-C. Kuo, "High-fidelity multichannel audio coding with Karhunen-Loeve transform," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 365–380, 2003.

[25] D. V. Compernolle, W. Ma, F. Xie, and M. V. Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, vol. 9, no. 5, pp. 433 – 442, 1990.

[26] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[27] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Interspeech*, 2016.

[28] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012.

[29] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, Sept 2013.

[30] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "SpecAugment on large scale datasets," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020. [Online]. Available: http://dx.doi.org/10.1109/ICASSP40776.2020.9053205