論文 / 著書情報
Article / Book Information

Title	SepVAC: Multitask Learning of Speaker Separation, Speaker Localization, Microphone Array Localization, and Room Acoustic Parameter Estimation in Various Acoustic Conditions
Authors	Roland Hartanto, Sakti Sakriani, Koichi Shinoda
Citation	Proc. Interspeech 2025, , , pp. 2480-2484,
Pub. date	2025, 8
Copyright	(c) 2025 International Speech Communication Association, ISCA
DOI	https://doi.org/10.21437/Interspeech.2025-2784

SepVAC: Multitask Learning of Speaker Separation, Speaker Localization, Microphone Array Localization, and Room Acoustic Parameter Estimation in Various Acoustic Conditions

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

[1]Institute of Science Tokyo, 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

This paper proposes a multitask learning method for speech separation, that **Sep**arates speech and estimates the recording conditions in **V**arious **A**coustic **C**onditions (SepVAC) jointly. Unlike the previous methods that aim to achieve robustness against the uncertainty caused by noise and reverberation, this method explicitly estimates speaker & microphone location and room acoustic parameters to disambiguate them from speech features. We introduce curriculum learning to learn the model parameters stably. In our evaluation using SMS-WSJ-Plus dataset, it outperforms the state-of-the-art SpatialNet baseline by 0.67 points in word error rate (WER).

Index Terms: speech separation, room acoustic parameter estimation

1. Introduction

Speaker separation is a technique to extract an individual speaker's voice when multiple people speak simultaneously [1, 2, 3]. While monaural speech separation relies only on spectral features from a single microphone, multi-channel speech separation utilizes both spectral features and spatial features captured by microphone arrays [4, 5, 6, 7, 8]. While it is expected to perform better, it is vulnerable to differences in the recording condition, including speaker and microphone location and room acoustics.

Nowadays, most multi-channel speaker separation methods have employed deep learning. Early methods use the same approach as monaural separation; they estimate masks in the time-frequency domain to separate speakers [4, 5, 6, 9, 10, 11]. Recent approaches [12, 13, 14, 15, 8] have tried to exclude the influence of noise and reverberation effectively. For example, TFGridNet [13] utilizes two models, one for separation and the other for dereverberation. Recently, C. Quan, 2024, [8] proposed SpatialNet that uses only one model for both of these and achieves state-of-the-art performance with less computational cost [13, 8].

Another approach to mitigate the influence of the recording condition is to identify the recording condition given and use it to exclude the impact of noise and reverberation. For example, several studies [16, 17, 14] show that the explicit use of the direction-of-arrival(DoA), the speaker direction relative to the microphone array, has improved separation. Hartanto *et al.* [18] propose multitasking learning of speech separation and DoA estimation and confirm its effectiveness. However, in these studies, only the relative position of speakers and microphones are estimated, and their positions in a room are unknown. Hence, they may not be effective in different room acoustics conditions.

On the other hand, some studies have explored room acoustic parameter estimation such as the volume [19, 20], geometry [21], total surface area of the room [20], the average absorption coefficient of room surfaces [20, 21], and reverberation time [21, 22, 23]. W. Yu, 2021, [21] performs their estimation using room impulse response (RIR), which is much simpler than speech signals. A. F. Genovese, 2019 [19], K. Zheng, 2022, [22], and Wu, 2017 [23] estimate them using speech signals from a single speaker. P. Srivastava, 2021, [20] performs their estimation using speech signals from multiple speakers. Wu, 2017 [23] simultaneously estimates reverberation time for dereverberation but only estimates room acoustic parameters and a source position. However, these studies do not examine dereverbaration for speech separation.

In this paper, we propose multi-task learning of speaker **Sep**aration and estimation of recording condition, including speaker location, microphone array localation, and room acoustics parameters in **V**arious **A**coustic **C**onditions (SepVAC). These factors are deeply entangled in the multi-channel source signals, and thus, it may not be easy to estimate them simultaneously. We employ curriculum learning to solve this problem. Evaluated on a highly reverberated dataset, SMS-WSJ-Plus, it outperforms the state-of-the-art SpatialNet by 0.67 points in word error rate (WER).

2. Previous Studies

2.1. Deep-learning-based Multi-channel Speaker Separation

Recent deep learning multi-channel speaker separation methods estimate speech real and imaginary frequency components from an input mixture [12, 13, 14, 15, 8]. They also simultaneously perform denoising and dereverberation by using clean speech as their training target [24]. Z. Q. Wang, 2023 [13], and C. Quan, 2024 [8] perform separation by exploiting cross-band and narrow-band information of frequency components. Narrow-band modeling captures spatial cues within a single frequency band, as phase differences between channels remain stable over time but vary across frequencies if the source is stationary. This also helps in handling reverberation, which differs across frequencies. Cross-band modeling, on the other hand, learns spectral patterns across frequencies at each time frame.

Z. Q. Wang, 2023 [13] proposes TFGridNet, which utilizes two networks: a separator network and a dereverberation network. The second network performs dereverberation for each source one by one using the input mixture and the separated speech signals from the separator. It performs separation and dereverberation in two stages, which is costly.

C. Quan, 2024 [8] proposes SpatialNet, which performs separation and dereverberation in one stage. SpatialNet consists of cross-band and narrow-band blocks arranged alternately. A narrow-band block contains a multi-head self-attention module,

leading to strong spatial information modeling and achieving state-of-the-art separation performance. SpatialNet is suitable for dealing with reverberation, which is related to spatial cues. Therefore, we choose SpatialNet as the backbone of our proposed method.

However, these separation methods rely on the spectral information of clean speech as training targets. It limits their ability to model the spatial information and deal with various recording conditions. In contrast, our method incorporates recording condition parameters, helping the model recognize room acoustics and improve spatial information modeling.

2.2. Speaker Separation utilizing Direction-of-Arrival

The direction of the speech signal from a speaker relative to the center of the microphone array (DoA), can help improve speech separation [16, 17, 14, 18]. For example, Location-Based Training (LBT) [14] leverages the direction-of-arrival (DoA) of speakers to organize target speech based on their DoA for loss computation. MSDET [18] performs multitask learning of speaker separation and DoA estimation, further exploiting DoA information. Multitask learning [25] enhances the performance of individual tasks, leveraging information from multiple tasks to learn a shared representation. MSDET performs separation by adding additional output to the separator backbone to output the DoA of each speaker.

Speech signals captured by microphones vary based on the DoA, distance, and the microphone array's position in the room. For instance, a microphone array in the center picks up different reverberation than one in a corner. Room size and wall materials also affect reverberation, making sounds seem to come from multiple directions in highly reverberant spaces. As a result, DoAs alone cannot fully handle diverse acoustic conditions.

2.3. Room Acoustic Parameter Estimation

Some previous works focus on estimating room acoustic parameters. Some of them estimate one parameter, and some others estimate multiple parameters simultaneously. A. F. Genovese, 2019, [19] focuses on estimating the volume of the room. K. Zheng, 2022 [22] focuses on reverberation time (RT60) estimation employing a speech dereverberation model. Wu, 2017 [23] simultaneously estimates reverberation time and performs speech dereverberation. P. Srivastava, 2021, [20] deals with volume of the room, total surface area of the room, and the average absorption coefficient of room surfaces. W. Yu, 2021, [21] estimates the room geometry, absorption coefficient, and reverberation time. These studies assume unknown microphone and source positions. However, the location of the microphone array in the recording room, along with its relative position to the speakers, affects the reverberation patterns. Since previous studies have not considered these factors, including them in room acoustic parameter estimation is necessary.

3. SepVAC

3.1. Architecture

We train a speaker separation model by performing multitask learning of separation (SS), speaker localization (SL), microphone array localization (ML), and room parameters estimation (RP). We employ SpatialNet [8] as our speaker separator backbone.

We present our proposed system in Fig. 1. Unlike [8], this architecture has a unit for SS and SL for each speaker, named speaker output unit, and a unit for ML and RP, named room acoustics unit. The speaker output unit consists of a linear layer to output speech spectrum and features for SL and a linear layer to output speaker location in (x, y), which is the coordinate relative to the center of the microphone array. We put SS and SL tasks in the speaker output unit to avoid permutation between the separated speech signals and speaker locations. The room acoustics unit consists of a linear layer to output features for ML and RP, a linear layer to output microphone array center coordinate in (x, y), and a linear layer to output room parameters. The x-coordinate denotes the distance from the microphone array center to the nearest longer wall, while the y-coordinate denotes the distance to the nearest shorter wall. The room parameters include reverberation time, early decay time, volume, surface area, width, length, average absorption coefficient of room surfaces, direct-to-reverberant ratio, and clarity. Although separate units handle speakers and room acoustics, the separator backbone shares parameters across tasks, allowing it to learn room acoustics while distinguishing them from speech features. We also add a convolutional layer to reconstruct the input mixture using the separated speech, the features of SL, ML, and RP.

3.2. Loss Function

The multitask loss is the weighted sum of the SS, SL, ML, and RP losses. The SS loss is the negative Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) loss between the ground truth and the separated speech in the time domain. The SL, ML, RP, and reconstruction losses are the mean squared error (MSE) loss. The multitask loss can be expressed as follows:

$$\begin{aligned} L = {} & w_{\text{SS\_SL}}(w_{\text{SS}}L_{\text{SS}} + w_{\text{SL}}L_{\text{SL}}) + w_{\text{ML}}L_{\text{ML}} + w_{\text{RP}}L_{\text{RP}} \\ & + (1 - w_{\text{SS\_SL}} - w_{\text{ML}} - w_{\text{RP}})L_{\text{recons}}. \end{aligned} \tag{1}$$

3.3. Curriculum Learning

Training our model to perform separation and estimate recording condition parameters is challenging since our model needs to simultaneously deal with four different tasks. Training all tasks simultaneously from scratch leads to degradation in all tasks. To solve this problem, we use curriculum learning by training the model in several steps. The first step trains each task iteratively. We train SS and SL and freeze the output layer parameters of ML and RP for j epochs. The loss weights applied are w_{SS} and w_{SL}. Tasks not yet introduced to the model have their loss weights set to zero. Next, we train ML, freeze the parameters of SS, SL, and RP, and then similarly train RP. When ML is introduced while freezing the other tasks, the applied weights include w_{SS}, w_{SL}, and w_{ML}. This prevents the model from forgetting the SS and the SL tasks. Once RP is introduced and the other tasks are frozen, all loss weights from Equation 1 are applied. After that, we repeat the process from training SS-SL, ML, to RP for k iterations. The second step trains two tasks simultaneously for each iteration. The training starts from SS-SL to ML-RP for l iterations. Finally, all tasks are trained simultaneously until convergence.

The task order also affects the model training. In the first and the second steps, SS and SL are trained together since they are directly related to the speakers. In our preliminary experiments, we found that training SS before SL causes SL to fail, and training SL before SS causes SS to fail. ML and RP are trained separately in the first step. Simultaneously training them initially causes ML to fail.

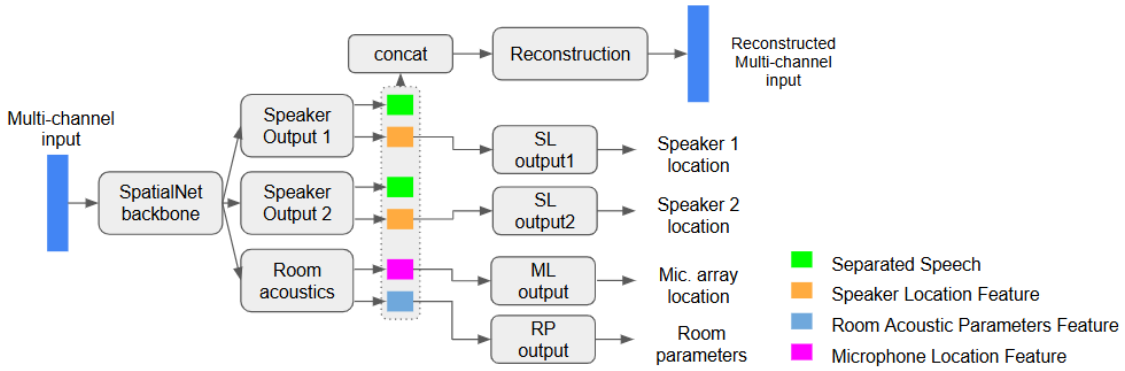

Figure 1: *Architecture of the proposed method based on SpatialNet.*

4. Experiments

4.1. Dataset

We utilize SMS-WSJ-Plus [8], a simulated dataset for noisy and reverberant speaker separation. It is an extension of SMS-WSJ dataset [26] with wider room dimensions, reverberation time (RT60), and speaker-to-microphone distance ranges. It contains two-speaker speech mixtures. It simulates a circular array with radius of 10 cm and six microphones. The reverberation time range is [0.1, 1.0] s. The room length and width range are [4, 10] m. The room height range is [3, 4] m. The distance between the speaker and the microphone array range is [1, 4] m. The dataset includes babble and white noises with a signal-to-noise ratio (SNR) of [0, 20] dB. The number of samples in the dataset is 33561 samples for training, 928 samples for validation, and 1332 samples for evaluation. The sampling rate is 8 kHz.

We also conduct an ablation study to evaluate the contribution of each task in the proposed method. In this study, we train all models with a subset of SMS-WSJ-Plus dataset, containing 2000 samples for training. The models are evaluated on the SMS-WSJ-Plus test dataset.

4.2. Experiment Settings

We follow [8] for the SpatialNet backbone configuration. We extract short-time Fourier-Tranform (STFT) features from speech mixture for the model inputs. The Hanning window is used as the analysis window, with a length of 32 ms and a hop length of 8 ms. We train the model by performing seven iterations of the first step with ten epochs for each task in each iteration, three iterations of the second step. Finally, the model is trained until convergence with maximum total number of epochs (including the first and the second steps) of 600. We set the learning rate to 0.001 and the number of blocks in SpatialNet to 8 (SpatialNet-small). The SS and SL tasks are trained using Permutation Invariant Training (PIT).

For multitask training, we assigned the weight of 0.9, 0.1, 0.91, 0.03, and 0.03 for w_{SS}, w_{SL}, $w_{\text{SS\_SL}}$, w_{ML}, and w_{RP}, respectively. We assign smaller weights for SL, ML, RP, and reconstruction loss since the main task is SS. In the first two steps of curriculum learning, the learning rate starts at 0.001 and halves when there is no improvement. In the third step, it begins at 0.0005 and decays exponentially by 0.99 per epoch until the maximum number of epochs.

Table 1: *Speech separation performance on SMS-WSJ-Plus dataset*

System	SI-SDR (dB)	WER (%)
SpatialNet + PIT [8]	**13.8**	15.89
SpatialNet + Proposed	13.4	**15.22**

Table 2: *Speech separation performance on mixtures with low reverberation (RT60 $<$ 0.5 s)*

System	SI-SDR (dB)	WER (%)
SpatialNet + PIT	**15.9**	12.64
SpatialNet + Proposed	14.9	**12.08**

Table 3: *Speech separation performance on mixtures with high reverberation (RT60 $\geq$ 0.5 s)*

System	SI-SDR (dB)	WER (%)
SpatialNet + PIT	12.3	18.48
SpatialNet + Proposed	**12.3**	**17.30**

4.3. Evaluation Metrics

We use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [27] and the Word Error Rate (WER) to evaluate speech separation performance. For SI-SDR, the higher value is better. For WER, the lower value is better. We evaluate in WER to observe how our method benefit automatic speech recognition (ASR). The ASR model used for this evaluation is the hybrid Hidden Markov Model with Temporal Deep Neural Network (HMM-TDNN) provided on Kaldi [28].

4.4. Results and Discussion

We show our experiment results in Table 1. We present separation results for two models: (1) SpatialNet + PIT, trained with only PIT and (2) SpatialNet + Proposed, our proposed method. Our proposed method achieves slightly lower SI-SDR than SpatialNet + PIT by 0.4 points. However, it achieves better WER by 0.67 points.

Our proposed method separates speaker outputs from room acoustic features. The room acoustic features represent the re-

Table 4: *Ablation study*

Tasks				SI-SDR	WER (%)
SS	SL	ML	RP	(dB)	
✓	✓	✓	✓	**11.9**	20.15
✓	-	✓	✓	**11.9**	**18.91**
✓	✓	-	✓	11.4	20.32
✓	✓	✓	-	11.3	19.88

Table 5: *Speaker localization error*

System	Euclidean Distance (m)
SL only	0.93
All	**0.30**

Table 6: *Microphone array localization error*

System	Euclidean Distance (m)
ML only	1.01
All	**0.83**

verberation information. Improvements are mostly observed when separating mixtures with high reverberation (RT60 $\geq$ 0.5 s), as shown in Table 3. Extracting reverberation information from the input yields less reverberant speech than SpatialNet + PIT. It leads to better separation results in WER, although the SI-SDR is slightly lower. SI-SDR degrades most when separating mixtures with low reverberation (RT60 $<$ 0.5 s) (Table 2). On the other hand, SI-SDR is the same when separating mixtures with high reverberation (RT60 $\geq$ 0.5 s). In low reverberation, ML and RP does not perform well. Low reverberation makes room acoustics harder to identify, though speech remains intelligible with slight WER improvement. SI-SDR stays above average the SI-SDR values obtained by both SpatialNet + PIT and SpatialNet + Proposed in Table 1. This explains the slight SI-SDR drop despite significant WER improvement.

4.5. Ablation Study

We also perform an ablation study to see the contribution of each task in the multitask learning. The results are presented in Table 4. The loss weight of a task that is not included in training (marked with "-") is set to zero.

The proposed method achieves the SI-SDR of 11.9 dB and the WER of 20.15%. The best performance is obtained when ML and RP are included in training, with the SI-SDR of 11.9 dB and WER of 18.91%. These results show the effectiveness of including ML and RP in multitask training. However, SL degrades the separation performance. Since SL is part of the speaker output in the architecture, separation performance is sensitive to errors in speaker location estimation.

We investigate the benefit of our proposed method for SL, ML, and RP. The results are presented in Table 5, Table 6, and Table 7 for SL, ML, and RP, respectively. We compare our proposed method with models trained exclusively on SL loss (SL only), ML loss (ML only), and RP loss (RP only). The results indicate that our proposed method enhances SL, ML, and RP.

5. Conclusion

We have proposed SepVAC, a multitask learning method for speech separation, that **Sep**arates speech and estimates recording conditions in **V**arious **A**coustic **C**onditions. It estimates speaker and microphone positions, along with room acoustics, to distinguish them from speech features. We also introduce curriculum learning to ensure stable model training. It outperforms the state-of-the-art SpatialNet by 0.67 points in WER. Future works include handling more speakers and various noises in the environment.

Table 7: *Room acoustic parameter estimation error*

Parameters	Mean Absolute Error (MAE)	
	RP only	All
RT60 (s)	0.15	**0.08**
Early decay time (s)	0.67	**0.62**
Volume (m^3)	30.92	**27.63**
Surface Area (m^2)	27.23	**23.23**
Width (m)	0.84	**0.70**
Length (m)	0.95	**0.87**
Absorption coefficient	0.09	**0.09**
Direct to Reverberant ratio (dB)	2.58	**1.58**
Clarity (dB)	3.79	**3.18**

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23H00490.

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] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.

[9] 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.

[10] 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

[11] 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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] R. Hartanto, S. Sakti, and K. Shinoda, "Msdet: Multitask speaker separation and direction-of-arrival estimation training," in *Interspeech 2024*, 2024, pp. 2170–2174.

[19] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 231–235.

[20] P. Srivastava, A. Deleforge, and E. Vincent, "Blind room parameter estimation using multiple multichannel speech recordings," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 226–230.

[21] W. Yu and W. B. Kleijn, "Room acoustical parameter estimation from room impulse responses using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021.

[22] K. Zheng, C. Zheng, J. Sang, Y. Zhang, and X. Li, "Noise-robust blind reverberation time estimation using noise-aware time–frequency masking," *Measurement*, vol. 192, p. 110901, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224122001828

[23] Y. M. L. K. e. a. Wu, B., "A reverberation-time-aware dnn approach leveraging spatial information for microphone array dereverberation," *EURASIP J. Adv. Signal Process*, vol. 81, 2017.

[24] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[26] 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.

[27] 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

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.