Multichannel environmental sound segmentation

with separately trained spectral and spatial features

Yui Sudo¹ · Katsutoshi Itoyama¹ · Kenji Nishida¹ · Kazuhiro Nakadai^{1,2}

Accepted: 3 March 2021 / Published online: 30 March 2021
© The Author(s) 2021

Abstract

This paper proposes a multichannel environmental sound segmentation method. Environmental sound segmentation is an integrated method to achieve sound source localization, sound source separation and classification, simultaneously. When multiple microphones are available, spatial features can be used to improve the localization and separation accuracy of sounds from different directions; however, conventional methods have three drawbacks: (a) Sound source localization and sound source separation methods using spatial features and classification using spectral features trained in the same neural network, may overfit to the relationship between the direction of arrival and the class of a sound, thereby reducing their reliability to deal with novel events. (b) Although permutation invariant training used in autonomous speech recognition could be extended, it is impractical for environmental sounds that include an unlimited number of sound sources. (c) Various features, such as complex values of short time Fourier transform and interchannel phase differences have been used as spatial features, but no study has compared them. This paper proposes a multichannel environmental sound segmentation method comprising two discrete blocks, a sound source localization and separation block and a sound source separation and classification block. By separating the blocks, overfitting to the relationship between the direction of arrival and the class is avoided. Simulation experiments using created datasets including 75-class environmental sounds showed the root mean squared error of the proposed method was lower than that of conventional methods.

Keywords Environmental sound segmentation · Sound source separation · Inter-channel phase difference · Semantic segmentation

1 Introduction

Various methods such as sound source localization (SSL), sound source separation (SSS), and classification have been proposed in acoustic signal processing, robot audition, and machine learning for use in real-world environments containing multiple overlapping sound events [1–3].

Conventional approaches use the cascade method, incorporating individual functions based on array signal processing techniques [4–6]. The main problem in this method is accumulation of errors generated at each function. Because each function is optimized independently

regardless of the overall task, each output might not be optimal for subsequent blocks.

Recently, deep learning-based end-to-end methods using a single-channel microphone have been proposed [7–9]. Environmental sound segmentation, which simultaneously performs SSS and classification, has been reported to achieve segmentation performance superior to the cascade method by avoiding accumulated errors [10, 11]. However, performance deteriorates with overlapping sounds from multiple sources, because a single-channel microphone obtains no spatial features.

Multichannel-based methods have been proposed for automatic speech recognition (ASR) [12–14]. Such methods integrate SSL, SSS, and ASR. In addition to the magnitude spectra, using interchannel phase difference (IPD) between microphones as spatial features have been reported to improve ASR performance in overlapping sounds containing multiple speakers. These methods use permutation

✉ Yui Sudo
sudo@ra.sc.e.titech.ac.jp

invariant training (PIT) [15, 16], where the model has multiple output layers corresponding to different speakers and is trained to cope with all possible combinations of speakers. While these studies assume mixtures of two or three speakers, it is impractical to extend them to many classes of sounds, such as environmental sounds.

A multichannel environmental sound segmentation method has been proposed [17]. This integrated method deals with SSL, SSS, and classification in the same neural network. Although the method implicitly intends that SSL, SSS and classification are trained simultaneously in a single network, loss function with respect to the direction of arrival (DOA) is not used in training. Thus, spatial features may not be used effectively. Deep learning-based methods for sound event localization and detection (SELD) have been proposed [18–21]. These methods simultaneously perform SSL and sound event detection (SED) of environmental sounds. Many SELD methods have two branches that perform DOA estimation and SED. These methods are trained using not

only loss function to SED outputs but also DOA outputs. However, the DOA and the class do not correlate unless the position and orientation of the microphones remains fixed. If a sufficient dataset is not available, the network overfits to the relationship between the DOA and the class.

Throughout the multichannel-based method, various features, such as complex values of short time fourier transform (STFT), IPD, sine and cosine of IPD, have been used as spatial features [14, 18, 20], but there are no studies comparing them.

This paper proposes a multichannel environmental sound segmentation method comprising two discrete blocks, a sound source localization and separation (SSLS) block and a sound source separation and classification (SSSC) block as shown in Fig. 1. This paper has the following contributions:

- It is not necessary to set the number of sound sources in advance, because sounds from all azimuth directions are separated simultaneously.

Fig. 1 Proposed framework of the environmental sound segmentation method. Spectral and spatial features were input to the sound source localization and separation block. Then each separated spectrogram was segmented into each class

- Because the SSLS block and the SSSC block are discrete there is no overfitting to the relationship between the DOA and the class.
- Comparison of various spatial features revealed the sine and cosine of IPDs to be optimum for sound source localization and separation.

2 Related work

This section describes multichannel-based approaches to sound source localization, sound source separation and classification.

2.1 Multichannel autonomous speech recognition

Multichannel-based methods have been proposed for ASR [12–14]. These methods perform SSL, SSS and ASR simultaneously. In addition to magnitude spectra, using IPD between microphones as spatial features have been reported to improve ASR performance in overlapping sounds containing multiple speakers. These methods use PIT, where the model has multiple output layers corresponding to different speakers and can account for all possible combinations of speakers. The computational complexity of PIT is of the order of $O(S!)$, where S is the number of speakers. While these studies assumed combinations of two or three speakers, it is impractical to extend them to many classes of sounds.

2.2 Multichannel environmental sound segmentation

A multichannel environmental sound segmentation method has been proposed [17]. This method uses magnitude spectra and the sine and cosine of the IPDs as input features to train SSL, SSS and classifications simultaneously in the same network. Although this method implicitly intends that SSL, SSS and classification are trained simultaneously in a single network, the loss function with respect to DOA is not used in training thus spatial features may not be used effectively. Normally, unless the position and orientation of the microphone is always fixed, DOA and the class do not correlate. If a sufficient dataset is unavailable, the network overfits to the relationship between DOA and the class.

A combination of SSL, and SSS for classification of bird songs has been proposed [24, 25]. The method comprises SSL, SSS, and classification blocks, and uses the SSL results as spatial cues for bird song classification. However, spatial cues are ineffective if the position and orientation of the microphone differs from that during training.

2.3 Sound event localization and detection methods for environmental sound

For detection and classification of acoustics of scenes and events (DCASE) [22], deep learning-based methods for SELD have been proposed [18–21]. These methods simultaneously perform SSL and SED of environmental sounds containing many classes. Many SELD methods have two branches that perform DOA estimation and SED, and calculate the loss to DOA and SED outputs, respectively. A simple SELD method optimizes both of these losses simultaneously, but typically DOA and class do not correlate unless the microphone position and orientation is always fixed. If a sufficient dataset is unavailable, the network overfits to the relationship between the DOA and the class. Therefore, many SELD methods have reported improved performance by training these two branches separately. However, these methods reduce the frequency dimension of the features by using frequency pooling [23] and cannot perform SSS. Additionally, various features, such as complex values of STFT coefficients, IPD, sine and cosine of IPD, have been used as spatial information, but no study has compared them.

2.4 Issues of related works

Conventional multichannel-based methods have drawbacks.

- For environmental sounds containing many classes, it is impossible to set a maximum number of sound sources in advance.
- If a sufficient dataset is unavailable, the network overfits to the relationship between the DOA and the class.
- Various features, such as complex values, IPD, sine and cosine of IPD, have been used as spatial information, but no study has compared them.

To address these issues, this paper proposes a multichannel environmental sound segmentation method that includes discrete SSLS and SSSC blocks. By separating the blocks, this method prevents overfitting to the relationship between DOA and the class. The SSLS block separates sound sources by each azimuth direction among the complex mixture of sounds, so that it is not necessary to set the number of sound sources in advance. Additionally, we compared multiple types of spatial features.

3 Proposed method

Figure 2 shows the overall structure of the proposed method which consists of four blocks: (a) feature extraction,

Fig. 2 Complete architecture of the proposed method comprising feature extraction, SSLS, SSSC and reconstruction. STFT was applied to the waveforms. The SSLS block predicted spectrograms of each direction from the input spectrograms. Since the SSLS block could not

separate sound sources that arrived from a close direction, the SSSC block performed not only classification but separated magnitude spectrograms for each class from the output of the SSLS block. Inverse STFT was applied to reconstruct the time domain signal

(b) sound source localization and separation (SSLS), (c) sound source separation and classification (SSSC), and (d) reconstruction. (a) STFT was applied to the mixed waveforms. STFT coefficients were decomposed into magnitude, sine and cosine of IPDs, and phase spectrograms. (b) The magnitude spectrograms and sine and cosine IPDs were input into the SSLS block. This block separated the magnitude spectrograms for each azimuth direction from the mixture. (c) The outputs of the SSLS block were input into the SSSC block. Because the SSLS block could not fully separate magnitude spectrograms for each azimuth direction from the mixture, it additionally separated magnitude spectrograms for each class from the output of the SSLS block. (d) The time domain signals were reconstructed using inverse STFT.

Normally, there is no correlation between DOA and the class unless the position and orientation of the microphone array is fixed. If a sufficient dataset is unavailable, conventional environmental sound segmentation method using the single network overfits to the relationship between the DOA and the class. In contrast, the proposed method prevents such overfitting by explicitly separating the SSLS block from the SSSC block. The method was trained in two stages. Initially, we trained with the SSLS block alone, then the SSSC block was trained using the output of the SSLS block as input with the weights of the SSLS block fixed. It is not necessary to set the number of sound sources in advance, because all sound sources in all azimuth directions are separated simultaneously. Although the SSLS block could not separate sound sources that arrived from a close direction, the SSSC block performed not only classification

but simultaneously separated the magnitude spectrograms for each class from the output of the SSLS block.

3.1 Feature extraction

We used the following spectral and spatial features proposed in [26, 27]. The input signals were multichannel time-series waveforms with a sampling rate of 16 kHz. STFT was applied using a window size of 512 samples and a hop length of 256 samples. A reference microphone, p , and other non-reference microphones, q , were selected. Magnitude spectrograms of the reference microphone were used as spectral features. The magnitude spectrograms were normalized to the range of [0, 1]. Meanwhile, sine and cosine of IPDs were used as spatial features as,

$$\sin IPD(t, f, p, q) = \sin(\theta_{t,f,p,q}), \quad (1)$$

$$\cos IPD(t, f, p, q) = \cos(\theta_{t,f,p,q}), \quad (2)$$

where $\theta_{t,f,p,q}$ is the IPD between the STFT coefficients $x_{t,f,p}$ and $x_{t,f,q}$, at time, t , and frequency, f , of the signals at reference microphone, p , and non-reference microphones, q .

3.2 Sound source localization and separation

Figure 3 shows the overview of the SSLS block. The SSLS block predicted $360 / n$ spectrograms of each azimuth direction at angular resolution, n . We used Deeplabv3+, which has been originally proposed for semantic segmentation of images [28], for our proposed method. Deeplabv3+ has been reported to improve the

Fig. 3 The overview of the SSLS block. The SSLS block predicted $360/n$ spectrograms of each azimuth direction at angular resolution, n

segmentation performance for environmental sounds with various event sizes [17].

Figure 4 shows the structure of Deeplabv3+ used for SSLS block. Similar to U-Net [7], which is often used as a conventional model, it has an encoder-decoder structure. The encoder block is a convolutional neural network that extracts high-level features. Xception [29] module was used

for feature extraction and it outputs a feature map that is $1/16$ of the original spectrogram size. The biggest difference with U-Net is a pyramid structure with dilated convolution [30] layers of different rates called Atrous Spatial Pyramid Pooling (ASPP) [31, 32] module performed after Xception module. Dilated convolution is a convolution applied to input with defined gaps as shown in Fig. 5. Dilation rate k

Fig. 4 Architecture of Deeplabv3+ which predicted spectrograms of each direction from the input spectrogram with angular resolution, n

Fig. 5 Dilated convolution. Blue area shows convolution filter. By increasing the dilation rate, the receptive field can be expanded without increasing the number of parameters

means skipping k pixels where $k = 1$ is a normal convolution. Because this technique can extract larger range of contexts without increasing the number of parameters, the model can be trained efficiently for environmental sounds with various event sizes. For example, the duration of a gunshot is short and thus the size of the spectrogram is small, whereas the duration of a musical instrument is long and thus the size of the spectrogram is large. The ASPP module can efficiently extract feature maps of environmental sounds of such various event sizes. In this paper, a kernel size of 3×3 , and dilation rates of 6, 12, and 18 were used in the ASPP module. The encoder features are first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level features from the Xception module that have the same resolution. After the concatenation, 3×3 convolutions are applied to refine the features followed by another bilinear upsampling by a factor of 4. The decoder block obtains a mask spectrogram the same size as the input spectrogram. By skipping the connection in the middle layer, a high-resolution function map is easily transmitted.

SSLS block predicted $360 / n$ spectrograms of each azimuth direction at angular resolution, n . While PIT requires the number of sound sources to be set in advance [15, 16], this method does not require the number of sound sources to be set in advance. The spectrograms in the direction where no sound sources exist were zero. If multiple sources exist in the same direction, the SSLS block cannot separate the sound sources. But the sources that could not be separated by the SSLS block were separated by the SSSC block described below. Note that the SSLS block predicts a spectrogram for each azimuth angle regardless of the class, so the network does not overfit to the relationship between the DOA and the class.

Equation (3) represents the loss function used in the training. X denotes the input spectrograms of the mixed signal and Y_{ssls} denotes the magnitude spectrograms of the target sounds. The mean squared error (MSE) between the output of the network and the target sounds was used for the

training:

$$L(X, Y_{ssls}) = \|f(X) \circ X_{mag} - Y_{ssls}\|_2, \quad (3)$$

where $f(X)$ is the mask spectrograms generated by the model and X_{mag} is the magnitude spectra of the reference microphone. The model was trained for 100 epochs at a learning rate of 0.001 using the ADAM optimizer [33].

3.3 Sound source separation and classification

Figure 6 shows the structure of the SSSC block. The output of SSLS block has n outputs for the n directions, but spectrograms separated by SSLS block were input into SSSC block one by one and segmented into each class. Although, the spectrograms of all directions could be input to the SSSC block at the same time as multichannel inputs, the relationship between the direction and class of the sound sources would be overfitted. The output in the direction where a sound source did not exist was not input to the SSSC block as,

$$X_{sssc} = \{Y_{sslsn} | \max\{Y_{sslsn}\} > 0.2\}, \quad (4)$$

where n represents the angle index of the SSLS output. In the normalized magnitude spectra, 0.2 corresponded to a maximum volume of approximately -96 dB. The spectrograms of the mixed sounds were concatenated and input to Deeplabv3+ because the spectrograms separated by SSLS block might be missing necessary information. Unlike Deeplabv3+ described in Section 3.2, the mixed spectrogram and separated spectrogram by SSLS block were input and the spectrograms for each class were output. Note that the outputs of the SSLS block were input to the SSSC block one by one, so that it has no spatial features. Since the SSSC block predicts the spectrogram for each class regardless of the DOA, the network does not overfit to the relationship between the DOA and the class. Equation (5) represents the loss function used in the training. X_{sssc} denotes the input spectrograms and Y

Fig. 6 Architecture of the SSSC block. Spectrograms separated by SSS block were input to this block one by one and segmented into each class

denotes the magnitude spectrograms of the target sound. The MSE loss between the output spectrograms and the targets was used in the training:

$$L(X_{SSC}, Y) = \|f(X_{SSC}) \circ X_{mag} - Y\|_2, \quad (5)$$

where $f(X_{SSC})$ is the mask spectrograms generated by the model and X_{mag} is the magnitude spectra of the reference microphone. The model was trained for 100 epochs at a learning rate of 0.001 using the ADAM optimizer [33].

4 Evaluation

This section describes the dataset, metrics, experimental conditions and experimental results.

4.1 Dataset

To train the network that deals with SSL, SSS, and classification, it was necessary to prepare pairs of mixed sounds and separated sound source signals of which DOA and class are known. Similar datasets, DCASE2020 Task3 dataset [34] and the Free Universal Sound Separation (FUSS) dataset [35] are used for SELD and sound separation in DCASE 2020 Task4, but, DCASE2020 Task3 dataset do not have the separated sound source signals and FUSS dataset do not have DOA labels, respectively. We created a dataset with dry sources of 75 classes using a method previously described [17]. Figure 7 shows the experimental settings for the simulations. Specifically, an 8-channel circular microphone array of radius 0.1 m was used. Dry sources were randomly selected from the corpus described in Table 1 and used as single point sound sources. Distance, d , between the center of the microphone array and a sound source was 1.0 m. The DOA of the sound source, θ , was randomly selected at 5 degree intervals. The impulse response from the

sound source to the microphone array was described as $\mathbf{h}(t, d, \theta) = [h_1(t, d, \theta), h_2(t, d, \theta), \dots, h_m(t, d, \theta)]$, where m indicates the number of microphones. The impulse response in free space was simulated. The impulse response was convolved to each dry source, $s_i(t)$, with a random time delay, t_r , as,

$$x_i(t) = \mathbf{h}(t, d, \theta) * s(t - t_r), \quad (6)$$

in which $*$ indicates a convolution operator. Each simulated sound was added, then sounds recorded in a restaurant and hall were mixed as background noise, $\mathbf{n}(t)$, as,

$$\mathbf{x}(t) = \mathbf{x}_1(t) + \mathbf{x}_2(t) \cdots + \mathbf{x}_I(t) + \mathbf{n}(t), \quad (7)$$

where I indicates the number of sound sources in a mixture. $I = 3$ was used in this experiment. The background noise, $\mathbf{n}(t)$, was added to all time frames of 8 channels to obtain an

Fig. 7 Experimental setting: 8-channel circular microphone array of radius 0.1 m was used. The distance between sound sources and the center of the microphone array was 1.0 m. The DOA of the sound sources were randomly selected at 5 degree intervals

Table 1 Corpus list for creating datasets

Database set	Contents	# of classes
ATR words	Male, female	2
RWCP	Bell, buzzer, clock, coin, phone, ping pong, ball, alarm, bottle, castanet, whistle, rap, maracas, air pump, book, handclaps, spray, tear	18
RWC-MusicDatabase	Cembalo, electric guitar, timpani, violin	4
Japan wild bird science	Birds	1
Bird research	Birds	1
Singing voices of grasshopper, and cricket	Insects	1
Sound database	Baby, bird, cat, footstep, frog, bathroom, clock, dog, golf, tennis, trampoline	11
Japanese cicada	Cicada	1
Freesound	Tearing, shattering, gunshot, fireworks, scissors, writing, computer keyboard, microwave oven, keys jangling, drawer opening or closing, knock, phone ringing, saxophone, oboe, flute, clarinet, acoustic guitar, tambourine, oboe, glockenspiel, snare drum, bass drum, hi-hat, double bass, electric piano, harmonica, trumpet, violin, cello, chime, cough, laughter, applause, finger snapping, gong, farting, burping, cowbell, bark, meow	43
DCASE 2016	Throat clearing, coughing,	
Task2 dataset	drawer, door slamming, keyboard, knocking, keys dropping, page turning, phone, speech, laughter	11
Total		75

average signal-to-noise ratio of about 15 dB. These sounds were assumed to be diffuse noise. Each sound was of 4.192 s duration. Figure 8 shows an example of the created dataset. The simulated signals were multichannel signals, but only one channel is shown in Fig. 8. The upper spectrogram was that of the dry source and was defined as the ground truth (Fig. 8a). The lower spectrogram was that of the mixed sound input into the neural network (Fig. 8b). The training and evaluation sets were 10,000 and 1,000 data points, respectively. The dry sources that was used to create the training set was not used to create the evaluation set.

4.2 Metrics

Environmental sound segmentation, in which SSL, SSS, and classification were performed simultaneously, was evaluated by root mean squared error (RMSE) as,

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=0}^N (Y_n - \hat{Y}_n)^2} \quad (8)$$

where Y and $\hat{Y}$ represent the magnitude spectra of the ground truth and separated sound, respectively. N represents the number of time-frequency bins. Silent sections were excluded from the evaluation.

4.3 Comparison between spatial features using SSSL block

To compare sound source separation performance by spatial features, preliminary experiments were conducted. In addition to sine and cosine of IPDs described in Section 3.1, we compared the complex STFT coefficients and IPD. The complex STFT coefficients were defined as,

$$STFT_{re}(t, f) = Re(x_{t,f}), \quad (9)$$

$$STFT_{im}(t, f) = Im(x_{t,f}), \quad (10)$$

where $x_{t,f}$, represents the STFT coefficients of each microphone at time frame, t , and frequency bin, f . IPDs were

Fig. 8 Dataset creation procedure. Each dry source was added after the impulse response was convolved with a random time delay. Mixed sounds were input data, and the sound source of each class before mixing was used as the ground truth. The dry sources that was used to create the training set was not used to create the evaluation set. **a** Ground truth **b** Mixed spectrogram

defined as,

$$IPD(t, f, p, q) = \theta_{t,f,p,q}, \quad (11)$$

where $\theta_{t,f,p,q}$ is the IPD between the STFT coefficients $x_{t,f,p}$, and $x_{t,f,q}$, at time, t , and frequency, f , of the signals at reference microphone p and non-reference microphones q . Note that STFT coefficients and IPDs were normalized to the range of $[-1, 1]$.

These spatial features were input into the SSLS block and evaluated against the output of the SSLS block using the RMSE. The angular resolution n was set to 45 degrees, and U-Net was used for the SSLS block in addition to Deeplabv3+. Here, sine and cosine of IPDs are continuous function, while $\theta_{t,f,p,q}$ is a periodic function which may make learning unstable.

Table 2 summarizes the results of the simulation experiments. Regardless of the model, the RMSE was smaller using sine and cosine of IPDs. Therefore, this results show that using sine and cosine of IPDs which are continuous functions as spatial features improves the sound source separation performance. We used sine and cosine of IPDs as spatial features in the following evaluation.

4.4 Analysis of the overfitting to the relationship between the DOA and the class

We conducted a preliminary experiments to analyze the overfitting to the relationship between the DOA and the class. Specifically, the RMSE of single-channel and multichannel input were compared. Unlike single-channel input, multichannel input has spatial features due to the DOA in addition to spectral features due to the class. If the network overfits to the relationship between the DOA and the class, the performance should be worse than that of the single-channel input. In addition to the 75-class dataset described in Section 4.1, a 3-class dataset was used. If the number of data is the same, performance degradation due to overfitting is less likely to occur in the 3-class dataset than in the 75-class dataset. In this experiment, we used a single network structure as shown in Fig. 10a to analyze the overfitting when SSL, SSS, and classification are implicitly performed in a single network. Deeplabv3+ and U-Net were used, and sine and cosine of IPDs validated in Section 4.3 were used as spatial features, respectively.

Table 3 shows the results of the experiments trained with 10,000 data points. Comparing the Deeplabv3+ and U-Net, the RMSE of Deeplabv3+ was smaller than that of U-Net in all experimental conditions. Thus, Deeplabv3+ has higher segmentation performance than U-Net. Subsequently, the results of single-channel and multichannel input are compared. The RMSE was smaller for the multichannel input in the 3-class dataset, whereas the RMSE was larger for the multichannel input in the 75-class dataset. This suggests that the network overfitted to the relationship between the DOA and the class because the number of data was not sufficient for the variance of the 75-class dataset, whereas the spatial features were efficiently exploited in the 3-class dataset.

Figure 9 shows the difference in RMSE between single-channel and multichannel input on the 3-class dataset.

Table 2 Results of the RMSE of spatial feature using SSLS block

Model	Real, imaginary	RMSE	
		IPD	sinIPD, cosIPD
U-Net	11.36	11.21	10.67
Deeplabv3+	10.25	10.53	10.17

Table 3 RMSE difference between single-channel input and multichannel input

# of classes	Data points	Model	Single-channel	RMSE Multichannel	Difference
3	10,000	U-Net	18.73	17.92	-0.45
		Deeplabv3+	17.10	16.59	-0.51
75	10,000	U-Net	25.87	25.97	+0.10
		Deeplabv3+	21.76	22.20	+0.44

Positive values mean that the RMSE of multichannel input was larger than that of single-channel input. When the number of data is not sufficient, the performance is deteriorated compared to the single-channel input because the network is likely to overfit to the relationship between the DOA and the class. Therefore, we found that the network overfits to the relationship between the DOA and the class if sufficient dataset is not available.

4.5 Comparison between various model structures

In environmental sound segmentation, it is necessary to optimize both spectral and spatial features, since it is necessary to perform SSL, SSS, and classification simultaneously. However, as previously mentioned, there is no correlation between the DOA and its class, which may lead to overfitting if the dataset is insufficient. Therefore, we compared four different structures, as shown in Fig. 10.

Figure 10a and b show end-to-end structures which simultaneously perform SSL, SSS, and classification in the same neural network. The difference between them is the loss functions used. The single-loss end-to-end structure shown in Fig. 10a uses only per-class losses. It does not use losses related to DOA, which may prevent the use of spatial

features. The multi-loss structure as shown in Fig. 10b is an extension of the single-loss end-to-end structure. This is a common structure of SELD methods which simultaneously performs DOA estimation and SED. This structure uses losses related to DOA and the class. However, learning two uncorrelated losses at simultaneously may lead to overfitting to the relationship between the DOA and the class. In contrast, Fig. 10c and d consist of two blocks. Both structures separate the SSLS block, which performs SSL and SSS simultaneously using spatial features, from the block which performs classification using spectral features. The structure shown in Fig. 10c consists of a cascade of the SSLS block and classification block using a convolutional neural network. Since the SSLS block could not separate sound sources that arrive from a close direction, which degrades performance due to errors caused by the SSLS block. The proposed structure uses an SSSC block instead of classification block. Although it is difficult for the SSLS block using spatial features to completely separate sound sources arriving from a close direction, the SSSC block improves separation performance based on the spectral features.

4.6 Results and discussion

Table 4 summarizes the results of the simulation experiments. The baselines were single-loss end-to-end methods using U-Net and Deeplabv3+ [17]. To verify the performance improvement due to spatial features, single-channel inputs were compared with multichannel inputs. Apart from input and output layers, each model had the same structure to compare the performance of the proposed method. The angular resolution, n , in the SSLS block was 45 degrees. In the baseline single-loss end-to-end method, the RMSE was not reduced by the use of multichannel inputs in either model. The single-loss end-to-end structure does not utilize the losses to DOA, and therefore the spatial features may not be well exploited. The multi-loss end-to-end structure, an extension of the single-loss end-to-end method, had a larger RMSE than the single-loss end-to-end structure. Since there is no correlation between DOA and class, optimizing them simultaneously may have overfitted the DOA and class relationship. In contrast, the results of the SSLS + Classification

Fig. 9 Relationship between the number of training data and RMSE difference between single-channel and multichannel input. Positive values mean that the RMSE of multichannel input was larger than that of single-channel input. When the number of data is not sufficient, the performance is deteriorated compared to single channel because the network is likely to overfit the relationship between DOA and class

Fig. 10 Comparison between various model structures. **a** End-to-end. **b** Multi-loss end-to-end. **c** SSLS + end-to-end. **d** SSLS + SSSC

structure showed a relatively small RMSE regardless of the model. This structure separates the SSLS and the classification blocks, which perform SSL and SSS based on spatial features, and thus does not overfit to the relationship between the DOA and the class. However, the SSLS block does not completely separate sounds arriving from a close direction, and thus the errors caused by the SSLS block accumulate. Therefore, no performance improvement was observed compared to the single loss end-to-end using Deeplabv3+. The proposed structure, in which the classification block of the SSLS + Classification structure was

replaced by the SSSC block, clearly had a smaller RMSE; the SSLS + Classification structure did not have the ability to correct the errors that occurred in the SSLS, but the SSSC block with inclusion of the separation feature reduced the propagation of errors in the SSLS block.

Figure 11 shows an example of the segmentation results. For clarity, magnitude spectrograms predicted for each class are displayed as a single spectrogram colored for each class. Overlaps of multiple sound sources are displayed in different colors. With the end-to-end approach, performance at the overlap area was degraded. Presumably, the spatial

Table 4 Results of environmental sound segmentation

System architecture	Model	RMSE
Single-loss end-to-end [17] (Fig. 10a)	U-Net (1ch)	25.87
	U-Net	25.97
	Deeplabv3+ (1ch)	21.76
	Deeplabv3+	22.20
Multi-loss end-to-end (Fig. 10b)	U-Net	29.36
	Deeplabv3+	27.06
SSLS + classification (Fig. 10c)	U-Net + CNN [36]	21.88
	Deeplabv3+ + CNN	21.88
SSLS + SSSC (Fig. 10d)	U-Net + U-Net	22.61
	Deeplabv3+ + Deeplabv3+	18.59

Fig. 11 Examples of segmentation results. **a** Ground truth. **b** Deeplabv3+. **c** Multi-loss. **d** SSLS+CNN. **e** SSLS+SSSC

features were not trained effectively because they did utilise losses related to DOA. In contrast, the multi-loss end-to-end structure clearly degraded the performance of the system. Since the DOA and its class were not correlated, it is

possible that they did not converge on the global minima. Thus, the method that explicitly separated the SSLS block from the class classification relatively reduced the RMSE. However, as shown in the top panel of Fig. 11d, since the

Fig. 12 An example of SSLS outputs and SSSC outputs. SSLS block predicted spectrograms of each direction from the input spectrograms. Although the SSLS block could not separate sound sources arrived from a close direction, the SSSC block separated the magnitude spectrograms for each class from the output of the SSLS block. **a** Mixture. **b** Output of SSLS. **c** Output of SSSC

two sound sources represented by orange and blue existed in close directions, the SSLS block could not separate them. The proposed method reduced the RMSE further by providing the SSSC block with a separation function to correct the errors that were not separated by the SSLS block, as shown in Fig. 11e. Figure 12 shows an example of SSLS block output and SSSC output. As shown in Fig. 12b, the SSLS block using spatial features could not separate the sound sources arriving from a close direction, whereas the SSSC block using spectral features was able to improve the segmentation performance. Thus, by explicitly training the SSLS and SSSC blocks in two stages, the spatial features were effectively exploited in the SSLS block and the SSSC block compensated for the drawback of the SSLS block using spectral features.

5 Conclusions

This paper proposed a multichannel environmental sound segmentation method that does not require the number of sound sources to be set in advance and does not overfit to the relationship between the DOA and the class. This method prevents overfitting to the relationship between DOA and the class by explicitly separating the SSLS block from the SSSC block. Simulation experiments using the created datasets including 75-class environmental sounds showed that the proposed method improved the segmentation performance compared to conventional methods. When the sound sources came from a close direction, the SSSC block was able to correct the sound sources that could not be separated by the SSLS block.

Acknowledgements This work was supported by JSPS KAKENHI Grant No. JP19K12017, JP19KK0260 and JP20H00475.

References

- Nakadai K, Okuno HG, Kitano H (2002) Real-time sound source localization and separation for robot audition. In: Proceedings IEEE international conference on spoken language processing, pp 193–196
- Nakadai K, Matsuura D, Okuno HG, Kitano H (2003) Applying scattering theory to robot audition system: Robust sound source localization and extraction. In: Proceedings IEEE/RSJ international conference on intelligent robots and systems, pp 1147–1152
- Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD (2015) Detection and classification of acoustic scenes and events. *IEEE Trans Multimed Speech Signal Process* 17(10):1733–1746
- Gabriel D, Kojima R, Hoshiba K, Itoyama K, Nishida K, Nakadai K (2019) 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. *J Adv Robot* 33(7-8):403–414
- Nakadai K, Ince G, Nakamura K, Nakajima H (2012) Robot audition for dynamic environments. In: IEEE International conference on signal processing, communication and computing, pp 125–130
- Nakamura K, Nakadai K, Okuno HG (2013) A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition. *J Adv Robot* 27(12):933–945
- Jansson A, Humphery E, Montecchio N, Bittner R, Kumar A, Weyde T (2017) Singing voice separation with deep U-Net convolutional networks. In: Proceedings of the international society for music information retrieval conference, pp 323–332
- Tan1 K, Wang D (2018) A convolutional recurrent neural network for real-time speech enhancement. In: Proceedings of Interspeech, pp 3229–3233
- Kavalerov I, Wisdom S, Erdogan H, Patton B, Wilson K, Roux JL, Hershey JR (2019) Universal sound separation. In: Proceedings IEEE workshop on applications of signal processing to audio and acoustics
- Sudo Y, Itoyama K, Nishida K, Nakadai K (2019) Environmental sound segmentation utilizing Mask U-Net. In: IEEE/RSJ International conference on intelligent robots and systems, pp 5340–5345
- Sudo Y, Itoyama K, Nishida K, Nakadai K (2020) Sound event aware environmental sound segmentation with Mask U-Net. *J Adv Robot* 34(20):1280–1290
- Yoshioka T, Erdogan H, Chen Z, Xiao X, Allea F (2018) Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. In: Proc ISCA Interspeech
- Yoshioka T, Erdogan H, Chen Z, Allea F (2018) Multi-microphone neural speech separation for far-field multi-talker speech recognition. In: Proc IEEE int. conf. acoust., speech signal process, pp 5739–5743
- Chang X, Zhang W, Qian Y, Roux JL, Watanabe S (2019) Mimospeech: End-to-end multi-channel multi-speaker speech recognition. In: IEEE automatic speech recognition and understanding workshop (ASRU), pp 237–244
- Yu D, Kolbak M, Tan Z-H, Jensen J (2017) Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Proc. ICASSP, IEEE, pp 241–245
- Yu D, Chang X, Qian Y (2017) Recognizing multi-talker speech with permutation invariant training. *Proc Interspeech 2017*:2456–2460
- Sudo Y, Itoyama K, Nishida K, Nakadai K (2020) Multi-channel environmental sound segmentation. In: IEEE/SICE International symposium on system integration, pp 820–825
- Sudo Y, Itoyama K, Nishida K, Nakadai K (2019) Improvement of DOA estimation by using quaternion output in sound event localization and detection, Workshop on Detection and Classification of Acoustic Scenes and Events 244–248
- Kapka S, Lewandowski M (2019) Sound source detection, localization and classification using consecutive ensemble of CRNN models. In: Technical report in detection and classification of acoustic scenes and events 2019 (DCASE) Challenge
- Cao Y, Iqbal T, Kong Q, Galindo MB, Wang W, Plumbley MD (2019) Two-stage sound event localization and detection using intensity vector and generalized crosscorrelation. In: Technical

report in detection and classification of acoustic scenes and events 2019 (DCASE) Challenge

21. Nguyen TNT, Jones DL, Gan W (2020) A sequence matching network for polyphonic sound event localization and detection. In: IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE
22. Detection and classification of acoustic scenes and events [Internet]. [cited2020Oct7]. Available at: <http://dcase.community/challenge2020/index>
23. Cakir E, Parascandolo G, Heittola T, Huttunen H, Virtanen T (2017) Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Trans Audio Speech Language Process* 25(6):1291–1303
24. Kojima R, Sugiyama O, Suzuki R, Nakadai K, Taylor CE (2016) Semi-automatic bird song analysis by spatial-cue-based integration of sound source detection, localization, separation, and identification. In: *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems*, pp 1287–1292
25. Kojima R, Sugiyama O, Hoshiba K, Nakadai K, Suzuki R, Taylor CE (2017) Bird song scene analysis using a spatial-cue-based probabilistic model. *J Robot Mechatron* 29(1):236–246
26. Wang Z-Q, Le Roux J, Hershey JR (2018) Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In: *Proceedings of IEEE international conference on acoustics speech and signal processing*
27. Wang W-Q, Wang D (2018) Integrating spectral and spatial features for Multi-channel speaker separation. In: *Proc Interspeech*, pp 2718–2722
28. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *The European conference on computer vision (ECCV)*, pp 801–818
29. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*, pp 1251–1258
30. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions, arXiv:1511.07122
31. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell* 40(4):834–848
32. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation, arXiv:1706.05587
33. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: *3rd International conference for learning representations San Diego*
34. Politis A, Adavanne S, Virtanen T (2020) A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection, workshop on detection and classification of acoustic scenes and events
35. Fonseca E, Favory X, Pons J, Font F, Serra X (2020) FSD50k: an open dataset of human-labeled sound events, in arXiv
36. Sudo Y, Itoyama K, Nishida K, Nakadai K (2021) Multi-Channel Environmental Sound Segmentation Utilizing Sound Source Localization and Separation U-Net. In: *IEEE/SICE International symposium on system integration*, pp 382–387

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research Institute Japan, Co., Ltd. His research interests include AI, computational auditory scene analysis, and robot audition.

scene analysis based on statistical signal processing and machine learning techniques. He is a member of the IPSJ, ASJ, and IEEE.

he was assigned to Neuroscience Research Institute as a senior research scientist from 2001 to 2018. Currently, he is a specially-appointed associate professor for Tokyo Institute of Technology from April 2018.

Yui Sudo received a BE degree in system design engineering and a ME degree in integrated design engineering from Keio University in 2009 and 2011, and received a Ph.D. in Engineering from Tokyo Institute of Technology in 2021. He worked with Honda Engineering for 6 years as a staff engineer from 2012 to 2018. After that, he worked for Honda R&D as a staff engineer from 2019 to 2020. Currently he is a senior engineer for Honda

Katsutoshi Itoyama received the B.E. degree in engineering in 2006, the M.S. degree in informatics in 2008, and the Ph.D. degree in informatics in 2011 all from Kyoto University. He worked with Kyoto University from 2011 to 2018. After that, he is currently an Specially Appointed Associate Professor (Lecturer), Tokyo Institute of Technology. His research interests include musical information processing, robot audition, and computational auditory

Kenji Nishida received a B.E. in electrical engineering in 1981 from Keio University, and received a Ph.D. in Engineering in 2009 from University of Tsukuba. He worked for Electrotechnical laboratory (ETL) 20 years as a research scientist from 1981 to 2001 (1985–1989, he was sent to Institute for New-Generation Computing Technology (ICOT)). As an organization of National Institute of Advanced Industrial Science and Technology (AIST),

Kazuhiro Nakadai received a B.E. in electrical engineering in 1993, an M.E. in information engineering in 1995, and a Ph.D. in electrical engineering in 2003 from the University of Tokyo. He worked with Nippon Telegraph and Telephone for four years as a system engineer from 1995 to 1999. After that, he was worked on the Kitano Symbiotic Systems Project, ERATO, JST as a researcher from 1999 to 2003. Currently he is a principal scientist for Honda

Research Institute Japan, Co., Ltd. He has had a concurrent position at Tokyo Institute of Technology, as a visiting associate professor from 2006 to 2010, a visiting professor from 2011 to 2017, and a specially-appointed professor from July, 2017. He also had a concurrent position as a guest professor at Waseda University from 2011 to 2018. His research interests include AI, robotics, signal processing, computational auditory scene analysis, multi-modal integration and robot audition. He has been an executive board member for JSAI from 2015 to 2016, and for RSJ from 2017 to 2018. He is also a member of JSAI, RSJ, IPSJ, ASJ, HIS, ISCA, ACM and IEEE.

Affiliations

Yui Sudo¹ · Katsutoshi Itoyama¹ · Kenji Nishida¹ · Kazuhiro Nakadai^{1,2}

Katsutoshi Itoyama
itoyama@ra.sc.e.titech.ac.jp

Kenji Nishida
nishida@ra.sc.e.titech.ac.jp

Kazuhiro Nakadai
nakadai@jp.honda-ri.com

¹ Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552, Japan

² Honda Research Institute Japan Co., Ltd., 8-1, Honcho, Wako, Saitama, 351-0188, Japan