MIMO Speech Compression and Enhancement Based on Convolutional Denoising Autoencoder

You-Jin Li*, Syu-Siang Wang†, Yu Tsao‡, and Borching Su*

* Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
† Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
‡ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

***Abstract*—For speech-related applications in Internet of things environments, identifying effective methods to handle interference noises and compress the amount of data in transmissions is essential for achieving high-quality services. In this paper, we propose a novel multi-input multi-output speech compression and enhancement (MIMO-SCE) system based on a convolutional denoising autoencoder (CDAE) model to simultaneously improve speech quality and reduce the dimension of transmission data. Compared with conventional single-channel and multi-input single-output systems, MIMO systems can be employed for applications where multiple acoustic signals need to be handled. We investigated two CDAE models, fully convolutional network (FCN) and Sinc FCN, as the core models in MIMO systems. The experimental results confirm that the proposed MIMO-SCE framework effectively improves speech quality and intelligibility, and reduces the amount of recording data to one-seventh for transmission.**

I. Introduction

Multichannel speech enhancement (MCSE) and speech compression techniques benefit several real-time speech communication in an Internet of things system [1], [2], [3], [4], [5]. Conventional MCSE systems with a multiple-input single-output (MISO) configuration suppress environmental noises from multiple noisy inputs to provide decent sound quality and intelligibility on the single-channel output side [6], [7], [8], [9]. Generally speaking, most MCSE algorithms were derived on beam-forming-based approaches [10], [11], [12], [13], wherein either the spatial diversity of received signals or the maximum signal-to-noise ratio (SNR) criterion were exploited to perform a linear filter function to preserve the desired signal [14], [15]. Several attempts further combine deep learning (DL) with conventional beam-forming-based MCSE to provide a robust transfer function and to promote the system capability on dealing with non-stationary noises environments [16], [17], [18], [19], [20], [21], [22], [23]. In addition to beam-forming-based approaches, some researches enhanced noisy recordings directly through the DL models. For example, the work in [24] used a denoising auto-encoder (DAE) model to suppress noise in the time domain to preserve the speech signal in an specified spatial direction. Our previous work [25] utilized a fully convolutional neural network (FCN) and Sinc FCN (SFCN) on MCSE to achieve decent speech quality and intelligibility in both subjective and objective tests. Notably, an Sinc layer [26] used in SFCN provides more meaningful filters to decompose the model inputs for the following FCN model.

However, apart from the improved sound quality, multichannel inputs also increase bandwidth, power consumption, and hardware costs for signal transmission and storage. An effective acoustic signal compression method is required to reduce the amount of captured data [27]. For acoustic signal compression, speech coding (SC) approaches are applied to remove the statistical redundancies or perceptual irrelevancies of input audio signals [28], [29], [30], [31], [32], [33], [34], [35], [36]. Traditional SC approaches, such as sub-band coding [37] and code-excited linear prediction [38], are derived by considering temporal properties to compress an single-channel speech signal. On the other hand, multichannel SC approaches, such as spatial audio coding [28], [29], [39], and modified discrete cosine transform [40], [40], [41], are applied to encode input signals by considering both coherence and statistical difference across channels. Generally, some level of distortions can be observed in coded and restored speech signals and slightly degrade the speech quality and intelligibility accordingly. Recently, DL techniques have been introduced in signal compression algorithms to perform SC systems [42]. In [43] and [44], an utterance was first analyzed using deep neural networks to extract phonological and prosodic speech representations to build novel speech codecs. In [45] and [46], speech spectra are encoded by a deep auto-encoder that is trained with identical input and output signals. The associated codecs are then derived from output nodes of the middle hidden layer. Meanwhile, DL models have been used as post-filters to enhance coded speech [47], [48], and have been shown to yield decent speech quality.

In this study, we propose a multi-input multi-output speech compression and enhancement (MIMO-SCE) framework. Notably, the multiple enhanced outputs can further be used for other applications, such as sound-based indoor positioning [49] and multi-channel ASR system [50]. The proposed enhancing-compressing framework is based on a convolutional DAE (CDAE) [25] model, comprising encoder and decoder parts. During training, the CDAE is trained to process noisy multichannel speech signals in order to generate enhanced signals. Thereafter, the encoder and decoder of the trained CDAE are placed at the edge and server, respectively. During testing, the encoder part encodes noisy multichannel speech inputs into bottleneck features with reduced dimensions. The encoded bottleneck features are then transmitted to the server and processed by the decoder to recover the enhanced multichannel

978-988-14768-9-0/21/$31.00 ©2021 APSIPA

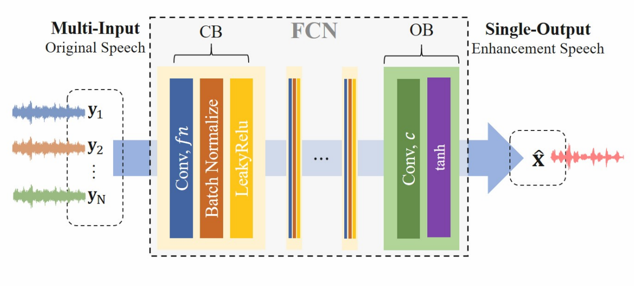

Fig. 1. Architecture of MISO-FCN. CB and OB in the figure represent convolutional blocks and the output blocks, respectively.

speech signals. Two CDAE models were implemented for the MIMO-SCE framework: an FCN-based (termed MIMO-SCE(F)) and an SFCN-based (termed MIMO-SCE(S)). Experimental results show that MIMO-SCE(F) and MIMO-SCE(S) can effectively reduce multichannel acoustic data by a factor of seven while improving speech quality and intelligibility.

The remainder of this paper is organized as follows: A review of related works is presented in Section II, and the concepts and architectures of the proposed MIMO-SCE(F) and MIMO-SCE(S) models are discussed in Section III. Section IV-A presents the experimental setup and results. Finally, the conclusions of the study are described in Section V.

II. Related Works

In this section, we first review MISO SE systems. Subsequently, we review two CDAE models: FCN and SFCN.

A. MISO SE system

For the multichannel noisy recording waveforms $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]$, where N denotes the number of channels, the MISO SE system aims to generate an enhanced speech signal $\hat{\mathbf{x}}$, where $\hat{\mathbf{x}} = f_\theta(\mathbf{Y})$; θ denotes the model parameters and is estimated by minimizing the difference between the generated speech $\hat{\mathbf{x}}$ and the clean reference. During the test, for a given noisy multichannel input, the MISO SE generates an enhanced single-channel output.

B. Two CDAE Models: FCN and SFCN

The CDAE model consists of an encoder and a decoder. In this study, two CDAE models were implemented. The first one is FCN, which comprises convolutional blocks (CBs), as shown in Fig. 1. Each CB consists of three components: convolution layer (Conv), batch normalization, and LeakyReLU. The filter number and filter length used in the convolution layer are fn and fl, respectively. A stack of CBs is concatenated for feature extraction and transformation. Finally, an output block (OB) consisted of a convolution layer and a tanh activation function is placed in the last part of the FCN. In an OB, the filter length of the convolution layer is defined as the output dimension c; for the MISO SE system, c is equal to 1.

In our previous work [25], we confirmed that SFCN can yield better MISO SE performance. The architecture of the SFCN is depicted in Fig. 2. The primary difference between the FCN and SFCN models is that SFCN adopts the Sinc convolution (SincConv) layer as the first CB. SincConv was designed and trained to provide various filter banks; thus, it can obtain band-pass information even for a limited amount and restricted diversity of training data. In addition, as SincConv contains fewer parameters, SFCN can be trained more efficiently.

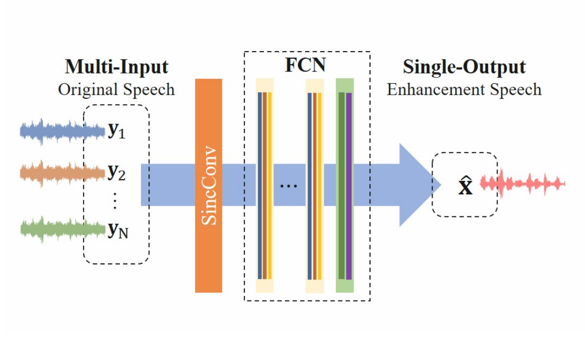

Fig. 2. Architecture of MISO-SFCN

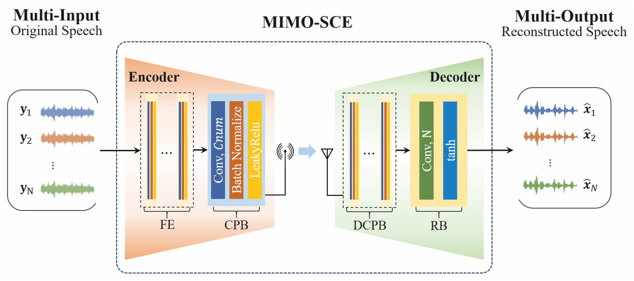

Fig. 3. Architecture of MIMO-SCE

III. Proposed MIMO Speech Compression and Enhancement Framework

In this section, we introduce the architecture of the proposed MIMO-SCE framework. Two CDAE models are used as the core units to build MIMO-SCE(F) and MIMO-SCE(S) systems. The goal of MIMO-SCE is to determine a function that transforms $\mathbf{Y}$ to multichannel clean speech signals, $\mathbf{X}$.

A. System architecture

The proposed MIMO-SCE system is presented in Fig. 3. The system is comprised of encoder and decoder parts. During training, for the noisy multichannel input $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]$, the MIMO system aims to generate enhanced speech signals $\hat{\mathbf{X}}$, where $\hat{\mathbf{X}} = f_\theta(\mathbf{Y})$; θ denotes the model parameters and is estimated by minimizing the difference of generated speech $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_N]$. Using the clean multichannel reference: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$, we train the model parameter θ to minimize the difference between $\hat{\mathbf{X}}$ and $\mathbf{X}$:

$$\hat{\theta} = \arg\min_{\theta} D(f_\theta(\mathbf{Y}), \mathbf{X}), \tag{1}$$

where $D(\cdot)$ denotes the cost function, which is defined in Eq. (2).

$$D(f_\theta(\mathbf{Y}), \mathbf{X}) = \sum_{i=1}^{N} (\hat{\mathbf{x}}_i - \mathbf{x}_i)^2 \,. \tag{2}$$

After training, we place the encoder and decoder parts of the trained model at the edge and server sides, respectively. During the test, for the noisy multichannel input, the MIMO system first encodes the data into a latent representation with the reduced dimension. The encoded representation vectors are then transmitted to the server side and finally reconstructed to multichannel outputs based on the decoder. Because the latent representations (instead of original multichannel inputs) are transmitted, the data size is reduced; thus, online transmission bandwidth costs can be reduced.

B. MIMO-SCE(F) and MIMO-SCE(S)

The proposed MIMO-SCE(F) and MIMO-SCE(S) process speech signals in the time domain. The main advantage of time-domain speech signal processing is that the phase information can be more accurately preserved, as compared with spectral-domain processing.

For MIMO-SCE(F) and MIMO-SCE(S), we designed a bottleneck architecture, where a middle layer has few dimensions, that is used to compress multichannel inputs termed as the compression block (CPB). By assigning the filter number of the CPB to C_{num}, the compression rate is $R_{comp} = N/C_{num}$, which is derived from the channel number before and after the encoder. The inputs of MIMO-SCE(F) and MIMO-SCE(S) are the same as those used in the MISO systems, as shown in Figs. 1 and 2, respectively, and the outputs of the two systems are multichannel signals, as shown in Fig. 3.

The MIMO-SCE(F) encoder consists of a feature inductor (FE) and CPB, where the FE is combined using four-layer CBs. All CBs have identical architectures, including Conv with filter number $fn = 30$, filter length $fl = 55$, Batch Normalization, and LeakyRelu. The CPB has a filter length of $fl = 55$, filter number $C_{num} = 1$, Batch Normalization, and LeakyRelu. Subsequently, the decoder consists of a decompression block (DCPB) and a reconstruction block (RB). The DCPB also has four-layer CBs that decompress the transmission signal. The CB set is the same as the encoder. The RB has a Conv layer with a filter length $fl = 55$, filter number $c = N$, and tanh activation function to rebuild the multichannel speech data, where $N = 7$ in this paper.

The encoder and decoder design of MIMO-SCE(S) is similar to that of MIMO-SCE(F). However, in MIMO-SCE(S), SincConv is added as the encoder's first CB to extract additional speech features, as shown in Fig.2. The rest part of MIMO-SCE(S) is identical to MIMO-SCE(F).

For MIMO-SCE(F) and MIMO-SCE(S), we use the cost function in Eq. (2) to estimate the model parameters. On the other hand, instead of using all seven-recording signals, the one-channel encoder output is transmitted to the far-end server terminal. Therefore, the compression rates of both systems are 7/1.

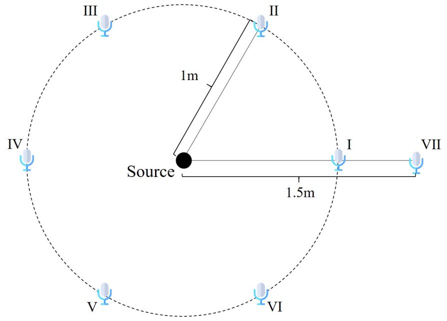

Fig. 4. Recording settings for the experiments. The speaker is placed at the center (source) and surrounded by seven microphones (I to VII). Microphones I to VI are placed 1 m away from the source, whereas microphone VII is placed at 1.5 m and behind microphone I.

IV. Experiments

In this section, we introduce the experimental setup and results of the proposed MIMO-SCE framework. The compression ratio (R_{comp}) was maintained at 7, and the speech quality (measured via the perceptual evaluation of speech quality, PESQ [51]) and intelligibility (measured via short-time objective intelligibility, STOI [52], [53]) of the enhanced multichannel outputs were measured and reported as the evaluation results. The PESQ score ranges from 0.5 to 4.5, and the STOI score typically ranges from 0 to 1. Higher PESQ and STOI scores indicate better speech quality and intelligibility, respectively.

A. Experimental Setup

The speech data used in this study were recorded using the setup shown in Fig. 4. The loudspeaker (head and torso simulator) was placed at the center (Fig. 4) and surrounded by seven microphones. Six microphones——I, II, III, IV, V, and VI——were placed at a distance of 1 m from the source, whereas microphone VII was placed 1.5 meters away from the source. All seven microphones were of the same model (Sanlux HMT-11). The transcript material is the Taiwan Mandarin Hearing in Noise Test dataset (TMHINT) [54], which is a phonetically balanced corpus consisting of 320 sentences and ten Chinese characters in each sentence. All utterances were pronounced by a native Mandarin male speaker for recording at 16 kHz sampling rate with seven microphones in the clean environment. We further split 320 utterances into two parts: 250 utterances for training and 70 utterances for testing. The training utterances from the seven microphones were contaminated with eight noise types: pink, fan, babble, gun, alarm bell, cough, buccaneer, and engine, at signal-to-noise ratios (SNRs) of -10, -5, 0, 5 and 10 dBs. Therefore, there are $35,000 = 250 \times 7 \times 5 \times 4$ noisy-clean utterance pairs in the training set. The testing utterances from the seven microphones were contaminated with another four noise types, namely sound of a water cooler, street noise, car noise, and the bell of a fire truck, at SNRs of -10, -5, 0, 5 and 10 dB, and thus provide $9,800 = 70 \times 7 \times 5 \times 4$ noisy testing samples.

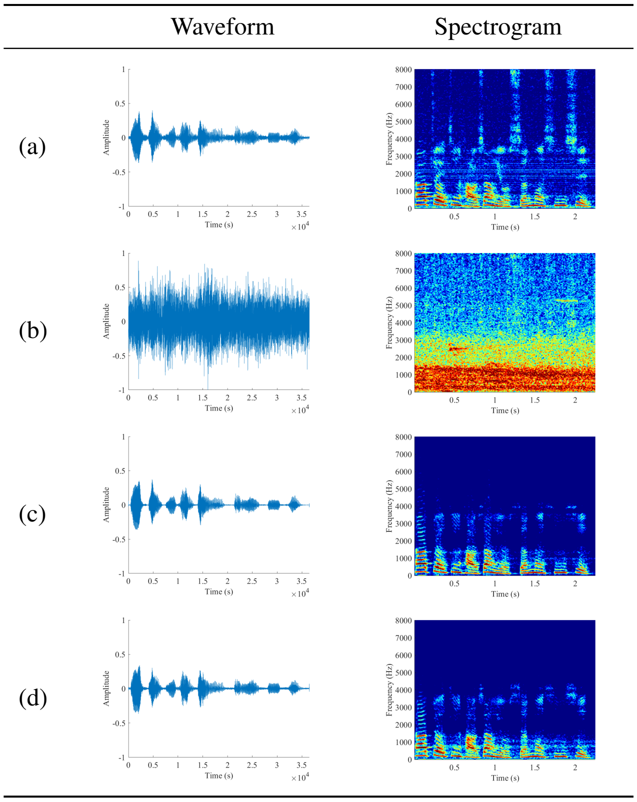

Fig. 5. Spectrogram and waveforms of (a) clean, (b) noisy (under street noise at 10 dB SNR), and (c) MIMO-SCE(S) with a compression ratio of 1, and (d) MIMO-SCE(S) with a compression ratio of 7. All utterances in the figure were selected from the microphone III.

B. Experimental results

The qualitative and quantitative results of the proposed MIMO-SCE(F) and MIMO-SCE(S) are presented in this section. Those results of the testing noisy that denoted as "Noisy" in the following sections are also listed as the baseline.

1) Qualitative spectrogram comparison: We first demonstrate the effects of the compression ratio (R_{comp}) on the proposed MIMO-SCE(S), as depicted in Fig. 5, in terms of the spectrum plots and the associated waveforms of a sample utterance recorded from the microphone III. Fig. 5 (a) and (b) present the clean and noisy utterances, respectively, whereas (c) and (d) depict the utterances derived from MIMO-SCE(S) with compression ratios of 1 and 7, respectively. On comparing Figs. 5 (c) and (d) with (b), it is evident that the noise components in the noisy spectrum and waveform were effectively suppressed. Furthermore, the harmonic structures of the spectrum and the envelope of waveforms in Figs. 5 (c) and (d) are preserved by MIMO-SCE(S), compared with those in Fig. 5 (a). These results indicate the effectiveness of the proposed model in enhancing speech subjected to noise environments and a high compression ratio. Therefore, the MIMO-SCE models with a compression ratio of 7 are used and evaluated, as described in the following section. On the other hand, we also noted that the quality of high-frequency components in Figs. 5 (c) and (d) are degraded when comparing those with the clean spectra in Fig. (a). One possible inference is highly distorted frequency-band signals remaining the hard task for FCN model.

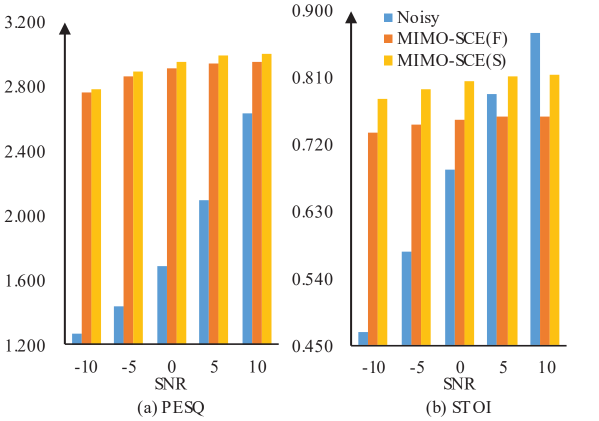

Fig. 6. Averaged (a) PESQ and (b) STOI scores of Noisy, MIMO-SCE(F), and MIMO-SCE(S) in -10, -5, 0, 5 and 10 SNRs.

2) Quantitative objective evaluation results: Table I lists the average PESQ and STOI results over all seven output channels and noise conditions for each of Noisy, MIMO-SCE(S) and MIMO-SCE(F). In addition, for each output channel, the associated clean signal in that channel was applied as a reference for performing PESQ and STOI metrics. From the table, we noted that MIMO-SCE(F) and MIMO-SCE(S) outperform Noisy in terms of the PESQ and STOI scores. MIMO-SCE(F) yields higher PESQ and STOI scores than MIMO-SCE(S), confirming the advantages of incorporating the SincConv layer in the enhancement system.

To further analyze the results listed in Table I, we present the detailed PESQ and STOI scores of Noisy, MIMO-SCE(F), and MIMO-SCE(S) at specific SNRs (-10, -5, 0, 5 and 10 dB) in Fig. 6. First, from Fig. 6 (a), we note that both MIMO-SCE(F) and MIMO-SCE(S) improve PESQ scores over Noisy, and more significant improvements were observed at lower SNR levels. Meanwhile, MIMO-SCE(F) marginally outperforms MIMO-SCE(S) consistently over different SNR levels. From Fig. 6 (b), we note that MIMO-SCE(F) and MIMO-SCE(S) improve the STOI scores over Noisy at low SNR conditions (-5 to 0 dB); however, both MIMO-SCE(F) and MIMO-SCE(S) do not provide further enhancements over noisy under cleaner conditions (at 5 and 10 dB SNRs). A possible inference for those reduced STOI scores is the distorted speech resulting from the data compression function of the proposed models.

TABLE I
AVERAGE PESQ AND STOI SCORES OF NOISY, MIMO-SCE (F), AND MIMO-SCE(S)

	Noisy	MIMO-SCE(F)	MIMO-SCE(S)
PESQ	1.825	2.890	2.927
STOI	0.678	0.750	0.801

V. Conclusions

In this paper, we propose a novel MIMO-SCE system to perform data compression for the simultaneous transmission and enhancement of speech signals. We investigated two CDAE models——FCN and SFCN——as core models in the proposed system, with a short-hand notation MIMO-SCE(F) and MIMO-SCE(S), respectively. The experimental results show that, under a high compression ratio of 7, the proposed MIMO-SCE(F) and MIMO-SCE(S) models improve speech quality and reproducibility under various SNR conditions. To the best of our knowledge, this is the first study to simultaneously perform data compression and SE based on DL-based CDAE models, in an MIMO scenario. In the future, we plan to explore MIMO systems for the integration of other heterogeneous data, such as visual and textual data, to further improve data compression and SE efficacy. More realistic microphone configurations are also going to be investigated for the proposed MIMO-SCE approach.

VI. Acknowledge

This work was supported in part by the grants AS-GC-109-05 and AS-CDA-106-M04

References

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[3] R. Cheng, C. Bao, and Z. Cui, "Mass: Microphone array speech simulator in room acoustic environment for multi-channel speech coding and enhancement," *Applied Sciences*, vol. 10, no. 4, p. 1484, 2020.

[4] J. Qi, H. Hu, Y. Wang, C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Tensor-to-vector regression for multi-channel speech enhancement based on tensor-train network," *arXiv preprint arXiv:2002.00544*, 2020.

[5] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. ICASSP*, 2019, pp. 451–455.

[6] F. de la Hucha Arce, M. Moonen, M. Verhelst, and A. Bertrand, "Adaptive quantization for multichannel wiener filter-based speech enhancement in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, p. 15, 2017.

[7] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. ICSLP*, 1996, pp. 929–932.

[8] S. van Ophem and A. P. Berkhoff, "Multi-channel kalman filters for active noise control," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2105–2115, 2013.

[9] A. Hyvarinen, "A family of fixed-point algorithms for independent component analysis," in *Proc. ICASSP*, 1997, pp. 3917–3920.

[10] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on signal processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[11] S. Emura, S. Araki, T. Nakatani, and N. Harada, "Distortionless beamforming optimized with ℓ_1-norm minimization," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 936–940, 2018.

[12] Y. T. Tsai, B. Su, Y. Tsao, and S.-S. Wang, "Adaptive subspace-constrained diagonal loading," in *Proc. APSIPA*, 2016, pp. 1–4.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[14] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[15] W. Kellermann, "Beamforming for speech and audio signals," in *Handbook of signal processing in acoustics*, 2008, pp. 691–702.

[16] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[17] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[18] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Proc. Interspeech*, 2018, pp. 1571–1575.

[19] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *Proc. ICASSP*, 2017, pp. 446–450.

[20] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. EUSIPCO*, 2018, pp. 1577–1581.

[21] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep learning based speech beamforming," in *Proc. ICASSP*, 2018, pp. 5389–5393.

[22] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[23] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," *arXiv preprint arXiv:1910.06522*, 2019.

[24] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," *Proc. Interspeech*, pp. 86–90, 2019.

[25] C.-L. Liu, S.-W. Fu, Y.-J. Lee, Y. Tsao, J.-W. Huang, and H.-M. Wang, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (Early Access)*, pp. 1–1, 2020.

[26] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. SLT*, 2018, pp. 1021–1028.

[27] A. Spanias, T. Painter, and V. Atti, *Audio signal processing and coding: Chapter 10*. John Wiley & Sons, 2006.

[28] F. Baumgarte and C. Faller, "Binaural cue coding-part I: Psychoacoustic fundamentals and design principles," *IEEE transactions on speech and audio processing*, vol. 11, no. 6, pp. 509–519, 2003.

[29] C. Faller and F. Baumgarte, "Binaural cue coding-part II: Schemes and applications," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 520–531, 2003.

[30] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 969–978, 1990.

[31] V. Ramamoorthy and N. Jayant, "Enhancement of adpcm speech by adaptive postfiltering," *AT&T Bell Laboratories Technical Journal*, vol. 63, no. 8, pp. 1465–1475, 1984.

[32] Y. Hiwasaki, S. Sasaki, H. Ohmuro, T. Mori, J. Seong, M. S. Lee, B. Kövesi, S. Ragot, J.-L. Garcia, C. Marro *et al.*, "G. 711.1: a wideband extension to itu-t g. 711," in *Proc. EURASIP*, 2008, pp. 1–5.

[33] J. He and W.-S. Gan, "Applying primary ambient extraction for immersive spatial audio reproduction," in *Proc. APSIPA*, 2015, pp. 1000–1009.

[34] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, 2013.

[35] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. WASPAA*, 2001, pp. 199–202.

[36] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1305—-1322, 2005.

[37] K. Sayood, *Introduction to data compression: Chapter 14*. Morgan Kaufmann, 2017.

[38] R. Jage and S. Upadhya, "CELP and MELP speech coding techniques," in *Proc. WiSPNET*, 2016, pp. 1398–1402.

[39] C. Faller, "Parametric multichannel audio coding: synthesis of coherence cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 299–310, 2005.

[40] K. Suresh and R. A. Raj, “MDCT domain parametric stereo audio coding,” in *Proc. SPCOM*, 2012, pp. 1–4.

[41] S. Chen, N. Xiong, J. H. Park, M. Chen, and R. Hu, “Spatial parameters for audio coding: MDCT domain analysis and synthesis,” *Multimedia Tools and Applications*, vol. 48, no. 2, pp. 225–246, 2010.

[42] A. Biswas and D. Jia, “Audio codec enhancement with generative adversarial networks,” *arXiv preprint arXiv:2001.09653*, 2020.

[43] M. Cernak, B. Potard, and P. N. Garner, “Phonological vocoding using artificial neural networks,” in *Proc. ICASSP*, 2015, pp. 4844–4848.

[44] M. Cernak, A. Lazaridis, A. Asaei, and P. N. Garner, “Composition of deep and spiking neural networks for very low bit rate speech coding,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2301–2312, 2016.

[45] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep auto-encoder,” in *Proc. Interspeech*, 2010.

[46] S. Kankanahalli, “End-to-end optimized speech coding with deep neural networks,” in *Prco. ICASSP*, 2018, pp. 2521–2525.

[47] Z. Zhao, S. Elshamy, H. Liu, and T. Fingscheidt, “A cnn postprocessor to enhance coded speech,” in *Proc. IWAENC*, 2018, pp. 406–410.

[48] Z. Zhao, H. Liu, and T. Fingscheidt, “Convolutional neural networks to enhance coded speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2018.

[49] H.-G. Kim and G. Y. Kim, “Deep neural network-based indoor emergency awareness using contextual information from sound, human activity, and indoor position on mobile device,” *IEEE Transactions on Consumer Electronics*, vol. 66, no. 4, pp. 271–278, 2020.

[50] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, “Unified architecture for multichannel end-to-end speech recognition with neural beamforming,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[51] 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. ICASSP*, 2001, pp. 749–752.

[52] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. ICASSP*, 2010, pp. 4214–4217.

[53] ——, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[54] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, “Development of the mandarin hearing in noise test (MHINT),” *Ear and hearing*, vol. 28, no. 2, pp. 70S–74S, 2007.