2023 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2023, ROME, ITALY

VOICE DIRECTION-OF-ARRIVAL CONVERSION

I-Chun Chern[1], *Steffi Chern*[1], *Heng-Cheng Kuo*[2], *Huan-Hsin Tseng*[3], *Kuo-Hsuan Hung*[2], *Yu Tsao*[2]

[1] Carnegie Mellon University, [2] Academia Sinica, [3] Brookhaven National Laboratory

ABSTRACT

The demand for augmented reality (AR) and virtual reality (VR) is steadily rising. To provide the best user experience in a virtual environment, their applications must ensure consistency between the visual and audio signals perceived by the users. For example, when an avatar (sound source) speaks while moving, the arrival direction of the avatar changes. In this paper, we introduce a voice direction of arrival (DOA) conversion task that aims to change the DOA of speech signals while keeping the remaining components unaltered. Furthermore, we propose DOAC-Net, a novel speech DOA conversion system that can perform causal speech DOA conversion. The results show that DOAC-Net can effectively convert the DOA of multi-channel speech signals with little distortion, while maintaining speech quality and intelligibility.

Index Terms: direction-of-arrival (DOA), audio conversion, DOA conversion, multi-channel audio conversion, stereo audio conversion

1. INTRODUCTION

The demand for augmented reality (AR) and virtual reality (VR) applications, such as virtual online conferences, has gained increasing attention. Smooth integration of audio and video signals plays a key role in AR and VR applications. In a virtual online conference, as we explore the surroundings or communicate with other participants, we expect the sound source to reflect the relative spatial positions of the avatars. On the other hand, when listening to oral presentations, it is preferable for the audio to continually originate from directly ahead to avoid distractions. Therefore, for audio signals, the direction of arrival (DOA) is a key component that indicates the sound source of interest in the augmented or virtual environments. In other virtual scenarios, such as virtual games, plays, or concerts, converting DOAs on the fly is also a desirable feature to achieve a more immersive experience. In this study, we investigate the causal DOA conversion based on deep learning (DL) for the first time.

The DOA is the arrival direction of the audio wave relative to a set of receivers, such as human ears or a microphone array [1]. The DOA estimation is widely utilized for locating and tracking signal sources (e.g., sonar, emergency search and rescue, etc). Numerous research works have been have been conducted to accurately estimate the DOA for a given set of received audio signals using either estimation-theory- [2, 3] or DL-based methods [4]. Besides, in other research [5, 6], the multiple audio waves are further combined into one signal based on beamforming techniques. The beamformed signals generally possess better audio quality with reduced noise and interference components; in the meanwhile, DOA information is discarded. On the contrary, another research work focuses on simulating a set of audio waves from a single audio source with a specified DOA information [7, 8]. With the simulated audio waves, listeners or microphone arrays can identify the direction of the audio signals.

To the best of our knowledge, no previous work has investigated the task of voice DOA conversion that involves changing the DOA of audio signals while keeping the other audio content unaltered. In this study, we implemented a novel voice DOA conversion system, called DOAC-Net based on the ResNet architecture. To perform the DOA conversion using traditional methods, a three-stage framework is conceivable: (1) inferring the source DOA of the received signals, (2) converting incoming multi-channel audio to a single-channel one by MVDR beamforming (or similar) techniques, and (3) forwarding the beamformed audio and the DOA parameter obtained from stage (1) to the Pyroomacoustics (PyRoom) simulator [7] to generate outputs of designated DOA. However, PyRoom simulator requires detailed information such as the room size, source location, and locations of the microphone array, which usually remain unknown to users. Thus, in the proposed DOAC-Net, we transfer the received set of audio waves to a specific DOA angle end-to-end. When performing voice DOA conversion, DOAC-Net directly generates the audio signals with a specified target DOA and leaves the remaining components of received audio signals unchanged. The comparison of the integration of the traditional methods and the proposed end-to-end system is illustrated in Fig. 1.

We used speech signals from the VCTK corpus to prepare audio data and the PyRoom simulator to generate paired training data with arbitrary DOAs. In the experiments, two DOA estimation methods were used to evaluate the resulting signals, the traditional MUSIC [2] and a DL method (termed DOAE-Net in the following). We note that, based on the proposed DOA estimation method, DOAC-Net can accurately convert the DOAs of audio signals according to our specifi-

979-8-3503-2411-2/23/$31.00

cations.

The remainder of this paper is organized as follows: In Section 2, the related works, including the DOA estimation, acoustic beamforming, and speech conversion, are reviewed. Section 3 introduces the proposed DOAC-Net system. The experimental setup and results are presented in Section 4. Finally, Section 5 concludes our work.

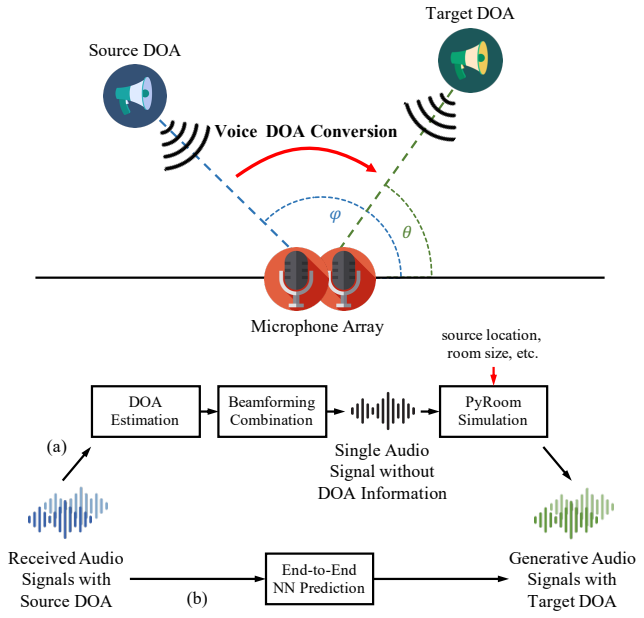

Fig. 1. The top panel is the visualization of voice DOA conversion; the bottom panel includes the flowcharts of (a) the traditional three-stage conversion and (b) the proposed end-to-end neural network (NN)-model.

2. RELATED WORKS

2.1. DOA Estimation

For DOA estimation, we generally assume that the received signal satisfies the far-field condition, and a microphone array is required. That is, assuming that the signal is a plane wave, the angle of the signal relative to the array is considered to be the DOA. Among the DOA estimation techniques, one representative class is MUSIC [2] and its various extensions [9, 10, 11]. Another is the rotation-invariant subspace-based methods that include ESPRIT [3] and its extensions [12, 13]. Notably, analytical algorithms such as MUSIC and ESPRIT set out from strong mathematical assumptions that restrict the possible application scenarios. In fact, recall that a m-channeled temporal signal $x(t) = \sum_i^K A_i\, e^{-i\omega_i t} + v(t) \in \mathbb{C}^m$ of K source signals and noise $v(t)$, the *covariance matrix* [14] of the combined signal x is given by $R_x = \mathbb{E}[x(t) \cdot x^*(t)]$ such that R_x is an $m \times m$ self-adjoint (Hermitian) matrix, $R_x^* = R_x$. Thus, R_x can be decomposed into,

$$R_x = \sum_{j=1}^{m} \lambda_j\, u_j \otimes u_j^* \tag{1}$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \in \mathbb{R}$ and the corresponding (orthonormal) eigenvectors $u_1, \ldots, u_m \in \mathbb{C}^m$ by spectral decomposition theorem [15]. There is also a natural *project operator* associated to the spectral decomposition, $P = \sum_{j=1}^m u_j \otimes u_j^*$. Therefore, there is a *cut-off* we need to decide which eigen-decompositions belong to the noise and which belong to the K sources. By choosing a threshold $j^* \in \mathbb{N}$, we *manually* determine below the *noise level* $\lambda_{j^*} \geq \cdots \geq \lambda_m$ is deemed as *noise*. Correspondingly, we have a noise projection operator $P_{\text{noise}} = \sum_{j>j^*}^m u_j \otimes u_j^*$. The MUSIC then utilizes a function $f(v) = \frac{1}{||P_{\text{noise}}(v)||^2}$ to detect DOA since

$$f(v) = \frac{1}{||P_{\text{noise}}(v)||^2} = \begin{cases} \infty & v \in \{u_1, \ldots, u_{j^*}\} \\ \text{finite} & v \in \{u_{j^*}, \ldots, u_m\} \end{cases} \tag{2}$$

In summary, there are two inherent constraints in MUSIC as follows: (1) it considers that the clean signal must be *louder* than the noise (eigenvalues) and (2) the number of noise signals and clean sources combined cannot exceed the microphone number m. Note that both conditions can be easily violated as in the real world there is no constraint on the noise volume and the number of noises and acoustic sources. By looking into the analytical methods, we need to bear in mind that these signal-processing-based approaches are generally designed based on strong assumptions of signal models. Although they have been widely used in real-world applications, their achievable performance drops when the assumptions fail.

Recently, DOA estimation based on DL techniques has gained significant attention. Generally speaking, these methods use DL models to directly characterize the relationship between the input signal and DOA [16, 17]. It has been shown that the DL-based approaches can provide high DOA estimation accuracy with reduced computational complexity as compared to MUSIC and ESPRIT; moreover, they have better generalization capabilities to be implemented in realistic scenarios, where inevitable noise and reverberation may be involved.

2.2. Acoustic Beamforming

Acoustic beamforming performs two major functions: signal extraction and source localization. When multi-channel sound signals are available, acoustic beamforming can retrieve the signal of interest with high quality. Among various acoustic beamforming techniques, MVDR [18, 19, 20] and its extensions (such as LCMV and GSC) [21, 20] constitute a representative group. These methods have found extensive usage across a diverse range of speech-related applications. The main idea of MVDR is to minimize the interference

and noise sources while keeping the desired signal distortionless in a specific direction (i.e., DOA of the desired signal). More recently, DL has been introduced to perform acoustic beamforming [22, 23]. Compared to traditional beamforming methods (such as MVDR), the DL-based approaches can yield better performance and possess the advantages of being trained along with downstream tasks [6, 24].

2.3. Speech Conversion

The purpose of speech conversion is to modify the properties of speech, such as speech content, emotion, accents, and speaker characteristics. Maintaining the remaining components unchanged while manipulating the target property poses a challenge. Unlike the discrete nature of texts or figures, it is difficult to annotate the continual speech signals of variable lengths. With the surge in DL techniques, neural network (NN)-based models have achieved state-of-the-art performances, particularly in speaker conversion and emotion transformation [25, 26]. The NN-based models were shown to have a better ability to disentangle the target conversion component and maintain unchanged components from the speech signals.

However, voice DOA conversion has not received ample attention. It is important to note that multi-channel speech signals carry additional spatial information. With such spatial information, humans can locate the sound source. Therefore, the DOA information is important for applications in AR and VR fields, where it is necessary to establish the spatial relationship of avatars and imitate moving sound sources in real time. In this study, we first considered the DOA as the main component of interest and other factors in speech as the remaining components to perform voice DOA conversion.

3. MODEL ARCHITECTURE

3.1. Audio Type

The input and label data are assumed to be m-channel audio signals in temporal form, denoted by $x = (x_1, \ldots, x_m) \in \mathbb{R}^{m\times T}$ and $y = (y_1, \ldots, y_m) \in \mathbb{R}^{m\times T}$, respectively, where each $x_i, y_i \in \mathbb{R}^T$ is a mono-channel audio of length T.

3.2. The DOA Conversion Net

The ground-truth DOA of an m-channel input x is denoted as θ_x. DOAC-Net is designed to be an end-to-end function $f : \mathbb{R}^{m\times T} \times \mathbb{R}^1 \to \mathbb{R}^{m\times T}$, given a designated DOA shift ϕ such that the output $f(x, \phi)$ produces target DOA $\theta_{f(x,\phi)} \cong \theta_x + \phi$. The DOAC-Net consists of three parts $f = \mathcal{F}^{-1} \circ h \circ \mathcal{F}$, where $\mathcal{F} : \mathbb{R}^{m\times T} \to \mathbb{C}^{m\times k\times \tilde{T}}$ is the *complex* Short-Time Fourier-Transform (STFT) converting temporal signals to a complex k-Fourier space with $\mathcal{F}^{-1} : \mathbb{C}^{m\times k\times \tilde{T}} \to \mathbb{R}^{m\times T}$ denoting its inverse operation. The STFT function $\mathcal{F}$ is fixed with no unknown parameters to be varied. Therefore, the parameters to be trained in f solely come from the network h. The main network h receives a target shift ϕ such that $h(\cdot, \phi) : \mathbb{C}^{m\times k\times \tilde{T}} \to \mathbb{C}^{m\times k\times \tilde{T}}$ converts from one spectrum to another, with the following structure (see Fig. 2).

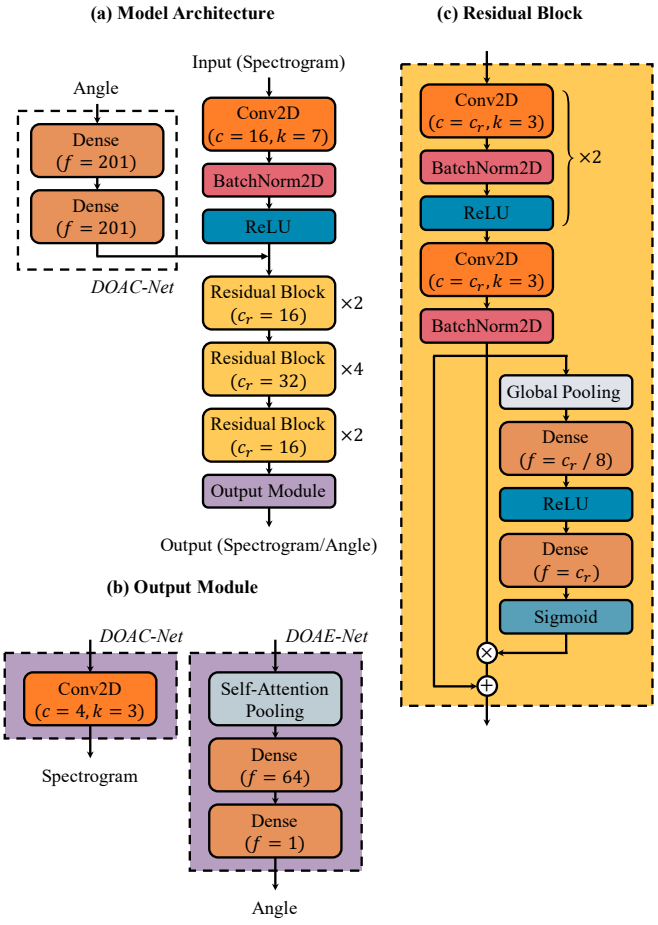

Fig. 2. The left panel is the structure of DOAC-Net/DOAE-Net. It should be highlighted that the angle embedding is only used in DOAC-Net. The right panel details the structures of the residual block. **Conv2D** represents the 2D convolution layer, where c denotes the number of filters and k denotes the kernel size. f is the size of the output feature in the dense layer.

The spectrum conversion net h utilizes the ResNet [27] as the building block, which is composed of various 2D convolution layers and batch-normalization layers (Fig. 2). We take stride 1 in every ResNet module and remove the pooling layers so that the STFT frame number can always match the time dimension. Under such construction, h takes a multi-channel complex spectrogram $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_m) \in \mathbb{C}^{m\times k\times \tilde{T}}$ and a target angle $\phi \in \mathbb{R}$ as input. This target angle (in radians) is first processed by an angle encoder to get an angle embedding, which then entangles with the complex spectrogram $\tilde{x}$ to *inject and integrate* the desired anglar information.

To train the end-to-end neural network f, we define a loss function in the k-Fourier space for h in that

$$\mathcal{L}(h) = \|h(\tilde{x}, \phi) - \tilde{y}\|_{\mathbb{C}^{m \times k \times \tilde{T}}} \tag{3}$$

where $\tilde{y} = \mathcal{F}(y)$ is the spectrum of the target audio y.

3.3. The DOA Estimation Net

The DOA Estimation Net (DOAE-Net) was designed to tackle the problem mentioned in Section 2.1, where the performances of the MUSIC algorithm were significantly hampered in the presence of noise or reverberation. Since extraction of the DOA information is the key concept no matter in the DOA conversion and DOA estimation tasks, we used a similar ResNet-based structure as the DOAC-Net with the target angle removed and a few linear layers concatenated at the end to predict the ground truth DOA θ_x. The training of DOAE-Net used the Mean Absolute Error (MAE) as the loss function.

$$\mathcal{L}(g) = |g(\tilde{x}) - \theta_x|_{\mathbb{R}} \tag{4}$$

where g denotes the DOAE-Net.

4. EXPERIMENTS

4.1. Experimental Setup

The speech corpus selected for our task was the Voice-Bank Corpus [28], consisting of *monaural* (single-channel) clean speech uttered by 110 English speakers. Each speaker spoke approximately 400 sentences. We randomly selected 2500 clean utterances as the *training set* and another 250 utterances for *testing* to build up our own dataset at a sampling rate of 16kHz. However, this dataset remained purely monaural due to the source of Voice-Bank. To conduct DOA conversion experiments, we required a multi-channel audio dataset with various DOA information within our control. Therefore, the Pyroomacoustics [7] (PyRoom) package was used to synthesize the multi-channel audio by generating a room impulse response (RIR) with virtual space settings. The generated RIRs were convolved with the mono-sound signals to simulate the spatial sound signals.

A virtual audio environment can be easily created in PyRoom, in which we build up a rectangular room of size $40 \times 40m^2$ along with a linear 2-microphone array placed at the room center, as shown in Fig. 3. The 2-microphone array has linear separation of $0.1m$ to receive a sound source in the room. The incident angle of a source to the microphone array is defined as the Direction of Arrival (DOA), which is calculated counterclockwise from the x-axis. The source is placed at a fixed distance of $5m$ from center of the microphone. Because of the front and rear symmetry of the 2D linear array, we only considered possible DOA angles ranging from $0°$ to $180°$ with a $10°$ separation.

The geometrical setting of the source with the Voice-Bank corpus resulted in a total of $47500 = 2500 \times (18 + 1)$ utterances for the training set and $3800 = 200 \times (18 + 1)$ utterances for the testing set, where all these utterances were from 2-channels due to the PyRoom simulation.

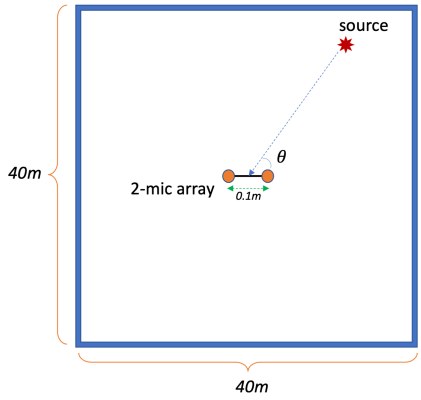

Fig. 3. Virtual room to generate our multi-channel speech data using PyRoom. A linear 2-microphone array is at the room center to receive a sound source from DOA $\theta \in [0°, 180°]$. The DOAC net aims to convert the original sound source incident at an angle θ to a target sound of an assigned DOA ϕ, as shown in Fig. 1.

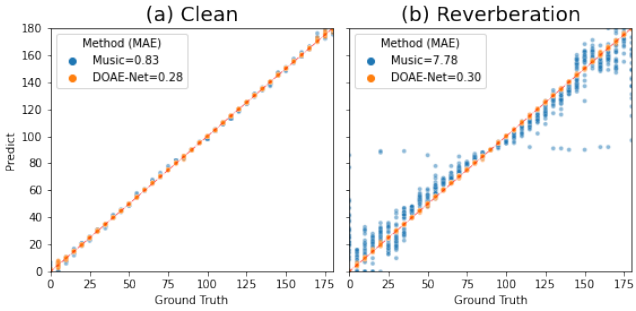

Fig. 4. Results of DOA prediction using MUSIC and DOAE-Net in with reverb and without reverb environments.

4.2. DOA Conversion Results

4.2.1. DOA Evaluation

To evaluate the efficiency of the proposed method, we first compared the performance of two DOA estimation methods (DOAE-Net and MUSIC). Then, we used these two methods to estimate the DOA of the resulting output. As shown in Fig. 4, the proposed DOAE-Net outperforms MUSIC in both clean and reverberant environments. It is worth mentioning that the DOAE-Net achieves a 0.30 MAE of DOA prediction in a reverberant environment, which is close to that in the clean environment (0.28).

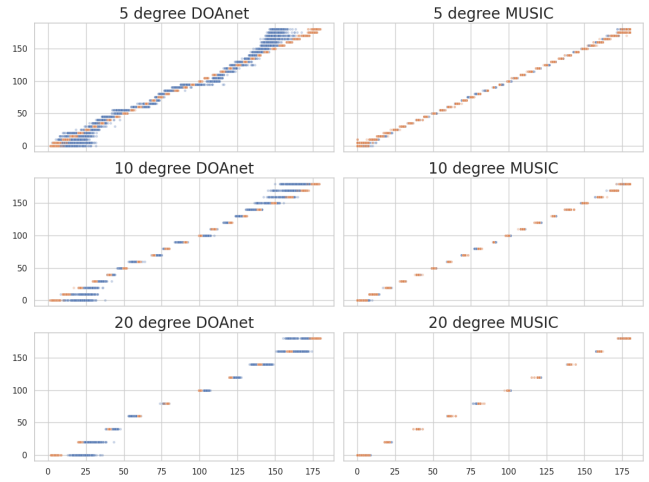

Fig. 5. Resulting plot of estimated DOA and target DOA using DOAE-Net and MUSIC. Blue and orange dots denote output signals of DOAC-Net and ground truth, respectively.

Table 1. Detailed PESQ, STOI, DOA prediction scores of DOAC-Net. For DOA prediction, difference and correlation are displayed.

	PESQ	STOI	DOAE-Net	MUSIC
5°	**3.81**	0.980	6.79/0.990	**2.84**/0.996
10°	3.70	**0.986**	**5.89/0.994**	2.85/**0.998**
20°	3.61	0.984	6.84/0.991	3.27/0.993

Next, we trained DOAC-Nets with different DOA intervals (5°, 10°and 20°), and compared the corresponding results as shown in Fig. 5; the blue dots represent the DOA estimated using DOAC-Net and the target DOA, whereas the orange dots represent the estimated DOA of the generated ground truth and the target DOA. Ideally, we would like to derive a 45°straight line to indicate that the proposed algorithm maps the multi-channel input to another with an assigned precise target DOA assigned. The results are also listed in Table 1, and the DOA prediction can be evaluated by the difference and the correlation. As shown above, indeed we obtain a distribution of the blue dots close to the diagonal. Another interesting finding is that the converted signals are more accurate for the MUSIC algorithm than for DOAE-Net, particularly at both ends 0°& 180°. We speculate that there is more distortion when the DOA conversion becomes larger. This distortion leads to poor model predictions.

4.2.2. Measuring Converted Audio Quality

To evaluate the fidelity of multi-channel converted audio signals, we first applied MVDR beamforming (from multi channels) to derive single-channel audio, and subsequently measured the PESQ and STOI scores, which are widely used as a speech quality and intelligibility metrics. As shown in Table 1, both PESQ and STOI acquired high scores, indicating that the signal still retains its high quality and intelligibility after conversion. We also compared the performances of different DOAs. Most of the scores improved mostly except for PESQ from 5° to 10° and all degrade from 10° to 20°. The reasons behind this trend is that for small DOA intervals, the model is relatively difficult to learn (distinguish) the difference between neighboring DOA signals, while for large DOA intervals, the DOA variability of training data is limited that cause the angle embedding difficult to learn the angle information. Therefore, 10° was determined to be the optimal DOA interval.

5. CONCLUSION

The DOA of audio signals is an important component for locating the desired sound source in an augmented or virtual environments. To improve users' hearing experiences in virtual games, concerts, or plays, the ability to convert DOAs instantaneously is crucial. Therefore, in this study, we proposed the DOAC-Net, a causal end-to-end DOA conversion system. In our model, speech signals were generated with the specified target DOA, while the content and quality of the received speech signals remained unchanged. Our results demonstrated that DOAC-Net effectively converts the DOA of multi-channel speech signals with minimal distortion. In the future, we will explore the use of visual signals to guide speech DOA conversion.

6. REFERENCES

[1] Zhizhang Chen, Gopal Gokeda, and Yiqiang Yu, *Introduction to Direction-of-arrival Estimation*, Artech House, 2010.

[2] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] Richard Roy and Thomas Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[4] Hongji Huang, Jie Yang, Hao Huang, Yiwei Song, and Guan Gui, "Deep learning for super-resolution channel estimation and doa estimation based massive mimo system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8549–8560, 2018.

[5] Rui Cheng and Changchun Bao, "Speech enhancement based on beamforming and post-filtering by combining phase information.," in *INTERSPEECH*, 2020, pp. 4496–4500.

[6] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong 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.

[7] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP 2018*. IEEE, pp. 351–355.

[8] Konrad Kowalczyk, Alexandra Craciun, Christian Dachmann, and Emanuël AP Habets, "Spatial perception of virtual xy recordings," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 129–133.

[9] Xiong Jing and Zi Cheng Du, "An improved fast root-music algorithm for doa estimation," in *Proc. ICASSP 2012*. IEEE, pp. 1–3.

[10] Debasis Kundu, "Modified music algorithm for estimating doa of signals," *Signal processing*, vol. 48, no. 1, pp. 85–90, 1996.

[11] Fei Wen, Qun Wan, Rong Fan, and Hewen Wei, "Improved music algorithm for multiple noncoherent subarrays," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 527–530, 2014.

[12] Feifei Gao and Alex B Gershman, "A generalized esprit approach to direction-of-arrival estimation," *IEEE signal processing letters*, vol. 12, no. 3, pp. 254–257, 2005.

[13] Xiaofei Zhang and D Xu, "Low-complexity esprit-based doa estimation for colocated mimo radar using reduced-dimension transformation," *Electronics Letters*, vol. 47, no. 4, pp. 283–284, 2011.

[14] Robert G Lorenz and Stephen P Boyd, "Robust minimum variance beamforming," *IEEE transactions on signal processing*, vol. 53, no. 5, pp. 1684–1696, 2005.

[15] Lawrence E Spence, Arnold J Insel, and Stephen H Friedberg, *Elementary linear algebra*, Prentice Hall, 2000.

[16] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. ICASSP 2015*. IEEE, pp. 2814–2818.

[17] Soumitro Chakrabarty and Emanuël AP Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. WASPAA 2017*. IEEE, 2017, pp. 136–140.

[18] Jack Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[19] Blair D Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Transactions on Aerospace and Electronic systems*, vol. 24, no. 4, pp. 397–401, 1988.

[20] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[21] Harry L Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2004.

[22] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP 2016*. IEEE, 2016, pp. 196–200.

[23] Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *Proc. ICASSP 2021*. IEEE, pp. 6089–6093.

[24] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *Proc. ICASSP 2017*. IEEE, pp. 5325–5329.

[25] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[26] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[28] Kirsten MacDonald Christophe Veaux, Junichi Yamagishi, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.