Experimental Evaluation of Speech Enhancement for In-Car Environment Using Blind Source Separation and DNN-based Noise Suppression

Yutsuki Takeuchi*, Taishi Nakashima*, Nobutaka Ono*,
Takashi Takazawa†, Shuhei Shimanoe†, and Yoshinori Tsuchiya†
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: takeuchi-yutsuki@ed.tmu.ac.jp
† MIRISE Technologies, Aichi, Japan

***Abstract*—In this study, we evaluate speech enhancement using multi-channel blind source separation (BSS) and single-channel DNN-based noise suppression in an in-car environment. Specifically, blind source separation is applied to a mixture of target speech and driving noise, and one of the outputs is fed into the DNN. This combination allows for speaker position-independent enhancement through blind source separation, while the single-channel DNN can be trained regardless of the microphone position. To further improve performance, we fine-tune the pre-trained model using the BSS outputs, including driving noise. Our evaluation criteria include signal-to-distortion ratio improvement (SDRi) for noise reduction performance, and accuracy and character error rate (CER) for speech recognition performance. Results demonstrate that SDRi for the proposed method reached 28.37 dB. Additionally, the average speech recognition accuracy improved from 0.785 to 0.826, and CER improved from 15.3% to 11.9%. These findings demonstrate the potential of combining BSS and DNN to enhance speech recognition capabilities in in-car environments.**

I. Introduction

Speech recognition technology has served as a vital interface for various systems, enhancing safety and convenience. In in-car environments, voice-based interfaces are particularly crucial as they allow drivers to operate systems without taking their hands off the steering wheel. Additionally, the development of speech recognition systems for rear seats or while reclined has gained attention owing to the difficulty of physically reaching the controls [1]. However, the in-car environment includes both stationary noises, such as engine and wind noise, and nonstationary noises, such as sounds from the car stereo and conversations among passengers. These noise sources significantly degrade the effectiveness of speech recognition technologies, making reliable noise reduction techniques essential.

When a microphone array is available, beamforming [2] is one of the valuable methods to enhance the speech coming from a target direction. Some previous studies demonstrate this technique [3]–[5]. However, although the driver's position can generally be assumed to remain static, the relative positions of speakers in the rear seats to the microphones may vary depending on the car type and seating arrangement. As a result, reliance solely on prior information is not feasible, and source direction estimation and steering vector estimation become necessary.

In contrast, multichannel blind signal separation (BSS) [6]–[8] works without prior information about the sound sources, making it a flexible framework for noise and speech separation. Thanks to an efficient update rule of parameters [9], this approach requires less computation, and online real-time applications have been developed [10]–[12]. Some previous works show the effectiveness of BSS in car environments [13], [14]. However, because it does not use prior information about sound sources, it may not perform well in complex noisy in-car environments.

Additionally, deep neural networks (DNNs) have been applied to single-channel or multichannel speech enhancement [15]–[19]. Related works show the high separation performance using DNNs in an in-car environment [20], [21]. Although most DNN-based source-separating processing is nonlinear and unsuitable for speech recognition, they can achieve higher noise reduction performance.

Motivated by these considerations, our focus in this study is to improve speech recognition accuracy in car environments by enhancing speech using microphone arrays. To achieve this, we explore a method that combines multi-channel blind source separation (BSS) and single-channel DNN-based speech enhancement. This approach enables enhancement that is independent of the speaker's position through blind source separation, while the single-channel DNN can be trained without regard to the microphone position. Additionally, by using BSS as a frontend, we expect to potentially eliminate the influence of multiple speakers and car stereo sounds, although this aspect is not evaluated in this study. To achieve better performance, we fine-tune the pre-trained model using paired data of BSS outputs and clean speech. We used auxiliary-function-based independent vector analysis (AuxIVA) [9] as the BSS method, as it is a standard multichannel BSS method, and considering future online evaluations [10], [11], [22], [23]. We also used Conv-TasNet [24] as the DNN model. The experimental results show the effectiveness of combining multi-channel BSS and single-channel DNN with fine-tuning.

II. Problem Settings

Let N be the number of microphones. Let $\boldsymbol{s}_t \in \mathbb{C}^N$ and $\boldsymbol{n}_t \in \mathbb{C}^N$ be the multi-channel observation of clean target speech signal and noise in the time domain, where t is the index of the discrete time. The observed signal $\boldsymbol{x}_t \in \mathbb{C}^N$ can be represented as $\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{n}_t$.

The objective of this study is to recognize speech in an in-car environment. Given that the speech recognition engine used in this study is treated as a black box and no retraining of the engine is considered, our aim is to enhance the speech as much as possible at the frontend through speech enhancement, that is, to estimate s_t from $\boldsymbol{x}_t$.

III. Proposed Approach

A. Combination of Multichannel BSS and Single Channel DNN-based Speech Enhancement

Recent studies demonstrate the effectiveness of DNN for speech enhancement, but in the context of speech recognition, DNN-based speech enhancement is not always effective due to nonlinear filtering, especially when the speech recognition engine is not retrained. To overcome this, we consider the combination of multi-channel BSS and single-channel speech enhancement with DNN. Fig. 1 shows the block diagram of the process. By applying a single-channel speech enhancement with DNN to the output of multichannel BSS, we expect less nonlinear distortion. Previous works show the effectiveness of the combination of DNN and source separation [25].

Let $\boldsymbol{X}_{lk} \in \mathbb{C}^N$ be the multi-channel observation in the time-frequency domain, which is obtained by short-time Fourier transform (STFT) from $\boldsymbol{x}_t$, where l and k are the indices of the time frames and frequency bins, respectively. The source separation by multi-channel BSS can be represented as

$$\boldsymbol{Y}_{lk} = \boldsymbol{W}_k \boldsymbol{X}_{lk} \tag{1}$$

where $\boldsymbol{W}_k \in \mathbb{C}^{N \times N}$ is the demixing matrix estimated by BSS at frequency index k and $\boldsymbol{Y}_{lk} \in \mathbb{C}^N$ denotes the separated signal. Note that $\boldsymbol{W}_k$ works as a time-invariant linear filter.

From $\boldsymbol{Y}_{lk}$, we select the separated signal that contains speech and obtain the time-domain signal y_t by applying inverse STFT (ISTFT). Note that y_t is a single-channel signal, not a multi-channel signal. We then apply DNN-based single-channel speech enhancement to y_t.

$$\hat{s}_t = f(y_t) \tag{2}$$

where $f(\cdot)$ represents the DNN-based single-channel speech enhancement.

B. Fine-Tuning of DNN Using BSS Outputs

Generally, it is essential to train DNN models on highly generalizable datasets and to guarantee performance on untrained data. However, in this study, we use BSS outputs as input data for the DNN, and pre-trained models may not perform adequately on such BSS outputs. Therefore, we propose fine-tuning the pre-trained model with BSS outputs. To verify the effectiveness of this approach, we compare two types of fine-tuning:

Fine-tuning A (FT-A):
As fine-tuning for driving noise in an in-car environment, the DNN is retrained by using mixed signals of speech and driving noise as inputs and clean speech as outputs.

Fine-tuning B (FT-B):
First, multi-channel observations of speech and driving noise are processed by BSS to obtain separated signals. These BSS outputs are then used as inputs, and clean speech as outputs, to retrain the DNN.

For both types of fine-tuning, the dataset is divided into training and validation sets. The DNN is retrained such that the loss function on the training data is minimized. Training is stopped when the loss function value on the validation data starts to increase, indicating overfitting.

IV. Experiments

A. Data Acquisition

In this study, evaluation test data were created from speech sources recorded in an anechoic room, impulse responses, and driving noise recorded in an actual car using a microphone array.

The vehicle Alphard (Toyota, 2022) was used as the actual car. In the car, a microphone array was fixed on the ceiling between the 1st row seats and the 2nd row seats (Fig. 2). The microphone array consisted of eight high-SNR MEMS microphones (TDK InvenSense, SNR=$74\,\mathrm{dBA}$), and the distance between microphones was $6\,\mathrm{mm}$ (Fig. 3). Analog–digital converters (ADCs) were mounted on the microphone printed circuit board (PCB) and controlled by a microcomputer (μC) to synchronize all microphone channel samplings. The time-stretched pulse (TSP) playback system and microphone array were independent regarding sampling timing. On the other hand, eight microphone channels were synchronized with each other (Fig. 4).

To create the target speech signal, 48 utterances of speech sources spoken by two males and two females were recorded in an anechoic room. The utterances likely to be spoken in a car situation were chosen.

Impulse responses were measured using a head and torso simulator (HATS; Type4128, Brüel & Kjær) and the TSP method. A TSP was played from HATS, which was mounted on the 2nd seat of the parked car, and recorded by the microphone array. The recorded TSPs were converted to impulse responses by convolving inverse TSPs. Because the position of HATS changes in accordance with the seat position, the evaluation seat positions were set to three conditions at the 2nd seat (Fig. 5). Pos. 1 is the nearest position between the microphone array and HATS. Pos. 2 is the normal seating position. Pos. 3 is the farthest position between the microphone array and HATS in the 2nd seat. Noise data were recorded in a car driving at $100\,\mathrm{km/h}$ on the Tokai–Kanjo Expressway (Aichi, Japan) and extracted to obtain steady noise with neither road connections nor over-taking cars.

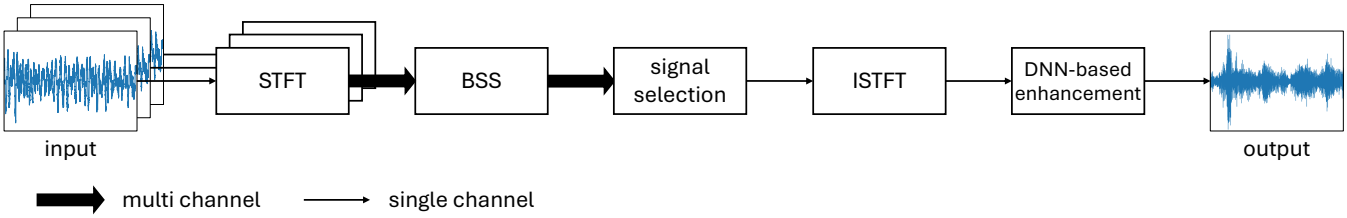

Fig. 1. Block diagram of processing

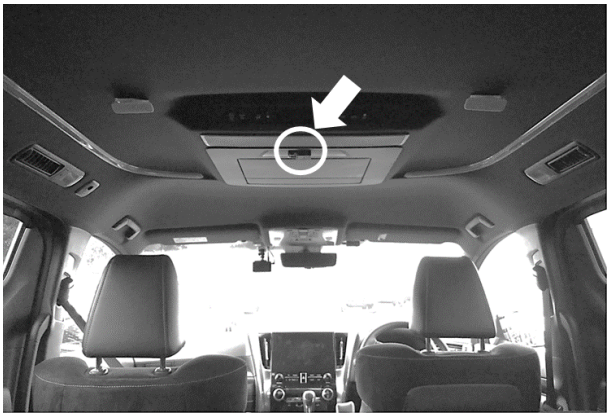

Fig. 2. Position of microphone array

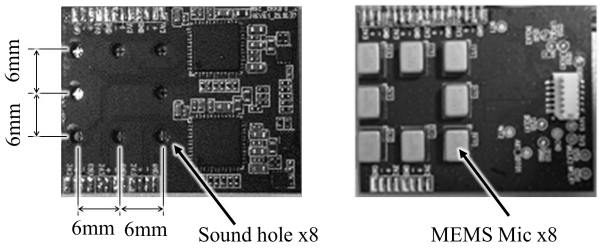

Fig. 3. Details of microphone array

The speech sources, recorded at a 44.1 kHz sampling, were downsampled to a 16 kHz sampling rate. The TSP responses were recorded at the 16 kHz sampling rate, so the impulse responses were also at a 16 kHz sampling rate. The target speech signal was created by convolving the impulse response. The amplitude of the target speech signal was compensated for so that the amplitude was almost the same as that of an actual

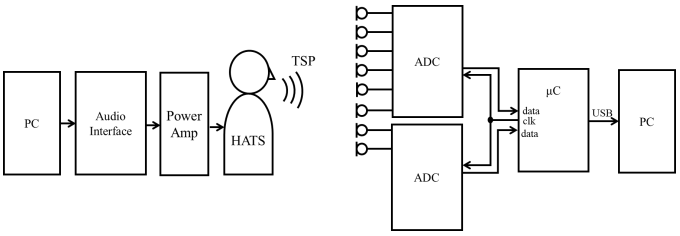

Fig. 4. Block diagram of TSP response recording

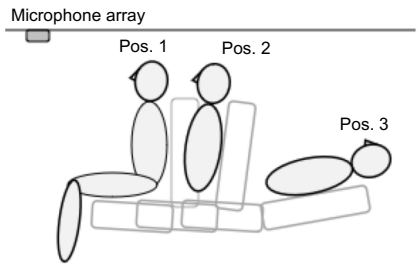

Fig. 5. Seat positions for impulse responses measurement

human voice in the car. The driving noise was also recorded at a 16 kHz sampling rate. The TSP response and driving noise were recorded in the same car and with the same microphone fixture.

B. Methods

The evaluation test data were created by superposing the target speech and driving noise signals with signal amplitude compensation to simulate the actual car noise level. As a result of this postprocess, three conditions with the same driving noise level, the same speech source, and different transfer function of evaluation test data were created. The length of these evaluation test data was 24 s and included 1–3 s of speech.

The training data for fine-tuning described in III-B were also created with these data. We prepared training data by separating the evaluation test data using AuxIVA, and ground truth with clean speech signal. From four speakers, we selected two speakers for training set, one for validation set, and the other for test set.

In this study, we used Conv-TasNet [24] for the training model. This model consists of three blocks: Encoder, Separation, and Decoder. The encoder module transforms the input mixture signal into intermediate representations for each segment. And for them, the separation module applies the mask vector to separate sources. Then, the decoder module reconstructs the signal.

We used the Conv-TasNet model pre-trained with Libri1Mix dataset [26], [27]. We selected this model for its compatibility with the dataset used in this study and its sufficient amount of training. We fine-tuned this model with the training set

TABLE I
EXAMPLES OF SPEECH CONTENT AND MEANING

Speech content	Meaning
Onryo wo ageru.	Turn up the volume.
Annai chushi.	Stop the navigation guidance.

described in IV-A.

For the BSS methods applying DNN as post-processing described in III-A, we employed AuxIVA based on prior experiments. The STFT frame length was set to 4096 samples, and the number of iterations was fixed at 100. In this study, we manually selected the separated signal containing speech from multiple separated signals. Developing an automatic selection method is planned for future work. To reduce low-frequency noise, we first applied a high-pass filter with a cut-off frequency of $100\,\mathrm{Hz}$, and then performed DNN-based post-processing.

Furthermore, to investigate the conditions under which the proposed approach is superior, we evaluated the performance with signal-to-noise ratio (SNR) in $5\,\mathrm{dB}$ increments, ranging from $-10\,\mathrm{dB}$ to $-30\,\mathrm{dB}$.

We used signal-to-distortion ratio improvement (SDRi) [28] to evaluate the performance of noise reduction. The result is obtained by subtracting the SDR before separation from the SDR after separation.

To evaluate speech recognition performance, we defined speech recognition accuracy as the percentage of correct speech recognition results. Since command speech is often used in car navigation systems, and the speech recognition result cannot be said to be correct if the actual action does not match the content of command speech, so we used speech recognition accuracy to evaluate speech recognition performance in this study. As a well-used speech recognition engine with high accuracy, we used Google's software (Cloud Speech-to-Text V1) for speech recognition and manually checked whether the recognition content matched the speech content. Table I shows the examples of speech content.

We also calculated character error rate (CER) to objectively evaluate the speech recognition performance. CER can be computed by dividing the sum of the number of substitutions, deletions, and insertions by the number of characters in the reference. We evaluated the performance after extracting speech segments manually.

C. Results

Fig. 6 shows the SDRi results for the driving noise condition. FT stands for Fine-Tuning. The average SDRi in ideal MaxSNR, which is designed by using true data, is $33.76\,\mathrm{dB}$ in pos. 1, and this is considered the limit value of the linear time-invariant filter. In contrast, the average SDRi value in AuxIVA+DNN with FT-B result in pos. 1 was $29.74\,\mathrm{dB}$, and the average SDRi in all conditions was $28.37\,\mathrm{dB}$. SDRi improved by about $10\,\mathrm{dB}$ compared with the AuxIVA results. This result would be attributed to the fact that Conv-TasNet, the nonlinear model, can better interpret more complex source characteristics and nonstationary noise. The characteristic to

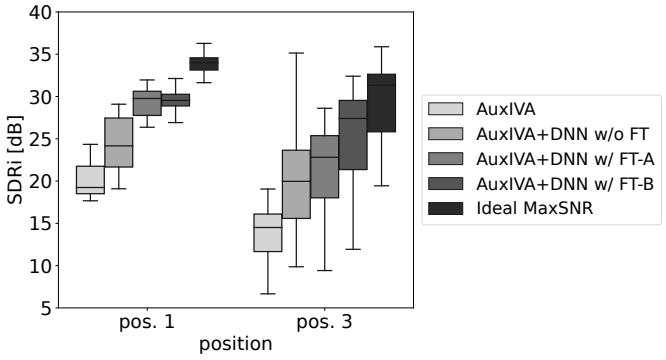

Fig. 6. SDRi for each method (FT-A: fine-tuning with mixture of speech and in-car noise, FT-B: fine-tuning with BSS outputs)

TABLE II
SPEECH RECOGNITION PERFORMANCE FOR EACH METHOD
(*DESINED USING CLEAN SOURCE IMAGES)

Methods		Accuracy		CER [%]	
		Pos. 1	Average	Pos. 1	Average
	Unprocessed	0.938	0.785	2.23	15.3
BSS	AuxIVA	0.917	0.819	4.47	**11.9**
BSS+DNN	AuxIVA+DNN w/o FT	0.875	0.597	15.6	42.1
	AuxIVA+DNN w/ FT-A	0.917	0.757	5.21	17.5
	AuxIVA+DNN w/ FT-B	**0.958**	**0.826**	**1.99**	12.7
	Ideal MaxSNR*	0.958	0.944	1.99	1.99

train significant features from large amounts of data would be one of the most compelling reasons. And comparing with the AuxIVA+DNN without FT result, SDRi improved about $5\,\mathrm{dB}$. From this result, we can confirm the improvement of noise reduction performance by using the BSS outputs as inputs of Conv-TasNet.

Table II shows the results of speech recognition performance under driving noise. The average accuracy in AuxIVA+DNN with FT-B result was 0.041 higher than that of the mixtures. Comparing AuxIVA+DNN with FT-B result with the AuxIVA result, the accuracy was equivalent to or higher score, and it was higher than that of the AuxIVA+DNN without FT result. In addition, the average CER in AuxIVA+DNN with FT-B result was equivalent to ideal MaxSNR result in pos. 1. Generally, the speech recognition performance with Conv-TasNet tends to worsen due to nonlinear filtering, but by using BSS outputs as input, the speech recognition accuracy was almost equivalent to or higher than that of BSS.

Fig. 7 shows the results of speech recognition accuracies for each SNR condition. The results using the proposed approach show the best performance in conditions above $-20\,\mathrm{dB}$. This results demonstrate that proposed approach works under high SNR. In contrast, below $-25\,\mathrm{dB}$, we obtained the best performance with AuxIVA, without Conv-TasNet. When the SNR is low, the noise remaining in the separated signal using AuxIVA is sometimes considered as speech, and the noise is sometimes enhanced as much as target speech. This may be the reason

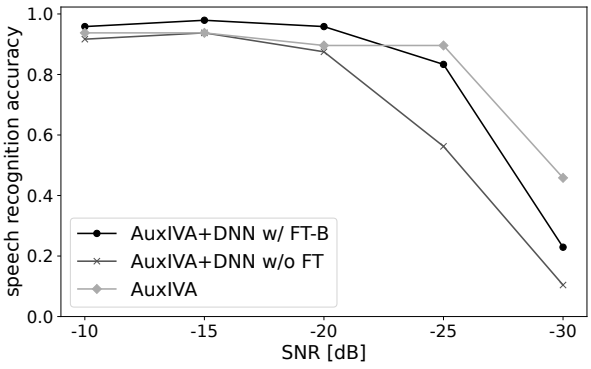

Fig. 7. Speech recognirion accuracy for each SNR (FT-B: fine-tuning with BSS outputs)

for the lower speech recognition rate.

Finally, Fig. 8 shows the spectrograms of the mixtures, the target speech signal, and the separated sound. It can be confirmed that low-frequency noise in the mixture is reduced or almost removed in AuxIVA+DNN with FT-B. However, as seen in before 20 sec of AuxIVA+DNN with FT-B result, the separation filter with Conv-TasNet may interpret the noise as speech. In addition, sparse parts of the actual speech are densely complemented by noise, which could not occur with a linear filter. We consider these results to be the reason for the low speech recognition performance.

V. Conclusion

In this study, we evaluated the performance of a speech enhancement system combining multi-channel BSS and single-channel DNN-based noise suppression in an in-car environment. Our approach demonstrated significant improvements, with an average SDRi of approximately 30 dB. Additionally, the speech recognition accuracy improved in both accuracy and CER criteria. These results indicate that the combination of BSS and DNN-based noise suppression can effectively enhance speech recognition performance in car environments. Future work will focus on leveraging more spatial information by incorporating multi-channel DNNs to further improve speech enhancement performance.

References

[1] M. Fukui, T. Watanabe, and M. Kanazawa, "Sound source separation for plural passenger speech recognition in smart mobility system," *IEEE Transactions on Consumer Electronics*, vol. 64, no. 3, pp. 399–405, 2018.

[2] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2002.

[3] C. Fox, G. Vitte, M. Charbit, J. Prado, R. Badeau, and B. David, "A subband hybrid beamforming for in-car speech enhancement," in *Proceedings of the 20th European Signal Processing Conference*, 2012, pp. 11–15.

[4] A. Chinatto, "In-car speech separation via MVDR," in *Proceedings of XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, Jan. 2017, pp. 122–126.

[5] S. Grimm and J. Freudenberger, "Wind noise reduction for a closely spaced microphone array in a car environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 7, Jul. 2018.

[6] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[7] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, Springer-Verlag, 2006, pp. 165–172.

[8] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proceedings of International Conference on Independent Component Analysis and Blind Signal Separation*, Springer-Verlag, 2006, pp. 601–608.

[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

[10] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, 2014, pp. 107–111.

[11] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2017, pp. 216–220.

[12] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 506–510.

[13] T. Yamada, A. Tawari, and M. M. Trivedi, "In-vehicle speaker recognition using independent vector analysis," in *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, 2012, pp. 1753–1758.

[14] K. Goto, L. Li, R. Takahashi, S. Makino, and T. Yamada, "Study on geometrically constrained IVA with auxiliary function approach and VCD for in-car communication," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 858–862.

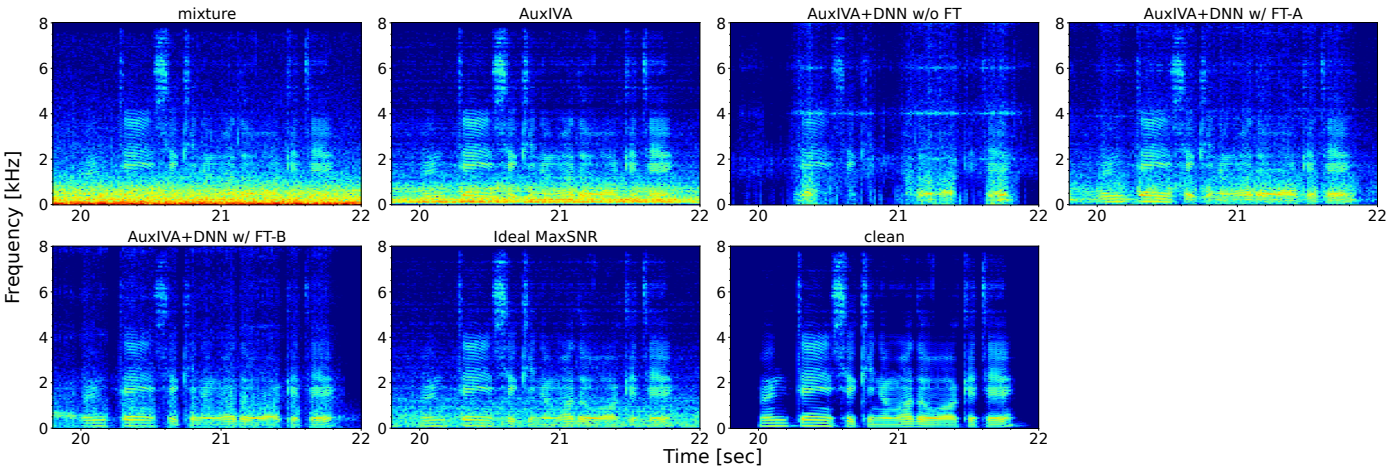

Fig. 8. Spectrograms of mixture, clean, and separated sound for pos. 1 (FT-A: fine-tuning with mixture of speech and in-car noise, FT-B: fine-tuning with BSS outputs)

[15] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3734–3738.

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

[17] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 286–290.

[18] 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.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[20] H. Wang, Z. Ye, and J. Chen, "A speech enhancement system for automotive speech recognition with a hybrid voice activity detection method," in *Proceedings of International Workshop on Acoustic Signal Enhancement*, 2018, pp. 1–9.

[21] V. Kothapally, Y. Xu, M. Yu, S.-X. Zhang, and D. Yu, "Deep neural mel-subband beamformer for in-car speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[22] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 749–753.

[23] Y. Kuriki, T. Nakashima, K. Yamaoka, *et al.*, "Efficient low-latency convolution with uniform filter partition and its evaluation on real-time blind source separation," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 765–769.

[24] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[25] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 56–60.

[26] M. Hu, *mhu-coder/ConvTasNet_Libri1Mix_enhsingle*, Dec. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4301955.

[27] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, *LibriMix: An open-source dataset for generalizable speech separation*, 2020. arXiv: 2005.11262.

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