INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-channel Speech Enhancement Using Time-Domain Convolutional Denoising Autoencoder

Naohiro Tawara[1], *Tetsunori Kobayashi*[1], *Tetsuji Ogawa*[1]

[1]Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan

Abstract

This paper investigates the use of time-domain convolutional denoising autoencoders (TCDAEs) with multiple channels as a method of speech enhancement. In general, denoising autoencoders (DAEs), deep learning systems that map noise-corrupted into clean waveforms, have been shown to generate high-quality signals while working in the time domain without the intermediate stage of phase modeling. Convolutional DAEs are one of the popular structures which learns a mapping between noise-corrupted and clean waveforms with convolutional denoising autoencoder. Multi-channel signals for TCDAEs are promising because the different times of arrival of a signal can be directly processed with their convolutional structure, Up to this time, TCDAEs have only been applied to single-channel signals. This paper explorers the effectiveness of TCDAEs in a multi-channel configuration. A multi-channel TCDAEs are evaluated on multi-channel speech enhancement experiments, yielding significant improvement over single-channel DAEs in terms of signal-to-distortion ratio, perceptual evaluation of speech quality (PESQ), and word error rate.

Index Terms: Time-domain denoising autoencoder, dilated convolutional network, multi-channel speech enhancement

1. Introduction

Speech enhancement is an important technology for processing speech in noisy environments. Speech enhancement techniques are categorized into single- and multi-channel approaches. Multi-channel approaches, such as beam-forming [1], and multi-channel Wiener filter [2], generally perform better than single-channel approaches due to the utilization of spatial information that can be deduced from the time it takes a signal to reach differently placed microphones.

Recently, DNNs have been widely applied to both single- and multi-channel speech enhancement, where they have outperformed the traditional statistical approaches. The Denoising auto-encoder (DAE) is one popular DNN structure which learns the mapping between noisy and clean spectrograms from single-channel input [3]. DAEs were also applied to multi-channel observations [4, 5]. In [4], for example, phase and magnitude ratios between microphones are used as multi-channel information to enhance the accuracy of mapping between spectrograms. While these studies focus on frequency domains, several recent studies attempt to apply DAEs in the time-domain [6, 7, 8]. Time-domain DAEs make direct mappings between wave-forms instead of mapping between spectrograms in the frequency-domain DAEs. Time-domain DAEs have an advantage over frequency domain DAEs in that phase information does not need to be estimated explicitly.

In a prior work, following these successes, we applied time-domain DAE to single channel speech enhancement [8]. Here, we demonstrated significantly improved performance of DAE when an auxiliary signal such as an enhanced signal, enhance noise, and noise itself were provided. The highest performance was achieved when an oracle noise-signal (i.e. the true noise signal imposed to observation) was provided as an auxiliary signal. These results suggest that the time-domain DAE can learn a time-differential filtering to calculate the difference between signals from different inputs. This paper goes a step further, applying time-domain DAE to the multichannel speech enhancement aiming at learning time-differential arrival from specific directions to enhance signals from the target. In the proposed time-domain DAE, the multi-channel signals are taken as an input to a time dilated convolutional network, and converted into a single-channel denoised signal. Here, we demonstrate that denoised signals obtained by multichannel TCDAEs significantly improve as the number of channels is increased.

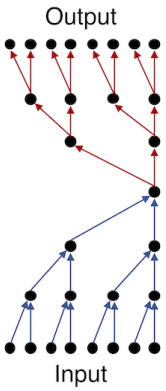


Figure 1: *One-dimensional dilated convolutions and deconvolutions in which shift- and window-size equal two. Each plot and array denote node and connection between nodes*

The rest of this paper is organized as follows. Section 2 explains the structure of a proposed TCDAE for multi-channel speech enhancement. Section 3 demonstrates the effectiveness of the proposed system on multi-channel speech signals. Finally Section 4 presents our conclusions.

2. Time-domain convolutional denoising autoencoders (TCDAEs)

This section gives a detailed explanation of TCDAEs for multi-channel speech enhancement.

2.1. Problem formulation

This research considers a multi-channel speech enhancement task in which a speech signal derived from a specific direction is enhanced. A C-channel noisy-corrupted signal at t-th frame

http://dx.doi.org/10.21437/Interspeech.2019-3197

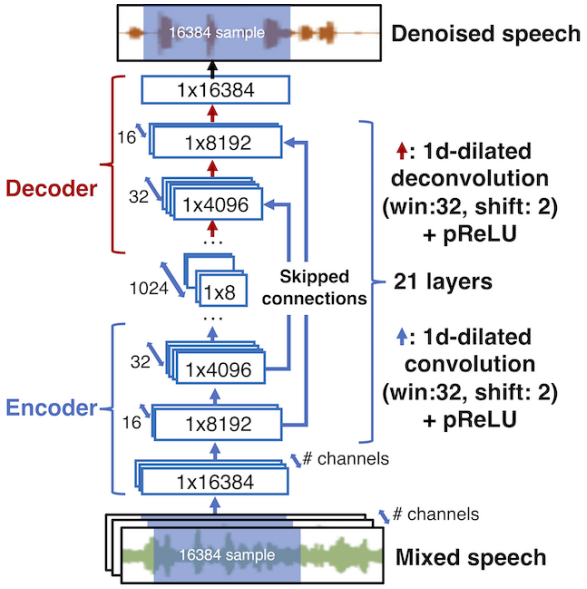


Figure 2: *Time-domain convolutional denoising autoencoder with spatial attention. Red and blue arrows represent one-dimensional dilated convolution and deconvolution, respectively.*

$\boldsymbol{Y}_t = [y_t^{(1)}, \cdots, y_t^{(C)}]$ is given as

$$\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{N}_t, \tag{1}$$

where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ denote the clean and noise signals, respectively. The our objective is to estimate the clean signal $\boldsymbol{X}_t$, attenuating the noise signal $\boldsymbol{N}_t$ for whole utterance $t = 1, \cdots, T$.

2.2. Structure of TCDAEs

The TCDAE is designed to generate a denoised one-second monaural waveform from one-second C-channel noisy waveforms using the whole one-second waveform as a context. Recurrent structure [9] is one possible technique to catch such context, but it requires a large number of parameters to be trained. Instead, a one-dimensional dilated convolutional structure[6, 10, 11] is used. Figure 2 shows an example of one dimensional dilated convolutions and deconvolutions in which window-and shift- size equal two.

Specifically, the t-th frames of the C-channel waveform with surrounding W frames $\boldsymbol{X}_{t:t+W} \in \mathbb{R}^{C\times 1\times W}$ are convoluted into H_1 nodes by a first s-dilated convolutional layer with a weight $\boldsymbol{W}_1^{\mathrm{enc}} \in \mathbb{R}^{H_1\times 1\times W}$ and a bias $\boldsymbol{b}_1^{\mathrm{enc}} \in \mathbb{R}^{H_1}$ as follows:

$$\boldsymbol{h}_t^{\mathrm{enc}_1} = g(\boldsymbol{W}_1^{\mathrm{enc}} \tilde{\boldsymbol{X}}_{s\cdot t:s\cdot t+W} + \boldsymbol{b}_1^{\mathrm{enc}}), \tag{2}$$

where g denotes a non-linear function and parametric ReLU (pReLU)[12] is used in this study. In addition, virtual batch normalization [13] is applied after each convolution to make the optimization faster. Then, obtained H_1-dimensional embedded frames $\boldsymbol{h}_t^{\mathrm{enc}_1}$ are repeatedly convoluted into higher layers as follows:

$$\boldsymbol{h}_t^{\mathrm{enc}_j} = g(\boldsymbol{W}_j^{\mathrm{enc}} \boldsymbol{H}_{s\cdot t:s\cdot t+W}^{\mathrm{enc}_{j-1}} + \boldsymbol{b}_j^{\mathrm{enc}}), \tag{3}$$

where j denotes a layer's index. Here, note that the number of nodes in higher layers gets smaller by shifting the window with shift-size s. Finally Embedding frames at the j-th layer $\left\{\boldsymbol{h}_1^{\mathrm{enc}_j}, \cdots, \boldsymbol{h}_{\frac{16384}{s^j}}^{\mathrm{enc}_j} \mid \boldsymbol{h}_t^{\mathrm{enc}_j} \in \mathbb{R}^{H_j}\right\}$ are obtained.

The embedded vector $\boldsymbol{h}_J$ which is finally obtained is taken as an input for the decoder. Each layer converts the nodes from lower layer with one-dimensional deconvolution with window of w-length shifting by s-steps. Here, in order to make the decoding easy, skipped connections from the corresponding encoding layer get concatenated with the output from the lower layer as proposed in U-net [14]. The output from the j-th decoding layer is as follows:

$$\boldsymbol{h}_t^{\mathrm{dec}_j} = g(\boldsymbol{W}_j^{\mathrm{dec}} \boldsymbol{H}_{s\cdot t:s\cdot t+W}^{\mathrm{dec}_{j-1}} + \boldsymbol{b}_j^{\mathrm{dec}}), \tag{4}$$

where $\boldsymbol{W}_j^{\mathrm{dec}} \in \mathbb{R}^{H_j\times 2\times W}$ and $\boldsymbol{b}_j^{\mathrm{dec}} \in \mathbb{R}^{H_j}$ denote the weight and bias of the j-th decoding layer. $\boldsymbol{H}_t^{\mathrm{dec}_{j-1}} = \left[\boldsymbol{H}_t^{\mathrm{dec}_{j-1}}, \boldsymbol{H}_t^{\mathrm{enc}_{j-1}}\right] \in \mathbb{R}^{H_{j-1}\times 2\times W}$ denotes an matrix concatenating outputs from the $(j-1)$-th decoding and encoding layers.

Figure 2 shows the whole network structure used in this study. Encoder and decoder are composed of 11 dilated convolution and deconvolution, respectively. In the encoder, the depth of the filter increases layer by layer and an eight-dimensional feature map is obtained at the bottle-neck with a depth of 1024. In the decoder, the depth of the filter decreases layer by layer and a one-dimensional feature with the same length as the input is obtained. The time-length $\times$ depth of the outputs of the layers are $16384 \times C, 8192 \times 16, 4096 \times 32, 2048 \times 32, 1024 \times 64, 512\times 64, 256\times 128, 128\times 128, 64\times 256, 32\times 256, 16\times 512$, and 8×1024, respectively.

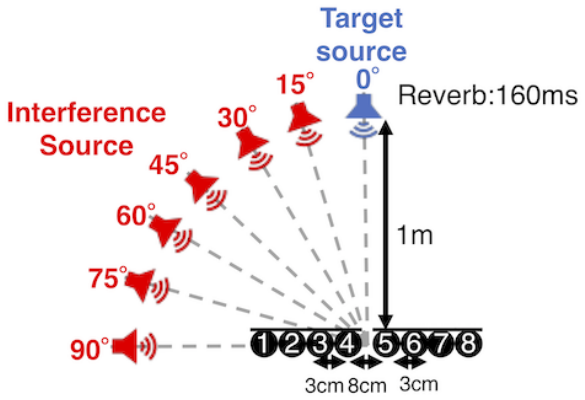


Figure 3: *Experimental environment with microphone array, a target source, and an interference source.*

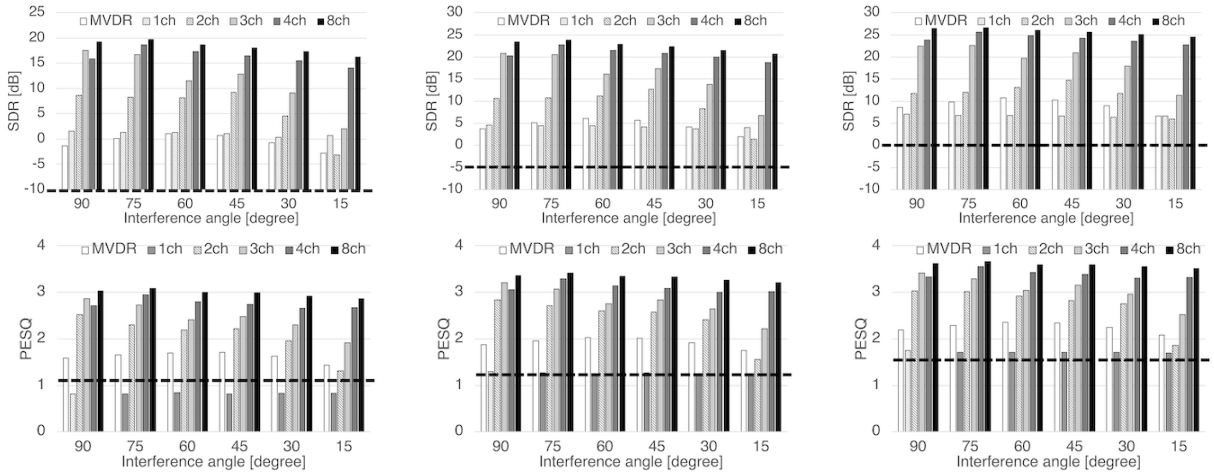
3. Multi-channel speech enhancement experiments

Multi-channel speech enhancement experiments were conducted using various combination of interference directions, type of noises, segments, and speakers.

3.1. Experimental setup

3.1.1. Speech material

Figure 3 shows the experimental environment. The target source was placed in front of eight channel microphones and the



(a) -10dB (b) -5dB (c) 0dB

Figure 4: *Speech enhancement performance of developed systems. Dashed line shows the SDR and PESQ scores evaluated on unprocessed noise-corrupted signals.*

Table 1: *Channel IDs used for multi-channel TCDAE. Each ID corresponds to one described in Figure 3*

# Channels	Channel ID
ch1	4
ch2	4,5
ch4	3,4,5,6
ch6	2,3,4,5,6,7
ch8	1,2,3,4,5,7,8

Table 2: *Interfering noise recorded. Noise signals are selected from the JEIDA noise database.*

DB id	noise type	use
09	exhibition hall (booth)	training
11	exhibition hall (aisle)	training
13	station (concourse)	training
14	station (aisle)	training
18	factory (machine)	training
20	factory (metal)	training
26	street	training
28	intersection	training
30	crowd	testing
47	elevator hall	testing

interference source was placed at a specific angle selected from -90 to 90 degrees excepting 0 degrees. The distance between each source and the microphones was 1 m and the microphones were spaced 3 cm apart except for a gap in the middle of 8 cm. The dry sources of noise and speech signals were convoluted with impulse responses taken from the multi-channel impulse response database (MIRD) [15]. Reverberation time was 160 ms. The mixed signals consisted of convoluted speech mixed with noise signals at five signal-to-noise ratios (SNRs) of -10, -5, 0, 5, and 10 dB.

For the training set, the dry sources of speech were 8000 utterances spoken by 78 females and 78 males randomly selected from the Japanese newspaper article sentences read speech corpus (JNAS) [16]. Approximately 50 different sentences were included for each speaker and noise condition. For the testing data, the dry sources of speech were 500 utterances spoken by 7 females and 7 males, including approximately 35 different utterances for each speaker and noise condition. The dry sources of noise were eight types of noise for training and two for testing selected from the JEIDA noise corpus [17]. These are listed in Table 2. Note that there is no over-wrap between types of speakers or noise conditions between the training and test sets.

To evaluate each method in more realistic environments, a data set was crated from the chime3 corpus [18]. All Utterances contained in tr05-simu set was used to train TCDAEs, and utterances in dt05-simu and et05-simu sets were used as a development and test sets, respectively.

3.2. Evaluation metrics

To evaluate the quality of enhanced signals, a signal distortion rate (SDR) between the estimated and the clean speeches was calculated using the BSS Eval toolbox [19]. A perceptual evaluation of speech quality (PESQ), based on the ITU standard P.862 [20], was used to measure the perceptual performance.

The word error rate (WER) was also calculated to evaluate the speech recognition performance. The 39-dimensional Mel-frequency Cepstral Coefficients (MFCCs) with their Δ and $\Delta\Delta$ parameters was used for the input feature. Linear discriminant analysis and feature-space maximum likelihood linear regression were applied to the MFCC features. Five layered DNNs with 1024 hidden nodes in each layer were used to calculate the posterior probabilities of the triphone states. The DNNs were trained using 31396 speech utterances selected from ASJ-JNAS. They were trained using clean input, and no techniques were used to adapt to noisy conditions. We used a phoneme bi-gram language model trained with 20000 speech utterances selected from ASJ-JNAS. For training and decoding, Kaldi [21] was used.

3.3. Evaluation items

The following four denoising systems were compared:

Table 3: *Word error rate (%) for each interference directions when SNR equals (left) -10 dB and (right) 0 dB.*

Method		Interference direction					
		90°	75°	60°	45°	30°	15°
Clean		10.7					
Noisy		82.6					
MVDR		78.7	76.2	76.9	77.5	79.1	80.4
TCDAE	1ch	67.7	68.2	68.4	68.1	68.8	68.8
	2ch	25.1	28.6	66.4	41.9	47.8	80.8
	4ch	16.7	18.5	34.2	28.4	44.5	72.2
	6ch	20.5	15.9	18.0	19.8	24.4	24.6
	8ch	**15.3**	**15.2**	**15.7**	**16.3**	**18.3**	**19.0**

Method		Interference direction					
		90°	75°	60°	45°	30°	15°
Clean		10.7					
Noisy		75.9					
MVDR		50.5	40.5	40.7	43.9	50.3	60.9
TCDAE	1ch	52.9	53.0	52.9	53.2	53.6	55.0
	2ch	18.0	19.7	36.6	27.4	29.3	57.7
	4ch	14.2	15.4	21.5	18.8	25.5	40.6
	6ch	15.9	13.6	14.8	15.7	17.8	18.2
	8ch	**15.4**	**13.0**	**13.7**	**14.4**	**15.7**	**16.4**

- **beamformit** TDOA-based beamformer [22];
- **MVDR:** minimum variance distortionless regression (MVDR) beamformer using eight channels;
- **TCDAE-1ch:** single-channel TCDAE [6]; and
- **TCDAE-{2-8}ch:** multi-channel TCDAEs using multiple channels. Channel IDs used are listed in Table 1.

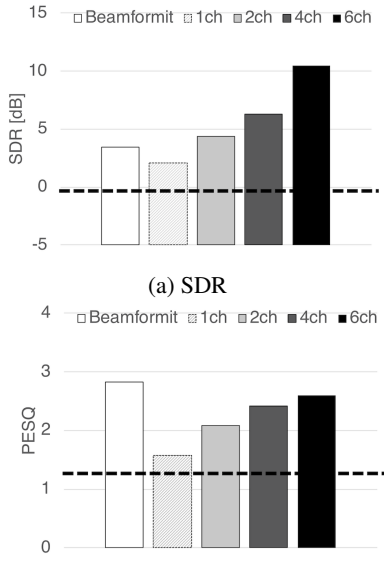
3.4. Experimental results

Figure 4 shows the speech enhancement performance evaluated with SDR and PESQ score averaged over 500 test utterances. Dashed line in each figure shows the scores evaluated on unprocessed mixed signals. This result shows that the enhanced signals obtained from the multi-channel TCDAE (i.e. ch $>=$ 2) yielded a significant improvement over single-channel TCDAE. The performance of TCDAE-2ch, however, was worse than that of TCDAE-1ch when the angle between target and interference sources was 15 degrees at SNR:-10dB. This was because the model could learn the time difference of arrival between microphones, but the difference was too small. The performance of multi-channel TCDAEs improved with the number of microphones used. For angles between sources of 15 degrees the performance degradation was attenuated by the use of more than six-channels. This result demonstrates the effectiveness of the use multi-channel information for TCDAE. Table 3 summarizes the WER calculated on clean, noise-corrupted, and enhanced signals with MVDR, single-channel TCDAE, and multi-channel TCDAEs. From this result, we can see that the performance of MVDR and single-channel TCDAE significantly deteriorated especially when the angle between target and interference sources was small. By contrast, multi-channel TCDAE keeps low WER for all conditions. These results show that multi-channel TCDAEs works well the noise source is placed near the target source.

Figures 5 (a) and (b) show the SDR and PESQ evaluated on chime3 dataset, respectively. These figures show that the proposed 8ch-TCDAE outperformed single-channel TCDAE and beamformit in SDR. However, the PESQ score obtained by 8ch-TCDAE is worse than beamformit. This is because due to the difficultly to training TCDAEs with the complex experimental environment in chime3 (e.g. the direction of target and interference frequently changed), and distorted signals were generated. Table 4 lists the WER evaluated on chime3 dataset. This result shows the TCDAE with 8ch-microphones outperformed the beamformit.

Table 4: *Word error rate (%) evaluated on chime3 dataset.*

Set	Noisy	beamformit	TCDAE-6ch
Dev	52.9	39.1	**24.8**
Test	66.6	61.6	**36.4**



(a) SDR

(b) PESQ

Figure 5: *Speech enhancement performance of developed systems evaluated on chime3 data. Dashed line shows the evaluation for mixed observation.*

4. Conclusions

This study proposed and developed a multi-channel time-domain convolutional denoising autoencoder (TCDAE) and evaluated its speech enhancement performance in a multi-channel configuration. The TCDAE directly maps noisy speech signals into clean signal in the time-domain, aiming to learn spatial information in an end-to-end manner. The multi-channel TCDAE was compared with a single-channel TCDAE and an MVDR beamformer in a simulated environments, and showed significant improvements in terms of SDR, PESQ, and WER. Experiments were also conducted with the chime3 dataset to evaluate the multi-channel TCDAE in a more realistic environment. Experimental results indicated that SDR and WER were significantly improved over single-channel TCDAE and beamformit while the obtained signals were relatively distorted compared to those obtained in the simulated environment.

5. References

[1] M. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2001.

[2] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[4] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 116–120.

[5] 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.

[6] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[7] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[8] N. Tawara, T. Kobayashi, M. Fujieda, K. Katagiri, T. Yazu, and T. Ogawa, "Adversarial autoencoder for reducing nonlinear distortion," in *APSIPA*, 2018, pp. 1669–1673.

[9] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2401–2405.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[15] "Multi-channel impulse response database," https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database/.

[16] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[17] S. Itahashi, "A noise database and japanese common speech data corpus," *The Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991.

[18] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[21] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[22] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.