Robust Sound Event Detection by a Two-stage Network in the presence of Background Noise

Jie Ou[1], Hongqing Liu[1], Yi Zhou[1], and Lu Gan[2]

[1] School of Communication and Information Engineering Chongqing University of Posts and Telecommunications Chongqing, China
[2] College of Engineering, Design and Physical Science, Brunel University. London UB8 3PH, U.K.
oj-cqupt@outlook.com, hongqingliu@outlook.com

Abstract. With the advent of deep learning, research on noise-robust sound event detection (SED) has progressed rapidly. However, SED performance in noisy conditions of single-channel systems remains unsatisfactory. Recently, there were several speech enhancement (SE) methods for the SED front-end to reduce the noise effect, which are completely two models that handle two tasks separately. In this work, we introduced a network trained by a two-stage method to simultaneously perform signal denoising and SED, where denoising and SED are conducted sequentially using neural network method. In addition, we designed a new objective function that takes into account the Euclidean distance between the output of the denoising block and the corresponding clean audio amplitude spectrum, which can better limit the distortion of the output features. The two-stage model is then jointly trained to optimize the proposed objective function. The results show that the proposed network presents a better performance compared with single-stage network without noise suppression. Compared with other recent state-of-the-art networks in the SED field, the performance of the proposed network model is competitive, especially in noisy environments.

Keywords: sound event detection, denoising block, two-stage method, neural network.

1 Introduction

Sound event detection (SED) is currently an important research topic in the field of acoustic signal processing. The purpose of SED is to detect specific sound events in different scenes and to locate the onset and offset times of target sound events present in an audio recording. This technique has great influences in many fields, such as anomaly detection, acoustic surveillance, and smart house [1–3]. To promote SED, the Acoustic Scene and Event Detection and Classification (DCASE) Challenge was launched as an international challenge in 2013.

Since DCASE released Task 2 in 2017, rare sound event detection has received more and more attention [4–6]. In [7], Lim et al. introduced a rare sound event detection system using the combination of one-dimensional (1D) convolutional neural network (1D ConvNet) and recurrent neural network (RNN) with long short-term memory units (LSTM), and ranked the first place in DCASE task 2. Kao et al. proposed a Region-based CRNN (R-CRNN) [8] to improve previous work. In [9], Zhang et al. proposed a multiscale time-frequency CRNN (MTF-CRNN) with low parameter counts for SED. In [10], He et al. proposed a time-frequency attention model for sound event detection to alleviate the problems caused by data imbalance. These models achieve excellent performance.hongqingliu@outlook.com

The negative impact of uncorrelated environmental noise on the performance of the SED system has attracted increasing attentions. With the developments of speech enhancement technology, in the field of acoustic signal processing, many researchers have conducted noise suppression preprocessing before training their model [11–13]. In the past single-channel SED tasks, similar methods have also been used. In response to this problem, Zhou et al. proposed robust sound event detection through noise estimation and source separation using NMF [14]. Later, Feng et al. proposed an adaptive noise reduction method for sound event detection based on non-negative matrix factorization (NMF) [15]. Wan et al. proposed noise robust sound event detection using deep learning and audio enhancement [16]. In wan's method, first, the noisy speech is denoised using the Log Spectrum Amplitude Estimation (OMLSA) audio enhancement method. Then the denoised audio is used as the input of the SED network. This scheme can improve the performance of the SED system, but at the same time there are some shortcomings in two aspects. First, two different models are used to handle two different tasks separately. Especially during the inference period, it is also necessary to run the two models separately in sequence, which is relatively redundant and complicated. Second, during training, the audio distortion caused by the denoising system will cause the performance of the SED system to seriously degrade. A method should be found to punish the degree of distortion caused by the denoising system during training. In this way, the two systems before and after can be related to each other.

To mitigate this problem, in this paper, inspired by a two-stage enhancement strategy [17] for impaired speech, we propose a two-stage deep learning network for SED. In the proposed network, we first train denoising network and the SED network separately so that the network learns the weights of the corresponding tasks. After that, we jointly train the two modules to make the network learn the ability to handle two tasks at the same time, which is relatively difficult for combining essentially two different tasks. Finally, an optimized system that can perform denoising and SED tasks at the same time is obtained. In addition, we propose a new weighted loss function, which can punish the distortion caused by the denoising network to associate the front and back models. This loss function plays an important role in the improvement of model performance.

The remainder of this paper is organized as follows. In Section 2, we first introduce the single-task signal model separately, and then, the proposed two-stage deep learning network is presented. The dataset, experimental setups, and evaluation metrics are illustrated in Section 3. The results and analysis are given in Section 4. Finally, we conclude our work in Section 5.

2 Methods

In this section, we first introduce the notations and then describe the frequency-domain denoising networks that we use in our experiments. Figure 1 shows a schematic diagram of the proposed system, which consists of a denoising block and a SED block.

2.1 Signal Model

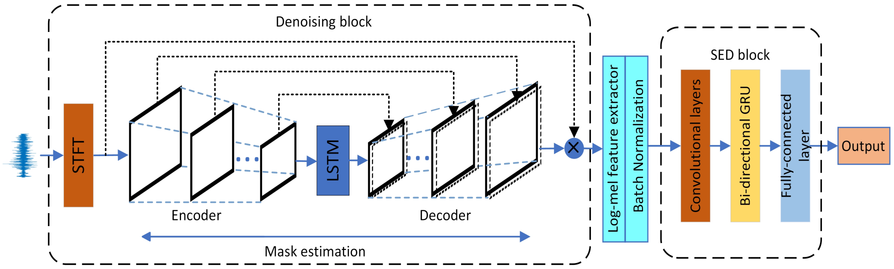


Fig. 1. *System diagram of the proposed two-stage model.*

Let us consider a single-channel microphone signal, denoted by $y(t)$

$$y(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ is the clean signal (the target event), $n(t)$ is background noise, and t is a time index. In this paper, the background noise does not include obvious target detection events. The purpose of the first level of the two-stage system is to recover a clean signal $s(t)$ from the corresponding noisy observations $y(t)$ to prepare for the SED.

2.2 Denoising block

The denoising block in Figure 1 is a schematic diagram of the denoising NN framework in the frequency domain. The denoising block is performed by 2 steps, (1) STFT, (2) mask estimation.

(1) STFT: When working in the frequency-domain [18], we usually compute the amplitude of the STFT coefficients to work on real numbers before inputting them to the mask estimation network. After the microphone signal is transformed by STFT, the time-frequency spectrum can be obtained, denoted by $\mathbf{x}_y$, where $\mathbf{x}_y$ is the amplitude spectrum of the noise signal $y(t)$.

(2) Mask estimation: We estimate a denoising mask using a mask estimation network. The structure of this mask estimation network is similar to the U-Net, which is a well known architecture composed as a convolutional autoencoder with skip-connections [19]. The mask estimation NN should account for the time context of the signal to distinguish speech from noise. This can be achieved by using BLSTM layers. In mask estimation, convolutional encoder is stacked by convolutional layers and pooling layers. It is used to extract high-level features from the original input signals. The structure of the decoder is basically the same as the encoder, but the order is reversed. The decoder maps the low-resolution feature map at the output of the encoder to the full feature map of the input size. The symmetrical encoder-decoder structure ensures that the output has the same shape as the input. Because the detection network only needs a clean sound spectrum with no change in size, this better prepares the conditions for the input of the latter stage.

The output of the mask estimation network is a single mask used to predict speech as

$$\mathbf{m}_s = MSnet(\mathbf{x}_y), \tag{2}$$

where MSnet($\cdot$) is a mask estimation network, and $\mathbf{m}_s$ is the masks associated with speech.

We apply ELUs [20] to all convolutional and deconvolutional layers except the output layer. In the output layer, we use softmax activation [21], which can constrain the network output to always be positive. The input size and output size of each layer are unified in the form of *Feature Maps* × *Time Step* × *Frequency Channel.*

For frequency-domain networks, we use the mean square error loss (MSE) as the loss

$$\text{Ł}_{s,y}(\theta) = \frac{1}{M}(\|\mathbf{x}_s - \mathbf{m}_s \odot \mathbf{x}_y\|_2), \tag{3}$$

where $\mathbf{x}_s$ is the amplitude spectrum of the target speech (the clean speech) signal, $\odot$ is an element-wise multiplication, θ is the parameter of the denoising network, $\|\cdot\|_2$ is Euclidean norm (L2 norm), M is the total number of pixels, and its size is *Batch Size* × *Time Step* × *Frequency Channel.*

2.3 SED block

For SED, we follow the state-of-the-art CRNN framework as a baseline [10]. In the SED block of Figure 1, the CRNN structure consists of three parts of CNN, RNN and Fully Connected Layer .

The convolution part of the network consists of four convolution layers, and each layer is followed by batch normalization [22], ReLU activation unit, and dropout layer [23]. Since we believe that the early convolutional layer is essential for feature learning, the first two convolutions of the network are stacked back to back. In order to maintain the most important information on each feature map, we use the max pooling layer on the time axis and frequency axis. The final feature map is reduced by four times in the time axis to match the frame resolution (80 ms) for computing the evaluation metrics. In the end of the CNN, the features extracted on different convolution channels are superimposed along the frequency axis.

In the RNN part, we use a bi-directional gated recurrent unit (bi-GRU) layer, which can better extract the time structure of acoustic events compared with unidirectional GRU. Bi-GRU encodes sequence feature into a sequence of feature vectors of size (375, U), where U is the number of GRU units. The returned features from the GRU layer is maintained and sent to a fully-connected layer (FCL) with an output size of C, where C is the number of event classes. After the FCL, sigmoid activation is used to produce classification result for each frame (80 ms), of which output represents the probability of the presence of the target sound event.

Finally, we set a binary prediction to a constant threshold of 0.5 for each frame. These predictions are post-processed through a median filter of length 240 ms. We choose the longest continuous forward prediction sequence to produce the start and offset of the target event.

2.4 Weighted multi-task loss

In many cases, the denoised audio will have a small amount of signal distortion compared with the clean target signal. Therefore, the target sound event in the denoised audio may be distorted, which will seriously affect the accuracy of the SED system in estimating the onset and offset of the target event. To mitigate this problem, in this paper, we propose an improved weighted multi-task loss function. Our weighted multi-task(WMT) loss $\text{Ł}_{wmt}(\theta)$ is defined as,

$$\text{Ł}_{wmt}(\theta) = \lambda E(\parallel \mathbf{x}_s - f(\mathbf{x}_y) \parallel_2) + \text{Ł}_{bce}(q \parallel p) \tag{4}$$

where f function represents denoising block, $\parallel \cdot \parallel_2$ is Euclidean norm (L2 norm), λ denotes the penalty factor of the denoising block, Ł_{bce} represents cross-entropy loss function, p and q are the output probability and label of the proposed model, respectively.

Our experiments have shown that penalizing the denoised block by measuring the Euclidean distance between the spectrogram of the clean speech and the amplitude spectrum of the denoised speech can solve this problem. This is also in full compliance with our assumptions. When the denoising block produces large distortion, $\text{Ł}_{wmt}(\theta)$ will punish the denoising block and force the network to learn in the correct direction so that the effect of distortion is minimized. Conversely, when the denoising block produces less distortion, the penalty produced by $\text{Ł}_{wmt}(\theta)$ will become smaller.

2.5 Two-stage Network

In Figure 1, we connect the denoising block and SED block into a larger network for joint optimization. In the denoising stage, we transform signal (noisy and clean) to 883-dimensional STFT. The obtained amplitude spectrum is used as the input feature. In the detection stage, 128 Mel-scale filters are applied to the amplitude spectrum output by denoising block on each frame, covering the frequency range of 300 to 22050 Hz. We add a batch normalization layer before sending the estimated features to the next level network to ensure that the input of the SED block is correctly normalized. During training, this layer keeps exponentially moving averages on the mean and standard deviation of each mini-batch. During testing, such running mean and standard deviation are fixed to perform normalization. By using the features processed by normalization and log-mel filters, we expect closer coupling between the separately trained denoising stage and SED stage, which can benefit joint training. After that, the normalized log-Mel features are directly sent to the SED block for detection. Since each step above is differentiable, we can derive the error gradients to jointly train the whole system.

Before joint training, the denoising block and SED block are trained separately, and the obtained parameters are used for the initializations of the two-level SED system.This network is considered to be a large network that can handle the dual tasks of removing noise and sound event detection at the same time, which is different from other methods.

3 Experiments

3.1 Data

We demonstrate the proposed model on DCASE 2017 Challenge task 2 [24]. The task dataset consists of isolated sound events for each target class and recordings of everyday acoustic scenes to serve as background. The task dataset consists of three target event categories: baby crying, breaking glass, and shooting. A synthesizer for creating mixtures at different event-to-background ratios (EBRs) is also provided. The dataset comprised of development dataset and evaluation dataset. The environmental noise and target sound format in the evaluation data set did not appear in the training data set. The development dataset also includes two parts: training subset and test subset. The detailed information about this task and dataset can be found in [24, 25].

We use the provided synthesizer to generate 5000 mixtures for each target class and generate the clean signal corresponding to each mixed audio as a label for denoising block pre-training. In order to simulate different EBR environments, we set EBR to three situations of -6dB, 0dB, and 6dB. In order to obtain more positive samples and alleviate the problem of data imbalance, the probability of occurrence of the event is set to 0.9 (the default value is 0.5).

3.2 Experimental setup

The training is divided into pre-training and joint training. Before joint training, the denoising block and SED block need to be pre-trained separately. For the pre-training of the denoising block, the amplitude spectrum feature of the noisy speech is used as the input of the denoising network, and the amplitude spectrum corresponding to the clean speech is used as the label. The number of GRU units U is 32 in SED block. During the joint training, the pre-trained neural networks are used to initialize the weights of the joint network to achieve a better optimization and accelerate the optimization process. The two-stage network block is trained with the Adam optimizer [26]. λ is set to 0.2. The learning rate is 0.001 for the first 60 epochs and is then decayed by 10 % after each epoch that follows. The training stops after 100 epochs. The batch size is 16 and sigmoid activation is used on the last layer of the FCL for our classification model. The output probability distribution is in the continuous range of $[0, 1]$.

3.3 Metrics

We follow the official evaluation metrics of DCASE Challenge. There are two types of event-based metrics: event-based error rate (ER) and event-based F-score. The definition of these two evaluation calculations can be found in [27]. The correct prediction only needs to consider the existence of the target event and its onset time. If the output accurately predicts the presence and onset of the target event, we express it as correct detection. The onset detection is considered accurate only when it is predicted within the range of 500 ms of the actual onset time. The ER is the sum of deletion error and insertion error, and F-score is the harmonic average of precision and recall.

4 Results

4.1 Experimental results

Table 1. *Performance of the proposed model, baseline method, and noise robust networks, $***$ indicates that class-wise results are not given in related paper. We compared other noise robust networks and there is no denoising block in the baseline.*

<table>
<tr><th rowspan="2">Model</th><th rowspan="2">Metric</th><th colspan="4">Development Dataset</th><th colspan="4">Evaluation Dataset</th></tr>
<tr><th>babycry</th><th>glassbreak</th><th>gunshot</th><th>average</th><th>babycry</th><th>glassbreak</th><th>gunshot</th><th>average</th></tr>
<tr><td>supervised NMF [14]</td><td>ER|F-score</td><td>******</td><td>******</td><td>******</td><td>******</td><td>0.17| 91.4</td><td>0.22| 89.1</td><td>0.55 |72.0</td><td>0.31|84.2</td></tr>
<tr><td>Subband-Weighted NMF [15]</td><td>ER|F-score</td><td>******</td><td>******</td><td>******</td><td>******</td><td>**0.10|94.8**</td><td>0.06|96.9</td><td>0.46|76.2</td><td>0.21|89.3</td></tr>
<tr><td>Baseline</td><td>ER|F-score</td><td>0.16|88.4</td><td>0.06|92.2</td><td>0.24|85.9</td><td>0.15|88.3</td><td>0.32|83.2</td><td>0.22|87.3</td><td>0.37|80.2</td><td>0.30| 83.5</td></tr>
<tr><td>Proposed</td><td>ER|F-score</td><td>0.09|95.8</td><td>0.03|98.6</td><td>0.04|96.7</td><td>0.05|97.0</td><td>0.16|91.3</td><td>0.06|97.1</td><td>0.12|94.1</td><td>0.11|94.2</td></tr>
<tr><td>**Proposed+MWT**</td><td>ER|F-score</td><td>**0.07|96.5**</td><td>**0.02|99.0**</td><td>**0.04|97.1**</td><td>**0.04|97.5**</td><td>0.14|92.9</td><td>**0.04|97.8**</td><td>**0.09|95.0**</td><td>**0.09|95.2**</td></tr>
</table>

The ER and F-score of the proposed model and other models are shown in Table 1. Results show that the proposed outperforms the baseline due to the noise

Table 2. *Performance of the proposed model and other state-of-the-art methods, * * * indicates that class-wise results are not given in related paper. We compare the following models:(1)1d-CRNN: DCASE 1st place model;(2)R-CRNN: Region-based CRNN;(3)MTF-CRNN: Multi-scale CRNN.(4)TFA: temporal-frequential attention CRNN;*

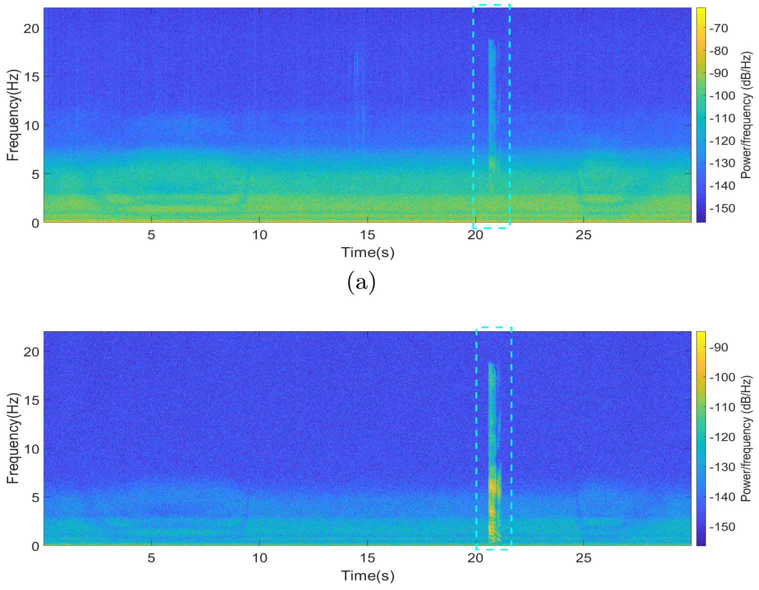
<table>
<tr><th rowspan="2">Model</th><th rowspan="2">Metric</th><th colspan="4">Development Dataset</th><th colspan="4">Evaluation Dataset</th></tr>
<tr><th>babycry</th><th>glassbreak</th><th>gunshot</th><th>average</th><th>babycry</th><th>glassbreak</th><th>gunshot</th><th>average</th></tr>
<tr><td>1d-CRNN [7]</td><td>ER |F-score</td><td>0.05 |97.6</td><td>0.01|99.6</td><td>0.16|91.6</td><td>0.07|96.3</td><td>0.15|92.2</td><td>0.05|97.6</td><td>0.19|89.6</td><td>0.13|93.1</td></tr>
<tr><td>R-CRNN [8]</td><td>ER|F-score</td><td>0.09|* * *</td><td>0.04|* * *</td><td>0.14|* * *</td><td>0.09|95.5</td><td>* * * * * *</td><td>* * * * * *</td><td>* * * * * *</td><td>0.23|87.9</td></tr>
<tr><td>MTF-CRNN [9]</td><td>ER|F-score</td><td>0.13|91.8</td><td>0.04|97.6</td><td>0.11|93.3</td><td>0.09|94.2</td><td>0.15|89.7</td><td>0.08|95.1</td><td>0.28|83.9</td><td>0.17|89.2</td></tr>
<tr><td>TFA [10]</td><td>ER|F-score</td><td>0.10|95.1</td><td>0.01|99.4</td><td>0.16|91.5</td><td>0.09|95.3</td><td>0.18|91.3</td><td>0.04|98.2</td><td>0.17|90.8</td><td>0.13|93.4</td></tr>
<tr><td><b>Proposed+MWT</b></td><td>ER|F-score</td><td>0.07|96.5</td><td>0.02|99.0</td><td>0.04|97.1</td><td><b>0.04</b>|<b>97.5</b></td><td>0.14|92.9</td><td>0.04|97.8</td><td>0.09|95.0</td><td><b>0.09</b>|<b>95.2</b></td></tr>
</table>

suppression. Compared with other noise robustness models, the performance of the proposed two-stage is also superior on evaluation datasets, which indicates that the proposed network is more robust to noise. In addition, using WMT as the loss function can further improve the performance of the proposed SED system, which verifies that the WMT mentioned above can reduce the distortion caused by the denoising block. We point out that the final model we get is a complete model, which is fundamentally different from the two models proposed by Wan et al., especially during model testing.

Table 2 shows performance comparisons of the proposed model and other state-of-the-art SED methods in terms of ER and F-score. Compared with other approaches, the performance of our model is also competitive. The average ER (0.09) and average F-score (95.2%) of the proposed model are better than those of all models. Note that of the top 1 teams adopt ensemble method. Although Lim et al. [7] achieves relatively good results, its final decision is made by combining the output probabilities of more than four models with different time steps and different data mixtures. However, the proposed model is treated as a single model.

4.2 Denoising block visualization

To better understand our proposed network, We visualized the output of the denoising block in the two-level network. Figure 2 shows that the two-stage model we proposed has actually learned how to suppress noise. We selected an audio with a baby crying to visualize and the baby's crying occurs from 20.49 seconds to 21.43 seconds, which is marked by a blue frame. The target event is under park noise. After noise suppression, the target sound becomes clearer. This confirms that our proposed network achieves the dual tasks of denoising and detection at the same time.



(a)

(b)

Fig. 2. *Visualization of denoising block output.(a) spectrogram of an noisy speech, where the blue box denotes target event. (b) spectrogram of denoising block output.*

5 Conclusion

In order to solve the impact of noise on SED, in this work, we propose a two-stage model to perform the joint task of signal denoising and SED. In addition, we propose a loss function that is conducive to network optimization. Our system can achieve the best performance on DCASE evaluation dataset. Compared with other noise robust networks, the joint network performance is better. Compared with other networks, the proposed model also outperforms them thanks to the noise suppression. The large improvement demonstrates the benefits of introducing the denoising block.

References

1. Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288, 2015.
2. Nguyen Cong Phuong and Tran Do Dat. Sound classification for event detection: Application into medical telemonitoring. In *International Conference on Computing*, 2013.
3. Chloé Clavel, Thibaut Ehrette, and Gaël Richard. Events detection for an audio-based surveillance system. *ICME*, pages 1306–1309, 2005.

4. Jan Baumann, Timo Lohrenz, Alexander Roy, and Tim Fingscheidt. Beyond the dcase 2017 challenge on rare sound event detection: A proposal for a more realistic training and test framework. *ICASSP*, pages 611–615, 2020.
5. Weiran Wang, Chieh-Chi Kao, and Chao Wang. A simple model for detection of rare sound events. *Interspeech*, 2018.
6. Shimada Kazuki, Koyama Yuichiro, and Inoue Akira. Metric learning with background noise class for few-shot detection of rare sound events. *ICASSP*, pages 616–620, 2019.
7. Hyungui Lim, Jeongsoo Park, and Yoonchang Han. Rare sound event detection using 1D convolutional recurrent neural networks. Technical report, DCASE2017 Challenge, September 2017.
8. Chieh-Chi Kao, Weiran Wang, Ming Sun, and Chao Wang. R-crnn: Region-based convolutional recurrent neural network for audio event detection. *Interspeech*, pages 1358–1362, 2018.
9. Keming Zhang, Yuanwen Cai, Yuan Ren, Ruida Ye, and Liang He. Mtf-crnn: Multiscale time-frequency convolutional recurrent neural network for sound event detection. *IEEE Access*, PP(99):1–1, 2020.
10. Yu-Han Shen, Ke-Xin He, and Wei-Qiang Zhang. Learning how to listen: A temporal-frequential attention model for sound event detection. *arXiv: Sound*, pages 2563–2567, 2019.
11. Kinoshita Keisuke, Ochiai Tsubasa, Delcroix Marc, and Nakatani Tomohiro. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. *ICASSP*, pages 7009–7013, 2020.
12. Morten Kolbæk. Single-microphone speech enhancement and separation using deep learning. *arXiv: Sound*, 2018.
13. Jahn Heymann, Lukas Drude, Christoph Böddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach. Beamnet: End-to-end training of a beamformer-supported multi-channel asr system. *ICASSP*, pages 5325–5329, 2017.
14. Q. Feng Z. Zhou. Robust sound event detection through noise estimation and source separation using nmf. *In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop(DCASE2017)*, 2017.
15. Qing Zhou, Zuren Feng, and Emmanouil Benetos. Adaptive noise reduction for sound event detection using subband-weighted nmf. *Sensors (Basel, Switzerland)*, 2019.
16. Tongtang Wan, Yi Zhou, Yongbao Ma, and Hongqing Liu. Noise robust sound event detection using deep learning and audio enhancement. *ISSPIT*, pages 1–5, 2019.
17. Yan Zhao, Zhong Qiu Wang, and De Liang Wang. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27:53–62, 2018.
18. Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. *Interspeech*, pages 3229–3233, 2018.
19. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.
20. Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *Computer ence*, 2015.
21. Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *AISTATS*, pages 315–323, 2011.
22. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, 2015.

23. Nitish Srivastava, E. Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958, 2014.
24. Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. *EUSIPCO*, pages 1128–1132, 2016.
25. A Mesaros, T Heittola, A Diment, B Elizalde, A Shah, E Vincent, and B and Raj. Dcase 2017 challenge setup: Tasks, datasets and baseline system. 2017.
26. P. Diederik Kingma and Lei Jimmy Ba. Adam: A method for stochastic optimization. *international conference on learning representations*, 2015.
27. A Mesaros, T Heittola, and T Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 2016.