Estimating Speaker's Seating Position from Monaural Speech in a Simulated Vehicle Interior Sound Field

Masataka Kaneko*†, Wen-Chin Huang*and Tomoki Toda*
* Nagoya University, Japan
† Forensic Science Lab., Aichi Pref. Police H.Q., Japan
E-mail: kaneko.masataka@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—We aim to establish a method to estimate the seating position of a speaker from monaural speech recorded by a driving recorder in a vehicle. There is a demand for identifying the seating position of a driver and passengers after the occurrence of a traffic accident. Since speech recorded with a microphone of the driving recorder is often available, we attempt to estimate the seating position using those speeches. If a person hears a sound directly with their ear in real space, it is possible to estimate the sound source direction to some extent from the characteristics of the pinna, even if it is one ear. However, the sound recorded with a mono microphone has a loss of information, making it difficult to estimate the direction. On the other hand, in a closed sound field, such as the inside of a vehicle, acoustic features of the monaural speech differ depending on the seating position from which the monaural speech reaches the microphone. Therefore, it is expected that the seating position estimation is still possible by modeling those subtle acoustic feature differences in the individual vehicles using machine learning techniques. In this paper, we investigate the possibility of estimating the seating position using the monaural speech by conducting experiments in a simulated sound field where the size of the vehicle interior and the arrangement of the microphone and seating positions are consistent during training and testing. The experimental results have demonstrated that 1) when the size of the vehicle interior and the arrangement of microphones and seats are almost the same during training and testing, the proposed method is effective for sound source localization using monaural audio, 2) if the microphone positions are different during training and testing, the performance degrades significantly, and 3) by using multiple utterances per person, the performance can be further improved.**

I. Introduction

When a traffic accident occurs, in order to prevent the perpetrator from hiding, it is necessary to know who was driving and where the passengers were sitting at the time of the accident objectively. There are several pieces of evidence to identify them, including DNA found in the vehicle, exterior security camera footage, and analysis of driving recorder footage. Of these pieces of evidence, few driving recorder footage shows the inside of the vehicle, and most only show the front and back of the vehicle. Therefore, it is usually not possible to determine who was sitting in which seat from the footage. Even if it is possible to identify who is speaking from sound such as a conversation, most driving recorders use mono microphones, making it impossible to employ sound source direction of arrival estimation methods using multi-channel microphones [1, 2], and therefore the localization of the speaker is indeed a difficult task.

Conventionally, it has been thought that sounds heard directly by one ear can be filtered by the pinna, making it possible to estimate the direction of the sound source to a certain extent in real space [3, 4]. On the other hand, when listening to sound signals recorded without passing through the pinna, it is thought to be difficult to estimate the direction of the sound source due to the lack of spectral cues [5].

However, in a closed sound field such as a vehicle, the reverberation and other features of the sound that reaches the driving recorder are significantly different for each seat from which the speaker speaks, and it is likely that different acoustic features are recorded as a result. Therefore, by using machine learning techniques, such as deep learning, which has been rapidly evolving in recent years, it is still possible to capture these acoustic feature differences.

In this paper, we investigate the possibility of estimating the seating position using the monaural speech by conducting experiments in a sound field where the size of the vehicle interior and the arrangement of the microphone and seating positions are consistent during training and testing. There are many types of vehicles, and the size of the interior and seat arrangement are different, so it is difficult to cover all of them in real-world experiments. In addition, to prove the feasibility of estimating the seating position from the monaural speech, it is necessary to attempt verification in an ideal sound field. Thus, we use a simulated vehicle interior sound field that allows for flexible design of the interior of the vehicle, and created data by playing sound from positions that represent each seat and recording monaural sound in a position that represents the driving recorder. We then use these data to classify the seat of the driver, passenger, rear of the driver, and rear of the passenger.

II. Related Works

Ando et al. [5] investigated clues for estimating monaural sound source direction by focusing on the Head Related Modulation Transfer Function (HRMTF) and analyzing the Monaural Modulation Spectrum (MMS) of sound signals heard by one ear. As a result, they found that it is difficult to estimate the direction of a sound source using only spectral cues, which are the features that humans mainly use to estimate the direction of a sound source. On the other hand, they also found that it is possible to use the difference in phase for each frequency

979-8-3315-7206-8/25/$31.00

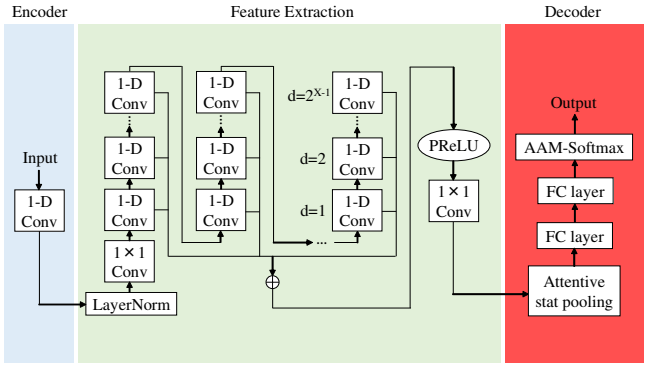

Fig. 1: A flowchart of proposed system. We used the encoder block and separation block of Conv-TasNet, added an Attentive statistics pooling layer and Fully Connected layers to the decoder block, and adopted AAM-Softmax loss for loss calculation

band, and that information on the modulation spectrum can be used to estimate the direction of a sound source.

Takashima et al. [6] proposed a sound source localization method using a monaural speech. According to their study, since it is difficult to estimate a position that has not been learned in advance, a single speaker localization method based on a regression model that predicts the position from the acoustic transfer function is discussed, and the effectiveness of the method was confirmed by sound source localization experiments performed in various indoor environments.

Apart from sound source localization, Luo et al. [7] proposed a sound source separation method from a monaural sound source. In this study, a machine learning model called Conv-TasNet is used to estimate the position of each speaker in time series information and achieve sound source separation by masking the original sound.

III. Proposed Seating Position Estimation with Monaural Speech

A. Framework

This study aims to classify the four seats mentioned above in the simulated vehicle interior sound field. To narrow the task, we define the environment in advance, such as the size and sound absorption of the vehicle inside, and the arrangement of seating positions. The experimental procedure is as follows: recording the sounds from each seat using a mono microphone, creating a classifier that estimates each seat position, training the classifier with the recorded sounds, and estimating each seat from monaural sounds using the trained classifier. In this framework, we will also focus on conditions such as the position of the driving recorder(mono microphone) and the change in sound source position within the same seat, and will clarify the effect that this has on the estimation performance.

B. Model Architecture

We detail the architecture of our model, which is based on Conv-TasNet [7]. Conv-TasNet is a commonly-used model structure for time-domain speech processing tasks. Some recent studies have considered processing in the time domain using deep learning, and their effectiveness has been shown in [7, 8]. This is because, as Luo et al. mentioned [7], most conventional speech and sound processing is done in the frequency domain using spectrograms, but the Short-Time Fourier Transform (STFT) is not necessarily optimal for sound source separation as information may be lost due to phase reconstruction errors. We assume that in a vehicle, since there is less internal reverberation, this loss of information may result in a large overall loss, and thus STFT may not be optimal for source localization as well.

Conv-TasNet is composed of three processing blocks: an encoder block, a separation block, and a decoder block. Sound source separation is achieved by inputting the product of the output of the encoder block and the output of the separation block into the decoder block. The input into the decoder block is shown in (1):

$$y = x \odot \mathcal{M}(x), \tag{1}$$

where x represents the output of the encoder block, $\mathcal{M}(\cdot)$ represents the mask function, y represents the input to the decoder block, and $\odot$ represents the Hadamard product. We utilize the encoder and separation blocks of Conv-TasNet to extend the system to classify the acoustic features of each seat from the original sound source.

Fig. 1 shows our proposed model architecture. The differences from previous studies are the configuration of the decoder block and the direct input of the mask representation $\mathcal{M}(x)$ into the decoder block. To deal with the specific task, this study is limited to classifying four seats in a vehicle, so unlike sound source separation, it is not necessary to retain time-series information, and we require the pooling to focus on important information in the time-series direction. Therefore, we introduce Attentive Statistics Pooling [9] and perform adaptive processing by calculating the weighted mean $\tilde{\mu}$ and weighted standard deviation $\tilde{\sigma}$ over a time sequence (from $t = 1$ to $t = T$) shown in (2) and (3):

$$\tilde{\mu} = \sum_{t}^{T} \alpha_t h_t \tag{2}$$

Table I: Detailed configuration of each block in the proposed model. T represents the number of frames in the time series direction.

Module	Detailed Configuration	Output Size	# params
Encoder	Conv1d	(1024, T)	614,400
Feature Extraction	Temporal ConvNet	(1024, T)	9,029,696
Decoder	Attentive Statistics Pooling	(2048, 1)	525,696
	FC layer-1	(512, 1)	1,049,088
	FC layer-2	(256, 1)	131,328
	AAM-Softmax	(4, 1)	1,028

Table II: Hyperparameters of the proposed network

Symbol	Description	Value
N	Number of filters in autoencoder	1,024
L	Length of the filters (in samples)	600
B	Number of channels in bottleneck and the residual paths' 1 × 1-conv blocks	256
Sc	Number of channels in skip-connection paths' 1 × 1-conv blocks	128
H	Number of channels in convolutional blocks	512
P	Kernel size in convolutional blocks	3
X	Number of convolutional blocks in each repeat	8
R	Number of repeats	4
m	Angular margin penalty of AAM-Softmax	0.01
s	Feature scale of AAM-Softmax	5

$$\tilde{\sigma} = \sqrt{\sum_{t}^{T} \alpha_t h_t \odot h_t - \tilde{\mu} \odot \tilde{\mu}}, \tag{3}$$

where h_t represents the feature for each frame, and α_t represents the attention weight for each frame.

Finally, we perform dimensionality reduction using a fully connected (FC) layer, and use the AAM-Softmax loss [10] to provide a margin between the four classes: driver's seat, passenger seat, rear of the driver's seat, and rear of the passenger seat. We detail the parameters of the blocks of the model in Table I.

IV. Experimental Setup

A. Data

Training data was created from *train-clean-100* and *train-clean-360* of LibriSpeech [11], and test data was created from *test-clean* of LibriSpeech. In addition, to confirm whether the accuracy of the estimate changes depending on the recording environment and language, the VCTK corpus [12], an English corpus as a different recording environment, and the JVS corpus [13], a Japanese corpus as a different language, were used for inference. When these data were input into the model, the model outputs the probability of the four seating positions.

We simulated four seating positions and created samples by playing speech from an arbitrary selected seat and recording it with a mono microphone. For training data, we created 8,000 samples with LibriSpeech for seating positions selected randomly. For testing data, we created 250 samples for each of the four seating positions, for a total of 3,000 samples with each corpus. All samples were resampled to 24 kHz and randomly trimmed to 3-seconds. In addition, to improve the robustness of the model, data augmentation was performed during training, including adding noise and adjusting the volume. For noise, Gaussian noise was added with an SNR of 40 to 60 dB, and the volume was adjusted from -6 to 0 dB, randomly.

B. Setting of the Simulated Vehicle Interior Sound Field

The design of the simulated vehicle interior sound field is shown in Fig. 2. The vehicle interior was simulated with a rectangular box using Pyroomacoustics [14]. Although in reality, the driving recorder is usually installed near the center of the windshield, the symmetry nature makes it impossible to distinguish left and right seating positions. Therefore, in our experiments, we placed the driving recorder on the left side of the vehicle. The sound source was placed in one of four positions, which were assumed to be the seat of driver, passenger, rear of the driver, and rear of the passenger, and the height was set to 800 mm for each position. After selecting the seating positions, the sound source was placed somewhere within a 200 mm square area of the seat.

C. Setting of Proposed Model

Table II shows the parameters of the separation block of Conv-TasNet and the parameters of AAM-Softmax in the model of the proposed system. A major difference from previous studies is that their approaches used a smaller receptive field (with *L*=16), whereas in this study we used a larger receptive field (with *L*=600, which is approximately 25ms), as we assume that a larger receptive field is needed to capture reverberation information.

V. Experimental Results

A. Main Results

The experimental results for each test data are shown in Table III. All accuracies were over 96 %, suggesting that with certain microphone and seat arrangements in the simulation, it may be possible to estimate position even with monaural speech. Furthermore, since the recognition results are comparable even when using Japanese, which is a different language family from English, it is confirmed that this method makes it possible to capture the acoustic features of speech regardless of language.

B. Robustness to the Change in Microphone Positions

To verify the robustness of the proposed system, we use a trained model with a microphone position of 1,000 mm to examine how much performance can be maintained when the microphone position is changed. The microphone positions were 760 mm near the center of the vehicle width (750 mm), 875 mm between the center and 1,000 mm, and 1,250 mm between 1,000 mm and the left side (1,500 mm). The location of each microphone is shown in Fig. 2 by the white microphones.

Table IV shows the accuracy of test data created for each position. At 875 mm, the accuracy of each seat was maintained at a high score, while the overall accuracy dropped significantly at 760 mm and 1,250 mm. However, if we focus on the seats on the left side, all of them maintained a relatively high score, which suggests that there is little change in reverberation and reflection due to the microphone position. Therefore, if the positional relationship between the microphone and the speaker is the same during learning and testing, a high accuracy will

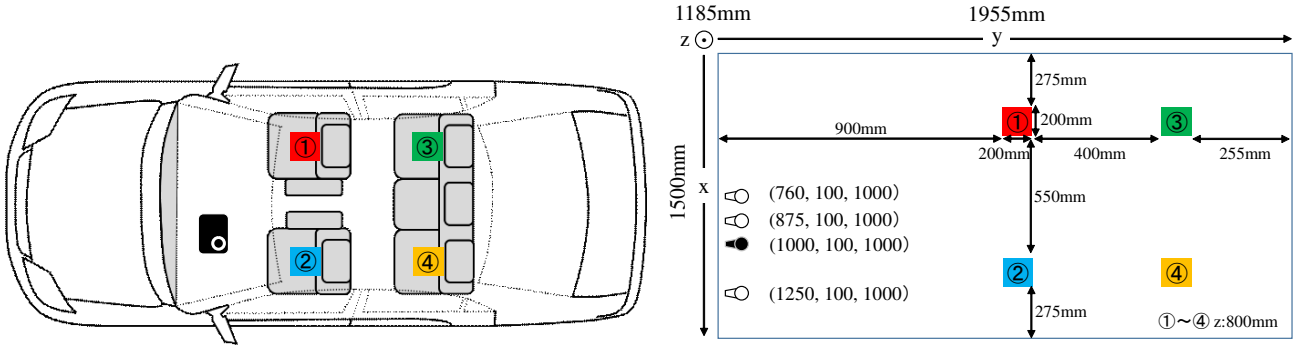

Fig. 2: Vehicle interior sound field environment settings. The figure on the left shows the assumed locations of the driving recorder and ① the driver's seat (Right, Front), ② the passenger seat (Left, Front), ③ the rear of the driver's seat (Right, Rear), and ④ the rear of the passenger seat (Left, Rear). The names of the seats and the location of the driving recorder vary depending on the position of the steering wheel, but in this paper we will set it up as follows assuming a right-hand drive vehicle. The figure on the right shows a simulated vehicle interior sound field designed based on the arrangement of the left. The microphone shown in black is used for training and testing, while the microphones shown in white are used for testing the experiment in Section V-B. In order to reduce the effect of reverberation from the front of the driving recorder, it was placed further forward than expected.

Table III: Results of source speaker localization using our proposed method. The numbers in the table show the predicted number of seats for each correct seat in the dataset. The rightmost column also shows the percentage of correct answers for each seat in each dataset.

Test Dataset	Correct Seat	# predicted seats				Accuracy
		①	②	③	④	
LibriSpeech	① R-Front	248	1	0	1	99.2%
	② L-Front	0	246	1	3	98.4%
	③ R-Rear	3	0	247	0	98.8%
	④ L-Rear	1	0	1	248	99.2%
VCTK	① R-Front	249	1	0	0	99.6%
	② L-Front	2	240	0	8	96.0%
	③ R-Rear	4	0	243	3	97.2%
	④ L-Rear	2	1	4	243	97.2%
JVS	① R-Front	244	1	1	4	97.6%
	② L-Front	0	248	0	2	99.2%
	③ R-Rear	1	0	247	2	98.8%
	④ L-Rear	0	0	0	250	100%

Table IV: Results of source speaker localization when changing the microphone location of test data

Test Dataset	Microphone Location	Correct Seat	# predicted seats				Accuracy
			①	②	③	④	
Libri-Speech	760 mm	①R-Front	43	156	32	19	17.2%
		②L-Front	34	167	31	18	66.8%
		③R-Rear	32	5	57	156	22.8%
		④L-Rear	38	8	37	167	66.8%
	875 mm	①R-Front	178	27	23	22	71.2%
		②L-Front	24	217	3	6	86.8%
		③R-Rear	34	6	155	55	62.0%
		④L-Rear	14	2	9	225	90.0%
	1,250 mm	①R-Front	79	75	81	15	31.6%
		②L-Font	8	229	2	11	91.6%
		③R-Rear	68	15	93	74	37.2%
		④L-Rear	5	11	54	180	72.0%

be shown, whereas if the positional relationship at the time of inference is different, performance is likely to drop significantly depending on the location. To address this, further learning is required for each position.

C. Influence of the Number of Enroll Utterances

In the above experiments, seating positions were estimated using a 3-second speech from each speaker. Although it would be ideal to identify seating positions using data of a shorter duration, conversations in real vehicles are rarely single 3-second speeches, and speakers generally speak intermittently multiple times while slightly changing the position of their face (and mouth). Therefore, it is expected that the accuracy of identification will be improved by using multiple utterances from the same speaker for one seating position, rather than being limited to identification at a single point.

The experimental procedure is as follows. First, three positions are randomly selected for each speaker within a 200 mm square area of one of the seats, and different speech data is played at each position to create three samples per speaker. The 256-dimensional output of the FC layer before being input to AAM-softmax is considered to be an embedding that represents the features of each seat, so we input three samples to the

Table V: Results of source speaker localization when using three utterances per speaker of test data

Test Dataset	Correct Seat	# predicted seats				Accuracy
		①	②	③	④	
LibriSpeech	① R-Front	250	0	0	0	100%
	② L-Front	0	250	0	0	100%
	③ R-Rear	0	0	250	0	100%
	④ L-Rear	0	0	0	250	100%
VCTK	① R-Front	250	0	0	0	100%
	② L-Front	1	248	0	1	99.2%
	③ R-Rear	0	0	249	1	99.6%
	④ L-Rear	0	0	0	250	100%
JVS	① R-Front	250	0	0	0	100%
	② L-Front	0	250	0	0	100%
	③ R-Rear	0	0	250	0	100%
	④ L-Rear	0	0	0	250	100%

proposed model individually and create a single normalized embedding by taking the mean value of the output embeddings. Finally, this value is input to AAM-softmax to output the estimation result for each seat.

Table V shows the results of calculating the accuracy using the data of three speeches per speaker for 1,000 speakers. The results improved in all datasets, confirming the effectiveness of this method.

VI. Conclusions

In this study, we estimated the speaker's seating position in a closed vehicle interior sound field as part of sound source localization using monaural speech. As a result, the proposed method showed high accuracy in estimating the seating position, and it was confirmed that the method is effective for sound source localization using monaural speech if the size of the vehicle interior and the arrangement of the microphone and seats were approximately the same during training and testing.

The experiment in this study is just a simulation, and does not accurately reflect the actual size of the vehicle interior or the reverberation, and there is a variety of noise inside and outside the vehicle. However, the structure inside the vehicle is not completely symmetrical, which causes different reverberations and echoes from each seat, and further improvement of the results can be expected by capturing these features, so the accuracy may be better than the simulation. In future research, we would like to make our system more robust by setting realistic conditions such as various interior size of the vehicle, inside music and outside noise, and ultimately attempt to identify speeches in an actual vehicle interior sound field.

Acknowledgment

This work is partly supported by JST AIP Acceleration Research JPMJCR25U5.

References

[1] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[2] B. D. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[3] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, pp. 1050–1063, 1997.

[4] K. Strelnikov, M. Rosito, and P. Barone, "Effect of audiovisual training on monaural spatial hearing in horizontal plane," *PLoS One*, vol. 6, no. 3, 2011.

[5] M. Ando, D. Morikawa, and M. Unoki, "Study on method of estimating direction of arrival using monaural modulation spectrum," *Journal of Signal Processing*, vol. 18, no. 4, pp. 197–200, 2014.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function," in *ICASSP 2013*, 2013, pp. 4295–4299.

[7] 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, 2019.

[8] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP 2018*, 2018, pp. 696–700.

[9] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[10] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, 2022.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP 2015*, 2015, pp. 5206–5210.

[12] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. DOI: 10.7488/ds/2645.6.

[13] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, *Jvs corpus: Free japanese multi-speaker voice corpus*, 2019. arXiv: 1908.06248.

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *ICASSP 2018*, 2018, pp. 351–355.