Interspeech 2024
1-5 September 2024, Kos, Greece

MSDET: Multitask Speaker Separation and Direction-of-Arrival Estimation Training

Roland Hartanto[1], *Sakriani Sakti*[2], *Koichi Shinoda*[1]

[1]Tokyo Institute of Technology, Japan, [2]Nara Institute of Science and Technology, Japan

roland@ks.c.titech.ac.jp, ssakti@is.naist.jp, shinoda@c.titech.ac.jp

Abstract

The information on the spatial location of speakers can be effectively used for multi-channel speaker separation. For example, Location-Based Training (LBT) uses the order of azimuth angles and distances of speakers to solve the permutation ambiguity problem. This location information can be used to improve the separation performance further. This paper proposes a multitask learning approach, **M**ultitask Speaker **S**eparation and **D**irection-of-Arrival **E**stimation **T**raining (MSDET), jointly optimizing speaker separation and Direction-of-Arrival (DoA) estimation. In our evaluation using SMS-WSJ dataset, it outperforms LBT by 0.13 points in SI-SDR and 0.35 points in ESTOI.

Index Terms: speaker separation, DoA estimation, multitask learning

1. Introduction

Speaker separation is the task of separating individual speaker voices from a mixture of multiple speakers' voices. It has been extensively studied for monaural and multi-channel speech processing [1, 2, 3]. Monaural separation uses signals recorded by a single microphone and relies solely on their spectral information. Multi-channel separation utilizes signals captured by a microphone array and leverages spatial cues of sources, leading to better separation performance [4, 5, 6, 7]. Deep learning based multi-channel speech separation has been extensively studied [4, 5, 6, 8, 9, 10, 11, 12]. Early studies have applied the successful monaural speaker separation methods to multi-channel cases, where they have attempted to estimate masks to separate speakers in spectral (time and frequency) domains [4, 5, 6, 8, 9, 10]. Recent studies directly estimate speech's real and imaginary frequency components from input mixture, called complex spectral mapping [11, 12, 13, 14], which further enhances separation performance.

In speaker-independent speech separation, assigning each of the separator's outputs to its corresponding speaker is necessary for the model training. For this purpose, in monaural speaker separation, deep clustering [1] attempts to learn time-frequency embeddings with a speaker permutation invariant objective function and estimate an ideal binary mask for each speaker by K-means clustering on the embeddings. Another technique is permutation invariant training (PIT) [3], widely used for monaural speech separation and adopted for multi-channel cases. It computes separation losses between all possible pairs of outputs and ground truths and then chooses the pair with the minimum loss. Since it has a factorial complexity, it becomes costly as the number of speakers increases.

Recently, Location-Based Training (LBT) [13] is proposed to solve this permutation ambiguity problem. It assigns each output to its corresponding speaker by using the order of its location: its azimuths and distance. It decreased the computational costs from PIT and achieved better results. While this method is significantly effective, it utilizes spatial location information only to solve the permutation ambiguity problem. We believe this information can be further utilized to improve the separation performance.

Another way to use spatial location information is a multitask learning approach. It has been proven to be effective in the training of multiple related tasks (e.g., facial landmark detection and head pose estimation [15]). Recently, multitask learning of speech separation and Direction-of-Arrival (DoA) estimation has been proposed for extracting a single speaker from a mixture of multiple speakers [16]. The same extraction process must be repeated for each speaker, which is computationally expensive and may be redundant to separate all the speakers from each other.

In this paper, we propose **M**ultitask Speaker **S**eparation and **D**irection-of-Arrival **E**stimation **T**raining (MSDET), a multitask learning method of speaker separation and DoA estimation. It solves the permutation ambiguity problem and further improves the separation performance using the DoA information. Different from [16], this method can simultaneously extract each speaker in a mixture. It utilizes a multitask loss, a weighted sum of separation loss and DoA estimation loss in the training phase. In our evaluation using SMS-WSJ dataset, it improved the separation performance over LBT by 0.13 points in Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) and 0.35 points in the extended short-time objective intelligibility (ESTOI).

2. Previous Studies

2.1. Deep-learning-based Multi-channel Speaker Separation using Complex Spectral Mapping

Complex spectral mapping estimates speech's real and imaginary frequency components from an input mixture [11, 12, 13, 14]. They are both essential [17] since they are both affected by environmental interferences.

Taherian et al. [13] employ Dense-UNet architecture proposed in [18] for complex spectral mapping. Dense-UNet consists of 9 densely connected convolutional blocks. The first 5 blocks (encoder) are interleaved with downsampling layers to map the input into higher dimension features. The other four blocks (decoder) are interleaved with upsampling layers so the model outputs have the same resolution as the input. This model employs skip connections that link the blocks in the encoder and the decoder at the same level. Each dense block consists of 5 convolutional layers. The middle layer of each dense block acts as a frequency mapping layer that models each frequency band.

10.21437/Interspeech.2024-2537

Recently, Wang et al. [12] proposed TF-GridNet architecture for complex spectral mapping. TF-GridNet consists of several blocks. Each block contains two layers of BLSTM followed by a self-attention module. The first BLSTM is utilized to model the full band of frequencies in each frame. The second BLSTM models the temporal information of each frequency band. The self-attention module learns the long context temporal relationships between frames in an utterance. The output of a block is fed to the next block, and the output of the final block is fed to a deconvolution layer to estimate the real and imaginary frequency components of the separated speech signals. Full band modeling and full utterance temporal modeling in each frequency band lead to the strong separation performance of TF-GridNet.

2.2. Deep-learning-based Multi-channel Speaker Separation using DoA information

Some other studies have used location information obtained by microphone arrays in addition to the spectral cues. For example, J. Wechsler et al. [7] perform speaker separation on pre-defined regions, assuming each region contains one speaker. An early study [19] uses DoA ground truth as an input for the separation model to improve the separation performance. This study assumes that DoAs are available in inference, but usually, they aren't easy to obtain. C. Han et al. [20] employ a DoA estimator model as a front end for speech separator. They first estimate the DoAs of all speakers and use them as additional input for the speech separation model. Its downside is that the speech separation performance depends on the DoA estimator performance. A previous study on unsupervised source separation [21] introduces spatial loss, which uses speakers' DoA estimated using a DoA estimator to constrain the demixing matrix estimation.

2.3. Location Based Training (LBT)

Assigning the outputs of the speaker separation model to the correct ground truths is crucial in training DNN-based speaker separators. The training will fail to converge if it is not assigned correctly. This problem is called permutation ambiguity.

For multi-channel speaker separation, a method called LBT [13, 22] leverages the DoA and speaker-microphone distance to solve this problem. It simply assigns each speaker to each of the outputs using the order of the ground truth DoA and distance in the training phase. While this method successfully reduces the computational cost of PIT [3], it utilizes the location information only to solve the permutation invariant problem. We believe this information can be more exhaustively used to improve the separation performance further.

2.4. Multitask Learning

Multitask learning [23] is a technique to optimize a model on multiple related tasks simultaneously. It enables learning a shared representation through information introduced from diverse tasks, thereby enhancing the performance of each task. Multitask learning is effective because training multiple related tasks concurrently allows each task to gain advantages from the training signals present in the other tasks. It has been widely applied in many fields of study, such as natural language processing [24], computer vision [15], speech processing [25].

Sun et al. [16] applied multitask learning to speech separation and source localization for a single target speaker. While yielding improvement in separation performance, this method cannot be directly utilized to separate multiple speakers. It needs a speaker recognizer to identify the target speaker. It also has to perform N times inference for N speakers.

3. Multitask Speaker Separation and Direction-of-Arrival Estimation Training (MSDET)

3.1. Architecture

We apply multitask learning by pairing each speaker with its corresponding DoA in the output layers and form a unit to estimate both the spectrum and DoA for each source. We focus on estimating DoA, assuming different speakers' DoAs cannot be the same. Our method can be applied to various speaker separation methods. We implemented our multitask learning method with two different methods in our evaluation. One is DenseUNet, which is used in [13], and the other is TF-GridNet [12].

Fig 1 illustrates our system when the number of speakers is two. To further improve the separation performance, different from [13] and [12], this method has a unit that includes a deconvolution layer and a DoA estimator for each source. The deconvolution layer outputs the speech spectrum and features for DoA estimation. The DoA estimation layer consists of a convolutional layer with a ReLU activation function followed by a linear layer. Each DoA estimator is trained by using a classification task. The linear layer size is the number of DoA classes determined by $(360/r)$, where r is the DoA resolution. In each class, a softmax function calculates the score of DoA output from the linear layer. The DoA class with the highest score is chosen.

3.2. Multitask Loss

We define a multitask loss for the multitask training, a weighted average of the speech separation loss and DoA estimation loss. We use L1 loss between the ground truth and the generated signals in the frequency domain for the separation loss, as defined in [13]. The DoA estimation loss is the cross-entropy loss between the output of the DoA estimator and the ground truth DoA. The DoA ground truth is a soft target probability distribution defined in [26].

Let L_{sep} and L_{DoA} be separation and DoA estimation losses, and w_{DoA} be the weight of DoA estimation loss. The multitask loss can be expressed as follows:

$$L_{\text{Multitask}} = (1 - w_{\text{DoA}})L_{\text{sep}} + w_{\text{DoA}}L_{\text{DoA}}. \quad (1)$$

3.3. Training procedure

We train this model from scratch. First, we assign the speech output of the first unit to one of the target speakers with the smallest DoA difference in degrees. Then, excluding this pair of the unit and the target speaker, We similarly assign the second unit to its target speaker. We repeat this process until each unit is aligned with its target speaker.

It is worth noting that the training process can also be started by training with only DoA estimation loss and then continued with the multitask loss. The training may need extra epochs. In this work, we conduct a straightforward approach for training which applies the multitask loss from the beginning.

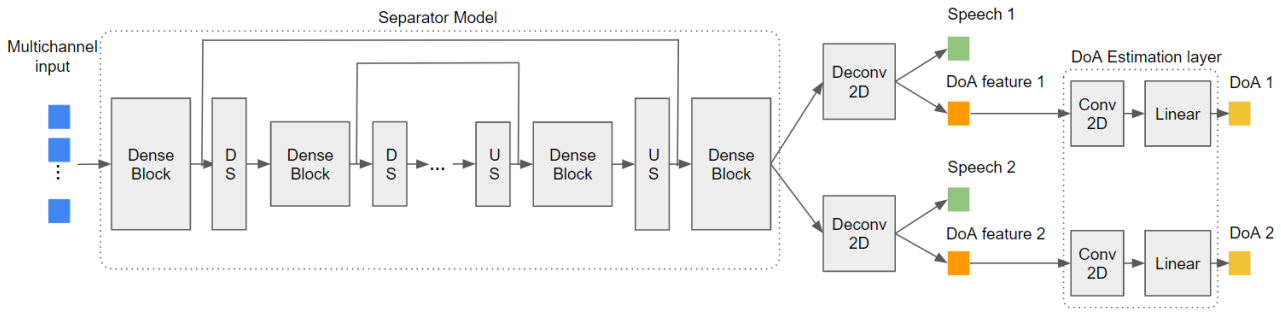

Figure 1: *Architecture of the proposed method based on Dense-UNet. DS and US blocks are downsampling and upsampling layers, respectively.*

4. Experiments

4.1. Dataset

We use SMS-WSJ [27], a simulated dataset with reverberation for multi-channel source separation derived from the WSJ corpus. The dataset contains two-speaker speech mixtures. It simulates a circular microphone array with six microphones. Its radius is 10 cm. The room impulse responses (RIRs) are simulated by using the reverberation time (RT60) ranging from [0.2, 0.5] s and the various room dimensions. The distance range between the speech sources and the array is [1,2] m. This dataset also includes white noise, added to the mixture with the signal-to-noise ratio sampled from the range [20, 30] dB. The dataset consists of 33561 samples for training, 928 samples for validation, and 1332 samples for evaluation. The sampling rate is 8 kHz.

4.2. Settings

We follow [13] for the Dense-UNet model implementation. We set the convolutional kernel size to 76. We set the DoA resolution to $r = 1$, thus the DoA estimation layer size is 360. We extract short-time Fourier-Tranform (STFT) features as model inputs from speech mixture. The analysis window is Hanning window, with a length of 32 ms and a hop length of 8 ms. We train all models with the same number of epochs for fair comparisons. The maximum number of epochs is 100. The learning rate is 0.0001, and the optimizer is Adam. We divide each mixture into 4-second chunks for training. We also use TF-GridNet implemented on the ESPNet toolkit [28] and follow [12] for the model and training settings. We set the learning rate to 0.001 and the number of TF-GridNet blocks to 4.

For multitask training, we assigned the weight of 0.05 and 0.01 to the DoA estimation (w_{DoA}) loss for Dense-UNet and TF-GridNet, respectively. According to our preliminary experiments, the DoA estimator requires fewer epochs than the speech separator to achieve convergence. Thus, we assign a smaller weight to the DoA estimation loss.

We use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [29] and the extended short-time objective intelligibility (ESTOI) [30] to evaluate speech separation performance. Higher SI-SDR and ESTOI are better. We also evaluate the DoA estimation performance by measuring the mean absolute error (MAE) of degrees. Lower MAE means better DoA estimation.

Table 1: *Speech separation performance on Dense-UNet*

System	SI-SDR (dB)	ESTOI (%)
Dense-UNet + PIT [13]	11.25	88.03
Dense-UNet + LBT (DoA) [13]	12.56	90.12
Dense-UNet + LBT (DoA) (ours)	13.12	89.56
Dense-UNet + LBT (DoA+distance) [13]	13.22	**90.96**
Dense-UNet + Multitask (Proposed)	**13.25**	89.91

Table 2: *Speech separation performance on TF-GridNet*

System	SI-SDR (dB)	ESTOI (%)
TF-GridNet + PIT [12]	19.90	96.60
TF-GridNet + LBT (DoA)	21.48	**97.20**
TF-GridNet + Multitask (Proposed)	**21.57**	**97.20**

4.3. Results and Discussion

4.3.1. Speech Separation Results on Dense-UNet

Tab. 1 shows the result. We implemented LBT [13] and obtained the SI-SDR of 13.12, 0.56 points higher than the result reported in [13], and the ESTOI of 89.56%, 0.56 points lower than ESTOI of 89.56% in [13]. Our proposed method, MSDET, achieved an SI-SDR of 13.25 and an ESTOI of 89.91. It is better than LBT by 0.13 points in SI-SDR and 0.35 points in ESTOI, which confirms its effectiveness.

We present the speech separation performance in different DoA differences between speakers in Tab. 3. In this evaluation, we first calculate the differences of DoAs between the two speakers using their DoA labels. Then, we categorize them into five classes according to their values. Finally, we calculate the average SI-SDR for each class. MSDET outperforms the separation performance of LBT (Dense-UNet + LBT (DoA)) in all classes. The separation performance is enhanced significantly when the DoA differences are less than 5 degrees, achieving an SI-SDR improvement of 2.25 points. Thus, our proposed method, MSDET, effectively separates speakers in general and those with small DoA differences.

Table 3: *Speech separation performance (SI-SDR) in various DoA differences (d) between speakers on Dense-UNet*

System	SI-SDR (dB)				
	$d < 5^\circ$	$5^\circ \leq d < 10^\circ$	$10^\circ \leq d < 20^\circ$	$20^\circ \leq d < 40^\circ$	$d \geq 40^\circ$
Dense-UNet + LBT (DoA) (ours)	6.33	12.13	12.31	13.00	13.46
Dense-UNet + Multitask (Proposed)	**8.58**	**12.19**	**12.37**	**13.01**	**13.53**

Table 4: *Speech separation performance (SI-SDR) in various DoA differences (d) between speakers on TF-GridNet*

System	SI-SDR (dB)				
	$d < 5^\circ$	$5^\circ \leq d < 10^\circ$	$10^\circ \leq d < 20^\circ$	$20^\circ \leq d < 40^\circ$	$d \geq 40^\circ$
TF-GridNet + LBT (DoA)	18.39	**19.90**	**20.21**	21.08	21.77
TF-GridNet + Multitask (Proposed)	**18.58**	19.75	20.13	**21.10**	**21.88**

Table 5: *DoA estimation results on Dense-UNet*

System	MAE (degrees)
Dense-UNet (DoA estimation loss)	0.70
Dense-UNet + Multitask (Proposed)	**0.40**

Table 6: *DoA estimation results on TF-GridNet*

System	MAE (degrees)
TF-GridNet (DoA estimation loss)	0.64
TF-GridNet + Multitask (Proposed)	**0.49**

4.3.2. Speech Separation Results (TF-GridNet)

We present the experiment results in Tab. 2. On TF-GridNet architecture, we implemented LBT (TF-GridNet + LBT (DoA)) and obtained an SI-SDR of 21.48 and an ESTOI of 97.20. Our proposed method further enhances the performance of TF-GridNet + LBT (DoA) by achieving an SI-SDR of 21.57, which is higher by 0.09 points. The ESTOIs are the same.

In different DoA differences between speakers (Tab. 4), MSDET exhibits the separation performance improvement over LBT (TF-GridNet + LBT (DoA)) in most cases. Consistent with the experiment result on Dense-UNet (Tab. 3), it achieves the highest SI-SDR improvement (by 0.19 points) when the DoA differences are less than 5 degrees. The separation performance drops when the DoA differences are between 5 and 20 degrees. However, there are more separation improvements in other DoA differences, leading to a better separation performance.

4.3.3. DoA Estimation Results

We also examine the effectiveness of multitask learning for estimating DoA. Tab. 5 and Tab. 6 show the results. For this purpose, we trained Dense-UNet and TF-GridNet using only DoA estimation loss (Dense-UNet (DoA estimation loss) and TF-GridNet (DoA estimation loss)). We obtained the MAE of 0.70 degrees and 0.64 degrees, respectively. On Dense-UNet, our proposed multitask learning method, MSDET, improves DoA estimation by achieving the MAE of 0.40 degrees, which is 0.30 points lower than Dense-UNet (DoA estimation loss). On TF-GridNet, it also enhances DoA estimation by achieving the MAE of 0.49 degrees, which is 0.15 points lower than TF-GridNet (DoA estimation loss).

5. Conclusion

We have proposed MSDET, a multitask learning approach for speech separation and DoA estimation. Our method fully leverages the DoA information to improve the separation performance. It outperforms LBT by 0.13 points in SI-SDR and by 0.35 points in ESTOI. Future works include improving separation performance in real recording conditions and handling the increase in the number of speakers.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23H00490 and NEC Corporation.

7. References

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[3] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM TASLP*, vol. 25, no. 10, 2017. [Online]. Available: https://doi.org/10.1109/TASLP.2017.2726762

[4] Z. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM TASLP*, vol. 27, no. 2, 2019.

[5] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[6] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, 2020, pp. 6394–6398.

[7] J. Wechsler, S. R. Chetupalli, W. Mack, and E. A. P. Habets, "Multi-microphone speaker separation by spatial regions," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6389–6393.

[9] H. Chen, Y. Yi, D. Feng, and P. Zhang, "Beam-Guided TasNet: An iterative speech separation framework with multi-channel output," in *Proc. Interspeech*, 2022, pp. 866–870. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-230

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *CSL*, vol. 46, pp. 605–626, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S088523081630122X

[11] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM TASLP*, vol. 29, pp. 2001–2014, 2020.

[12] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[13] H. Taherian, K. Tan, and D. Wang, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM TASLP*, vol. 30, pp. 2791–2800, 2022.

[14] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. Wang, "Multi-input Multi-output Complex Spectral Mapping for Speaker Separation," in *Proc. INTERSPEECH 2023*, 2023, pp. 1070–1074.

[15] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *ECCV*, 2014, pp. 94–108.

[16] W. Sun, M. Wang, and L. Qiu, "Spatial aware multi-task learning based speech separation," *arXiv:2207.10229*, 2022.

[17] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *Proc. Interspeech*, 2014, pp. 1623–1627.

[18] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 12, p. 2092–2102, dec 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2941148

[19] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE SLT*, 2018, pp. 558–565.

[20] C. Han and N. Mesgarani, "Online binaural speech separation of moving speakers with a wavesplit network," in *Proc. ICASSP*, 2023, pp. 1–5.

[21] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in *Interspeech*, 2022, pp. 241–245.

[22] H. Taherian and D. Wang, "Multi-resolution location-based training for multi-channel continuous speech separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[23] R. Caruana, "Multitask learning," *Machine Learning 28, 41–75*, 1997.

[24] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *Proc. 53rd Annu. Meeting of the ACL and 7th IJCNLP*, Jul. 2015, pp. 1723–1732. [Online]. Available: https://aclanthology.org/P15-1166

[25] W. Hou et al., "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Proc. Interspeech*, 2020, pp. 1037–1041.

[26] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *CSL*, vol. 75, p. 101360, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000079

[27] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv:1910.13934*, 2019.

[28] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[29] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630. [Online]. Available: https://api.semanticscholar.org/CorpusID:53246666

[30] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.