A JOINT SEPARATION-CLASSIFICATION MODEL FOR SOUND EVENT DETECTION OF WEAKLY LABELLED DATA

Qiuqiang Kong, Yong Xu*, Wenwu Wang, Mark D. Plumbley*

Center for Vision, Speech and Signal Processing, University of Surrey, UK
{q.kong, yong.xu, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

Source separation (SS) aims to separate individual sources from an audio recording. Sound event detection (SED) aims to detect sound events from an audio recording. We propose a joint separation-classification (JSC) model trained only on weakly labelled audio data, that is, only the tags of an audio recording are known but the time of the events are unknown. First, we propose a separation mapping from the time-frequency (T-F) representation of an audio to the T-F segmentation masks of the audio events. Second, a classification mapping is built from each T-F segmentation mask to the presence probability of each audio event. In the source separation stage, sources of audio events and time of sound events can be obtained from the T-F segmentation masks. The proposed method achieves an equal error rate (EER) of 0.14 in SED, outperforming deep neural network baseline of 0.29. Source separation SDR of 8.08 dB is obtained by using global weighted rank pooling (GWRP) as probability mapping, outperforming the global max pooling (GMP) based probability mapping giving SDR at 0.03 dB. Source code of our work is published.

Index Terms— Sound event detection, source separation, weakly labelled data.

1. INTRODUCTION

Sound event detection (SED) aims to detect specific audio events from an audio recording. SED has many applications in our daily life, for example, detecting a baby cry at home, detecting the tapping sound in an office and monitoring the fire alarm or gunshot [1] in a public area. On the other hand, source separation (SS) aims to separate individual sources from a recording [2] and can be used in SED [3].

Many current SED models are trained using supervised learning methods [4, 5, 6]. These supervised learning methods need labelled onset and offset time of the audio events, which we call *strongly labelled data* (SLD). Labelling the SLD is time consuming and difficult to scale [4]. In addition, the onset and offset time of some audio events are ambiguous due to the fade in and fade out effect, for example, the approaching and moving away of a car. In contrast to the SLD, many audio datasets contain only the tags, that is, the presence or absence of audio events in an audio recordings. This is referred to as *weakly labelled data* (WLD) [7]. Many audio tagging datasets are weakly labelled [8, 9, 10] and are often larger than strongly labelled SED datasets [4, 9]. To utilize the WLD, some methods including joint detection-classification (JDC) model [11], attention and localization model [12] and multi-instance learning methods [7] have been used. Source separation can be used for sound event detection [3]. Unsupervised source separation methods including computation audio scene analysis (CASA) uses time-frequency (T-F) masking to emulate how human performs source separation [13]. Supervised source separation methods need clean sources for training [2] and have achieved state-of-the-art performance.

In this paper, a *joint separation-classification* (JSC) model is proposed to train the source separation model on the WLD. The proposed framework consists of two parts. The first part is a *separation mapping* from the T-F representation of an audio signal to the T-F segmentation masks of each audio event. The second part is a *classification mapping* from each segmentation mask to its corresponding audio tag. In the source separation stage, separated sources of different classes can be obtained from the T-F segmentation masks.

The remainder of the paper is organized as follows: Section 2 discusses convolutional neural network. Section 3 proposes the source separation framework. Section 4 shows experimental results. Section 5 concludes and proposes the future work.

2. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural networks (CNNs) are used initially in image classification [14] and recently have been very successful in audio processing, including speech recognition and audio classification [15]. In audio classification, the waveform is transformed to T-F representations which are treated as an image and fed as input to a CNN [15]. A CNN consists of several convolutional layers and each contains several trainable filters trained to learn some local patterns in the feature map. Downsampling usually follows some convolutional lay-

* The authors contribute equally to this work.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

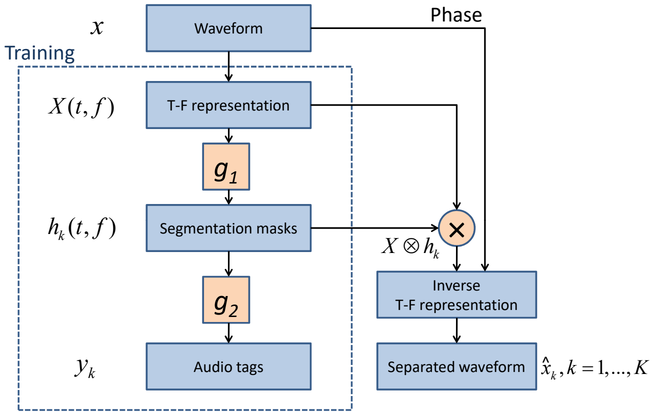

Fig. 1. Framework of the joint separation-classification model.

ers to reduce the size of the feature maps. Finally a global max pooling on each feature map [15] is usually used to select the most prominent T-F unit in each feature map followed by a fully connected neural network for classification [15].

3. PROPOSED JOINT SEPARATION-CLASSIFICATION MODEL

In this section, a joint separation-classification (JSC) model trained on WLD is proposed. This idea is related to the object localization from weakly labelled images [16, 17], where only the labels of an image are known, but the location of the objects are unknown. In [16] a class activation mapping (CAM) is applied to highlight the class-specific discriminative regions to localize objects from weakly labelled data.

Similar to the weakly labelled image data [16, 17], many audio datasets [4] only contain the tags of an audio recording, but the happening time of the events are unknown. The proposed separation-classification model is shown in Fig. 1. The input audio waveform x is transformed to a time-frequency (T-F) representation $X(t, f)$ such as a spectrogram or log Mel spectrogram. To simplify the notation we abbreviate $X(t, f)$ as X. The first part of the model is a *separation mapping* $g_1 : X \mapsto \mathbf{h}$ from the input T-F representation X to the T-F segmentation masks $\mathbf{h} = [h_1, ..., h_K]$, where K is the number of audio tags and h_k is the T-F segmentation mask of the k-th audio tag. The values on each segmentation mask are between 0 and 1 for source separation. The mapping g_1 can be parametrized by trainable parameters. The second part of the model is a *classification mapping* $g_2 : h_k \mapsto y_k, k = 1, ..., K$ from each segmentation mask to its corresponding audio tag, where $y_k \in [0, 1]$ represents the presence probability of the k-th event in an audio recording. A compound model $g_2 \circ g_1$ is a mapping from the input T-F representation X to the audio tags $y_k, k = 1, ..., K$. In the training phase, the model can be trained end-to-end from X to $y_k, k = 1, ..., K$. In the separation stage, the T-F representation of an input waveform is passed through the mapping g_1 to get the segmentation masks. Then the input T-F representation is multiplied by each segmentation mask to obtain the separated T-F representation of each event with corresponding audio tag. Then an inverse T-F transform is applied on each separated T-F representation of each audio tag using the phase of the original waveform to obtain its separated waveform of each audio tag (Fig. 1). Finally, SED result of each audio event can be obtained from its corresponding segmentation masks.

3.1. Separation mapping

We apply log Mel spectrogram as input T-F representation, which is a good representation for audio tagging [15]. We apply a CNN to model the separation mapping g_1. The CNN modeled JSC model is shown in Fig 2. We remove all the downsampling layers to keep the resolution of each T-F segmentation mask the same as the input T-F representation. This high resolution T-F segmentation mask is useful for source separation. The number of feature maps in the last convolutaional layer is the same as the number of audio events to separate followed. Then a sigmoid nonlinearity is applied on the feature maps to obtain the segmentation to ensure the values on segmentation masks are between 0 and 1. This segmentation mask of this T-F representation is similar to the class activation mapping (CAM) in weak image localization [16].

3.2. Classification mapping

The classification mapping g_2 maps each segmentation mask to the presence probability of its corresponding tag. Classification mapping can be modeled by, for example, global max pooling (GMP) [16], global average pooling (GAP) [18, 16] or global weighted rank pooling (GWRP) [19].

3.2.1. Global max pooling

Global max pooling (GMP) [15] is defined as follows:

$$g(h_c) = \max_{t,f} h_c(t, f) \qquad (1)$$

where h_c represents the c-th segmentation mask and t, f are indexes of time and frequency bin. GMP returns the highest value on each feature map. GMP performs well in classification but tends to underestimate the T-F units of events in each segmentation mask [16] because only the T-F unit with the highest value is passed to the next layer (Fig. 3).

3.2.2. Global average pooling

Global average pooling (GAP) [18] is defined as:

$$g(h_c) = \frac{1}{M} \sum_{t,f} h_c(t, f) \qquad (2)$$

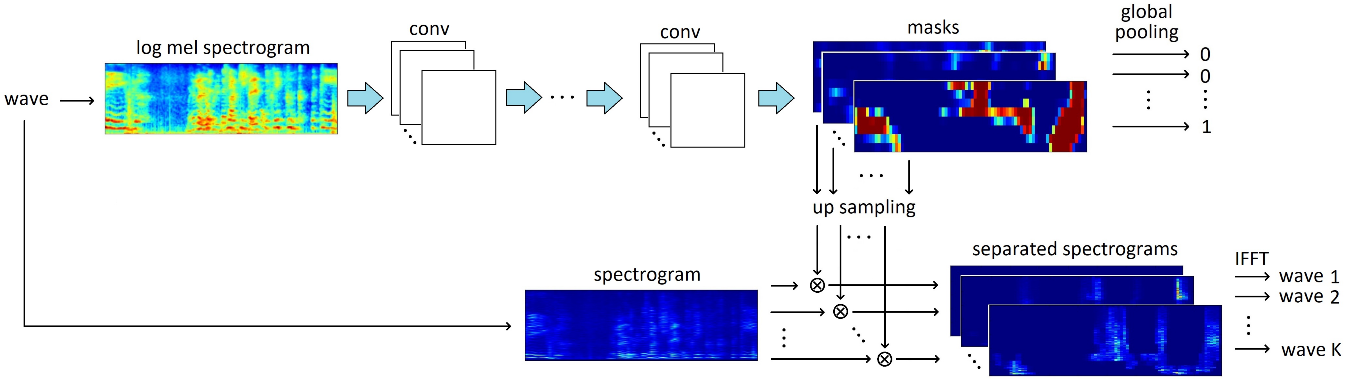

Fig. 2. Convolutional neural network (CNN) based weak source separation. Log Mel spectrogram is used as T-F representation. Separation mapping is modeled by a CNN. Classification mapping is applied on each T-F segmentation mask to obtain the prediction of audio tags. In separation stage, the separated waveforms are obtained from the segmentation masks.

where M is the number of time frames multiplied number of frequency bins. In contrast to GMP, GAP averages all the values of T-F units on a segmentation mask, which tends to overestimate the events in a segmentation mask [19] (Fig. 3).

3.2.3. *Global weighted rank pooling*

Global weighted rank pooling (GWRP) is proposed in [19] and is a generalization of GMP and GAP. Define $I^c = \{i_1, ...i_n\}$ as an index set in descending order of the values on feature map h_c, i.e. $(h_c)_{i_1} \geq (h_c)_{i_2} \geq ... \geq (h_c)_{i_n}$. Then GWRP is defined as

$$g(h_c) = \frac{1}{Z(d_c)} \sum_{j=1}^{N} (d_c)^{j-1} (h_c)_{i_j} \tag{3}$$

where $0 \leq d_c \leq 1$ is a hyper parameter and $N = TF$ is the number of T-F units in a segmentation mask and $Z(d_c) = \sum_{j=1}^{N} (d_c)^{j-1}$ is a normalization term. When $d_c = 0$ and $d_c = 1$, GWRP simplifies to GMP and GAP, respectively.

3.3. Sound event detection

The segmentation masks obtained from the JSC model contains the presence of the audio events in a T-F representation (Fig 3). Hence we achieved sound event segmentation in T-F domain. In this paper we simply average out the frequency axis to obtain the SED in the time axis.

4. EXPERIMENTS

In this section we apply the proposed JSC model on the modified detection of rare audio sound events dataset from Task 2 of DCASE 2017 challenge [10]. This dataset consists of rare events including "babycry", "gunshot" and "glassbreak". The background sounds come from the acoustic scene dataset from Task 1 of the DCASE 2017 data challenge [10]. To investigate WLD, we extract several rare audio events from the dataset and mix the rare audio events with 4 second clips from the acoustic scene dataset. Altogether 1008 clips are created for training, with 1/3 are single labelled and 2/3 are multilabelled. Only the presence or absence of the audio events in an audio clip is known. The audio mixtures are converted to monaural, and the sampling rate is 16 kHz. A log Mel spectrograms with 64 frequency bins are used as the T-F representation. In the Fourier transform a Hamming window with size of 1024 and overlap of 280 samples is used to ensure that there are 128 frames in each 4 seconds clip. We apply a Visual Geometry Group [14] like CNN consists of 8 convolutional layers. Each layer consists of 64 feature maps followed by batch normalization (BN) [20] and ReLU nonlinearity. Dropout rate of 0.3 is applied to regularize overfitting. The value of d_c in GWRP is set as 0.999. These hyper-parameters are chosen empirically, but they do not affect the result much.

The learned segmentation masks using different classification mappings are visualized in Fig 3. The first column shows the log Mel spectrogram of a "babycry", a "glassbreak" and a "gunshot". The second column shows the ideal binary mask (IBM) [21] of the audio events. Column 3 to 5 shows the segmentation masks learned using GMP, GAP and GWRP as classification mapping, respectively. It can be observed that GMP tends to underestimate the presence of the audio events in the T-F segmentation mask. GAP and GWRP performs better in learning the T-F segmentation mask on this dataset.

Table 1 shows the separation results of different audio tags evaluated on SDR, SIR and SAR [22]. The results of IBM [21] and without separation are listed as baselines. GWRP performs better in terms of SDR and SAR in babycry, glassbreak, gunshot and background than without separation, GMP and GAP. Table 1 shows that source separation using the proposed JSC model outperforms significantly the baseline without separation. Table 1 also shows how far JSC is from the

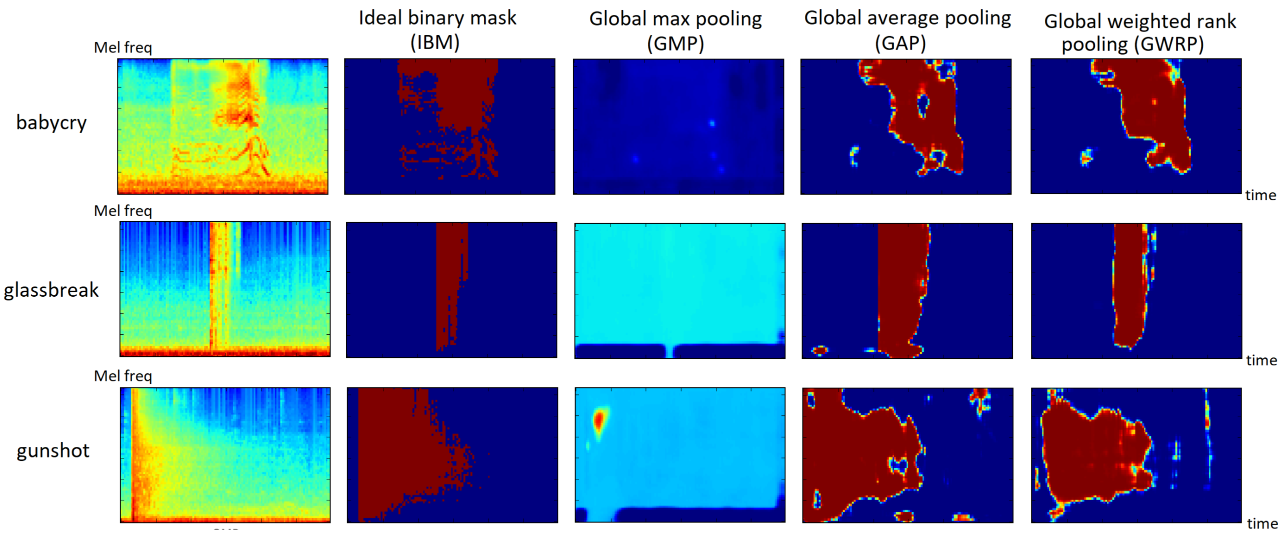

Fig. 3. Visualization of the segmentation masks using different global pooling strategy. The first column shows the log Mel spectrogram of babycry, glassbreak and gunshot sound in noisy background. The second column shows the ideal binary mask. The third to the fifth column shows the T-F segmentation masks learned using global max pooling (GMP), global average pooling (GAP) and global weighted rank pooling (GWRP), respectively.

Table 1. Separation results of mixed rare events with background sound using different methods.

	Babycry			Glassbreak			Gunshot			Avg.		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
w/o separation	-3.66	-3.66	inf	-7.52	-7.52	inf	-6.48	-6.48	inf	-5.89	-5.89	inf
IBM	20.14	34.73	20.32	18.62	37.35	18.70	15.24	33.04	15.35	18.00	35.04	18.12
Proposed GMP	2.99	15.43	5.85	-1.79	0.79	10.05	-1.11	1.66	**9.84**	0.03	5.96	8.58
Proposed GAP	9.58	**22.61**	10.21	6.35	17.81	8.49	**2.25**	13.05	4.73	6.06	17.82	7.81
Proposed GWRP	**13.36**	24.61	**14.20**	**12.29**	**28.06**	**12.86**	-1.41	**13.93**	-0.28	**8.08**	**22.20**	**8.93**

Table 2. Frame wise equal error rate (EER) of mixed rare events with background sound using different method.

	babycry	glassbreak	gunshot	avg.
baseline DNN	0.27	0.26	0.34	0.29
weak GMP	0.27	0.30	0.32	0.30
weak GAP	**0.11**	0.12	**0.19**	**0.14**
weak GWRP	**0.11**	**0.10**	0.20	**0.14**

IBM in source separation.

Table 2 shows the frame wise sound event detection equal error rate (ERR) using different global pooling strategies. GAP and GWRP outperforms the baselines DNN and GMP. The results are correspondent to the visualization of segmentation masks in Fig 3. and Table 1. We published source code of our work[1].

5. CONCLUSION

In this paper a joint separation-classification (JSC) model has been presented for sound event detection and source separation. A separation mapping from the input time-frequency representation to the segmentation masks and a classification mapping from each segmentation mask to each audio tag are proposed. We obtain frame wise sound event detection EER of 0.14, which outperforms the DNN baseline, and average source separation SDR of 8.08 using global weighted rank pooling compared to SDR of 0.03 using global max pooling. In future, we will research more on improving the source separation quality using the JSC model.

6. ACKNOWLEDGEMENT

This research is supported by EPSRC grant EP/N014111/1 “Making Sense of Sounds” and research scholarship from the China Scholarship Council (CSC). Thank to Sacha Krstulovic, Giacomo Ferroni and Adrian Stepien from Audio Analytic Ltd for discussions on audio event detection.

[1] https://github.com/qiuqiangkong/ICASSP2018_joint_separation_classification

7. REFERENCES

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2007*. IEEE, 2007, pp. 21–26.

[2] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[3] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Machine Listening in Multisource Environments*, 2011.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *EUSIPCO*. IEEE, 2016, pp. 1128–1132.

[5] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[6] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1, 2013.

[7] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[8] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "CHiME-home: A dataset for sound source recognition in a domestic environment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.

[10] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2017.

[11] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 641–645.

[12] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," *arXiv preprint arXiv:1703.06052*, 2017.

[13] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE press, 2006.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[17] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[18] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[19] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[21] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp. 181–197, 2005.

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