Overlapping Acoustic Event Classification Based on Joint Training with Source Separation

Soo Hyun Bae, Inkyu Choi, Hyung Yong Kim, Kang Hyun Lee, and Nam Soo Kim
Department of Electrical and Computer Engineering and INMC,
Seoul National University, Seoul, Korea
E-mail: {shbae, ikchoi, hykim, khlee}@hi.snu.ac.kr, nkim@snu.ac.kr Tel/Fax: +82-2-884-1824

***Abstract*—Overlapping acoustic event classification is the task of estimating multiple acoustic events in a mixed source. In the case of non-overlapping event classification, many approaches have achieved a great success using various feature extraction methods and deep learning models. However, in most real life situations, acoustic events are overlapped and different events may share similar properties. Simultaneously detecting mixed sources is a challenging problem. In this paper, we propose a classification method for overlapping acoustic events which incorporates joint training with source separation framework. Since overlapping acoustic events are mixed in multiple sources, we trained the source separation model and multi-label classification model for estimating the type of overlapping acoustic events. The source separation model is trained to reconstruct the target sources by minimizing the interference of overlapping events. Joint training can be conducted to achieve end-to-end optimization between the acoustic event source separation and multi-label estimation. To evaluate the proposed method, we conducted a number of experiments using artificially mixed data. We observed that the jointly trained neural network outperforms the baseline network with an identical structure except for the training method.**

I. Introduction

Acoustic event is a segment of environmental audio that easily occur in human life, such as coughing, phone ringing, clash sound and so on. Acoustic event classification (AEC) and detection (AED) aim to recognize the audio elements inside an audio clip. Recognizing acoustic events in audio can be utilized in various applications, including indoor environment recognition [1], surveillance systems [2] and automatic audio indexing [3]. Recently, as the interest in this area increases, large datasets [4] were released and challenges such as the detection and classification of acoustic scenes and events (DCASE) challenge have been held.

Research on AED can be separated into two main scenarios, overlapping and non-overlapping. Overlapping AED is a much more challenging problem due to the mixture of acoustic sources and is considered to be more important because acoustic events often overlap with each other in real life recordings.

For a decade, there have been many studies to address the problem of detecting overlapping events from audio. In [5], the author proposed context-dependent hidden Markov models (HMMs) with multiple path decoding. Also non-negative matrix factorization (NMF) approach has been utilized in order to separate overlapping events via dictionary learning [6]. Other approaches were proposed, such as using connectionist temporal classification (CTC) [7], linear dynamical systems for overlapping sound event tracking [8] and feature representation for AED [9]. More recently, various neural network models have been quite successful in AED area. In [10], the multi-label deep neural networks (DNNs) were proposed for detecting of temporally overlapping sound events, and the author in [11] used bi-directional long short term memory (BLSTM).

With regard to AED, although neural networks are able to learn the non-linear relationship between the input and output, they cannot fully utilize each source information from the mixture source. The additive property of sound sources makes it difficult to find the robust features to recognize them in overlapping audio. Thus, we propose a neural network for overlapping AEC which is optimized by the joint training with source separation model and multi-label classification model. The source separation model is trained to reconstruct the target sources from unknown overlapping event. It helps the model to decompose the mixture event. The classification model learns the properties of overlapping event from the reference sources. After that, two models are combined and jointly trained, so that the model can be optimized to minimize the interference of overlapping events and estimate labels of mixed events directly.

The remainder of this paper is organized as follows: section 2 presents the problem formulation of source separation for overlapping AEC. The proposed approach of using joint training for AEC is described in section 3. Section 4 presents the experimental results, and section 5 provides conclusions and future work.

II. Source Separation of Overlapping Acoustic Event

The main objective of source separation is to estimate one or more sources from a given mixed source signal. This can serve as an intermediate step for other tasks. Since overlapping acoustic events are also mixture of multiple signals, source separation framework can be applied to AEC. In [12], unsupervised source separation was used as a pre-processor for overlapping AED. Unlike this approach, the proposed system is trained as a single model including source separation and event classification.

In this section, we focus on source separation of overlapping acoustic events. Given target sources $s_1(t)$ and $s_2(t)$, we define $S_1(t,f)$, $S_2(t,f)$ and $Y(t,f)$ as the short time Fourier

transform(STFT) coefficients of $s_1(t)$, $s_2(t)$ and mixed signal $y(t)$, respectively, where t represents the frame index and f is the frequency-bins. Due to the linearity of the STFT, source separation problem can be defined as follows:

$$\begin{aligned} y(t) &= s_1(t) + s_2(t), \\ Y(t,f) &= S_1(t,f) + S_2(t,f). \end{aligned} \tag{1}$$

In the source separation framework, the magnitude spectrogram of the mixture signal can be approximated as the sum of the magnitude spectra of each source as follows:

$$|Y(t,f)| \approx |S_1(t,f)| + |S_2(t,f)|. \tag{2}$$

For a specific time frame t, the magnitude spectrogram can be written in vector form as follows:

$$\mathbf{y}_t \approx \mathbf{s}_{1_t} + \mathbf{s}_{2_t}, \tag{3}$$

where $\mathbf{y}_t \in \mathbb{R}^F$, $\mathbf{s}_{1_t} \in \mathbb{R}^F$ and $\mathbf{s}_{2_t} \in \mathbb{R}^F$ denote the magnitude spectrum of the mixture and the two target acoustic events at time frame t, respectively. F is the spectral magnitude dimension. Hence, the goal of event separation is to find $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$ using the mixture training data and reference event data.

III. Proposed Method Using Joint Training

In this section, we describe the proposed neural network training scheme for improving the AEC performance. The schematic of the proposed approach can be seen in Figure 1.

A. Source Separation Model

Various DNN based approaches have been proposed to address the monaural source separation problem [13], [14], [15]. In order to obtain the estimated single event from overlapping acoustic events, we exploit the DNN framework for source separation. Given the input mixture features $\mathbf{y}_t$ from the mixture, we obtain the output estimates $\hat{\mathbf{y}}_{1_t}$ and $\hat{\mathbf{y}}_{2_t}$ from the network. In the training process, the discriminative objective function is used in order to regularize the reconstruction error as defined in [13]

$$L(t) = \|\hat{\mathbf{y}}_{1_t} - \mathbf{s}_{1_t}\|^2 + \|\hat{\mathbf{y}}_{2_t} - \mathbf{s}_{2_t}\|^2 - \gamma\|\hat{\mathbf{y}}_{1_t} - \mathbf{s}_{2_t}\|^2 - \gamma\|\hat{\mathbf{y}}_{2_t} - \mathbf{s}_{1_t}\|^2, \tag{4}$$

where $\|\cdot\|$ indicates the l_2-norm and γ denotes the regularization parameter which adjusts the trade-off between the reconstruction error and the discrimination information. In order to estimate each source, the soft time-frequency mask $\mathbf{m}_t \in \mathbb{R}^F$ is calculated as follows:

$$\mathbf{m}_t = \frac{|\hat{\mathbf{y}}_{1_t}|}{|\hat{\mathbf{y}}_{1_t}| + |\hat{\mathbf{y}}_{2_t}|}. \tag{5}$$

Then Wiener filtering can be used to reconstruct the magnitude spectra of each acoustic event source as follows:

$$\begin{aligned} \hat{\mathbf{s}}_{1_t} &= \mathbf{m}_t \otimes \mathbf{y}_t, \\ \hat{\mathbf{s}}_{2_t} &= (1 - \mathbf{m}_t) \otimes \mathbf{y}_t, \end{aligned} \tag{6}$$

where the division is performed element-wise and $\otimes$ indicates element-wise multiplication. The source separation model is trained through the mixture source $\mathbf{y}_t$ as an input and reference source $[\mathbf{s}_{t_1}\mathbf{s}_{t_2}]$ as a target. This process is described in Figure 1 by the solid blue line box.

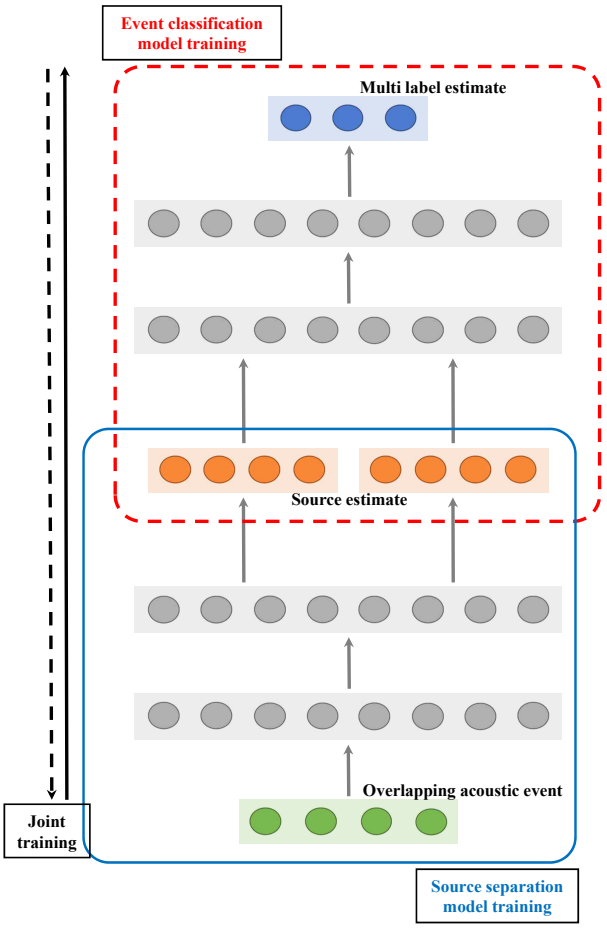

Fig. 1. Joint training structure for the proposed technique.

B. Multi-Label Classification Model

Multi label neural networks are utilized for detection of temporally overlapping acoustic events[10]. In the training stage of multi-label classification, the network learns the mapping between reference source $[\mathbf{s}_{t_1}\mathbf{s}_{t_2}]$ as an input and the corresponding target output $\mathbf{a}_t$, where $\mathbf{a}_t \in \mathbb{R}^I$ indicates true multi-label vector of overlapping acoustic events. I is the number of acoustic events. This process is shown in Figure 1 by the red dashed line box.

C. Joint Training Method

Jointly trained models have achieved improvement in various learning tasks, especially in the speech recognition area. Motivated by the good performance of the joint training scheme shown in [16], [17], [18], we use this technique in order to improve AEC performance. AEC is also suitable enough to adopt the joint training because source separation and event classification are trained through the difference objectives.

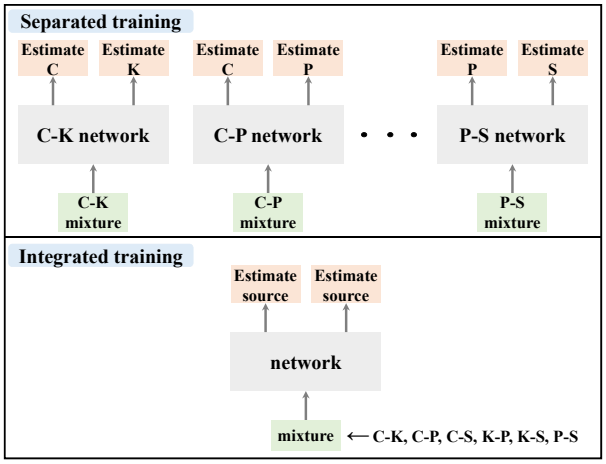

Fig. 2. Comparison between separated and integrated source separation models.

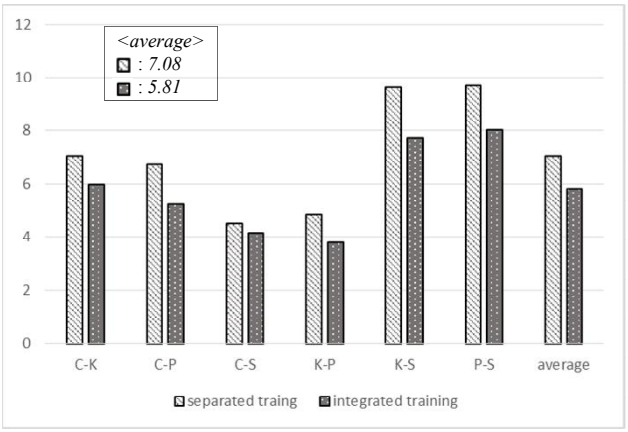

Fig. 3. The source separation performance (SDR [dB])

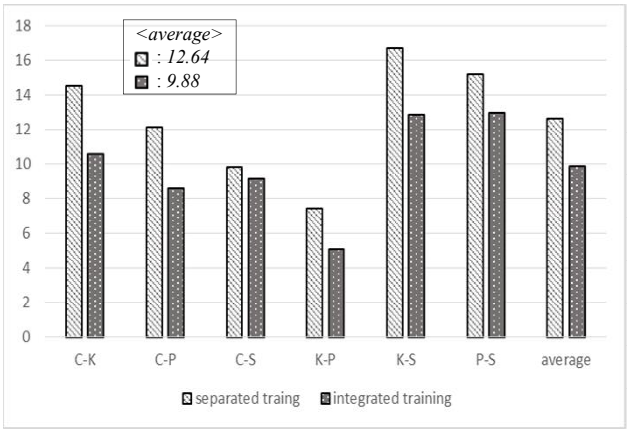

Fig. 4. The source separation performance (SIR [dB])

After two networks are trained, they are combined to form a single network and further trained jointly. In the training phase, the network is trained with mixture source $\mathbf{y}_t$ as input and true label $\mathbf{a}_t$ as output. As shown in Figure 1, the weights of the unified network are adjusted using back-propagation. As a result, the network is trained to utilize the information of separated source implicitly. This helps the network to estimate acoustic events from the mixture source.

IV. Experiments

A. Dataset and Data Augmentation

In order to evaluate the performance of the proposed method, we conducted a set of acoustic event source separation experiments using the IEEE DCASE 2016 Challenge Task 2 Train Datasets[19]. The training dataset consists of 20 isolated sound events per event class. We selected four acoustic events: coughing, keyboard typing, page turning and speech, and used them to construct a mixed source dataset. Six different types of dataset were generated in the source mixing process ($_4C_2 = 6$). Unlike most speech datasets which usually consist of hours of data or more, conventional sound event datasets are not sufficiently long enough to train a robust DNN model. In order to tackle the insufficient data problem to train a DNN model, data augmentation approach was used for training the DNN. To construct the diverse source mixtures from a small dataset, acoustic events were artificially generated by time stretching. Finally, various mixture combination of two acoustic events were produced with SNR 0 dB scale.

B. Experimental Setup

The dataset were sampled at 16 kHz, and the magnitude spectrograms were calculated using STFT. Hamming window with 512-point length and 75% overlap was applied and the FFT was taken at 512 points. Only the first 257 FFT points were used since the conjugate of the remaining 255 FFT point are symmetric with the first half.

As a regression model for source separation training, we built a DNN with two hidden layers with 1000 Rectified Linear Unit (ReLU). The input features were 257×7-dimension (current frame and the three previous and next frames of mixture source), and the output was 257×2-dimension (regression of two target sources) with sigmoid unit. Equation (4) was used as the loss function, and the number of randomly ordered mini-batches in each epoch was set to be 100. After processing each mini-batch, the weights were updated using Adam [20]. In order to mitigate the over-fitting problem in the training phase, dropout was applied with a probability of 30% for all hidden layer.

In order to predict the labels of overlapping acoustic events, we also trained a DNN consisting of two hidden layer with 1000 ReLU node. The input features were 257×2 (two separated source), and output was 4-dimensional (each acoustic event label) softmax layer. Mean squared error was used as the loss function and other setup was equivalent to the source separation model.

TABLE I
OVERLAPPING ACOUSTIC EVENT CLASSIFICATION PERFORMANCE.

Event class	Precision			Recall			F-score		
	2L-DNN	5L-DNN	Proposed	2L-DNN	5L-DNN	Proposed	2L-DNN	5L-DNN	Proposed
C	0.8677	0.9220	**0.9785**	0.9294	0.9067	**0.9333**	0.8975	0.9143	**0.9554**
K	**0.9085**	0.8987	0.8898	0.9308	0.9467	**0.9841**	0.9195	0.9221	**0.9346**
P	0.9477	0.9341	**0.9852**	0.8767	0.8821	**0.8967**	0.9108	0.9074	**0.9389**
S	0.9526	0.9377	**0.9685**	0.8667	0.9067	**0.9765**	0.9076	0.9219	**0.9725**
average	0.9191	0.9231	**0.9555**	0.9009	0.9106	**0.9477**	0.9089	0.9164	**0.9503**

After training the source separation model and the multi-label classification model, two networks were cascaded to form a single larger network and the weights of the unified network were adjusted using back-propagation.

C. Evaluation of Source Separation

In many two source separation tasks, a single network is trained to estimate a source pair. However, in the proposed source separation network, a single network estimates six source pairs (${}_4C_2 = 6$). This means that if the source separation network do not estimate the target sources well, the jointly trained network may show similar performance to the baseline network which has an identical structure including model size and hyper-parameters, but without applying the joint training scheme. In order to verify this point, we compared the source separation performance of the proposed method and networks which were trained using a mixture dataset, where each recordings consist of only two target sources as shown in Figure 2. Alphabets denote the acoustic event name, C: Coughing, K: Keyboard typing, P: Page turning and S: Speech. The 'C-K' means that the mixture source includes coughing sound and keyboard typing sound. The 'separated training' indicates that a single network was trained using only a mixture dataset. Thus, total six networks were produced. The 'integrated training' means that a single network was trained using whole six pair datasets.

The performance of source separation was evaluated in terms of the signal to distortion ratio (SDR) and signal to interference ratio (SIR) [21]. Figure 3 and Figure 4 show the source separation performance. As shown in the figures, although the performance is degraded, the proposed source separation network is enough to provide each source information to multi-label classification network.

D. Acoustic Event Classification Results

To evaluate the performance of proposed method, we calculated the number of $correct$, $missed$ and $false\ alarm$ events. The $precision$, $recall$ and $F\text{–}score$ are calculated as follows:

$$precision = \frac{correct}{correct + false\ arlam}, \tag{7}$$

$$recall = \frac{correct}{correct + missed}, \tag{8}$$

$$F\text{–}score = \frac{2 \times precision \times recall}{precision + recall}. \tag{9}$$

Table 1 shows the overlapping AEC performance. '2L-DNN' and '5L-DNN' denote DNN structures which have two and five hidden layers. These baseline networks did not apply the joint training with source separation. The proposed method was found to improve the classification performance and achieve an average $F\text{–}score$ of 0.9503. In the each acoustic source, the joint training with source separation achieved higher performance.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a neural network for overlapping AEC based on joint training between source separation model and multi-label classification model. By adopting the source separation framework into the overlapping AEC task, the jointly trained network can minimize the interference of overlapping events. From the experimental results, it has been found that the proposed technique outperforms the baseline networks which do not apply the joint training with source separation. The future work will focus on overlapping AED that is able to detect more various acoustic sources.

ACKNOWLEDGMENT

This work was supported by the research fund of Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and Agency for Defense Development of Korea.

REFERENCES

[1] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[2] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[3] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 4, no. 2, p. 11, 2008.

[4] J. F. Gemmeke, D. P. 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[5] 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. 1–13, 2013.

[6] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*. Springer, 2013, pp. 341–371.

[7] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
[8] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Polyphonic sound event tracking using linear dynamical systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1266–1277, 2017.
[9] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.
[10] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.
[11] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.
[12] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8677–8681.
[13] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.
[14] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Discriminative enhancement for single channel audio source separation using deep neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 236–246.
[15] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.
[16] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 92–101, 2015.
[17] K. H. Lee, S. J. Kang, W. H. Kang, and N. S. Kim, "Two-stage noise aware training using asymmetric deep denoising autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5765–5769.
[18] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.
[19] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.
[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[21] 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.