ON THE EFFECTIVENESS OF SPATIAL AND MULTI-CHANNEL FEATURES FOR MULTI-CHANNEL POLYPHONIC SOUND EVENT DETECTION

Thi Ngoc Tho Nguyen[1]*, *Douglas L. Jones*[2], *Woon Seng Gan*[1],

[1] Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore, {nguyenth003, ewsgan}@ntu.edu.sg
[2] University of Illinois at Urbana-Champaign, Dept. of Electrical and Computer Engineering, Illinois, USA, {dl-jones}@illinois.edu

ABSTRACT

Multi-channel log-mel spectrograms and spatial features such as generalized cross-correlation with phase transform have been demonstrated to be useful for multi-channel polyphonic sound event detection for static-source cases. The multi-channel log-mel spectrograms and spatial features are often stacked along the channel dimension similar to RGB images before being passed to a convolutional model to detect sound events better in multi-source cases. In this paper, we investigate the usage of multi-channel log-mel spectrograms and spatial features for polyphonic sound event detection in both static and dynamic-source cases using DCASE2019 and DCASE2020 sound event localization and detection datasets. Our experimental results show that multi-channel log-mel spectrogram and spatial features are more useful for static-source cases than for dynamic-source cases. The best use of multi-channel audio inputs for polyphonic sound event detection in both static and dynamic scenarios is to train a model that use all the single-channel log-mel spectrograms separately as input features and the final prediction during the inference stage is obtained by taking the arithmetic mean of the model's output predictions of all the input channels.

***Index Terms*—** DCASE, moving sound sources, multi-channel input, spatial features, sound event detection.

1. INTRODUCTION

Sound event detection (SED) has wide applications in urban sound sensing [1], wild life monitoring [2], and surveillance [3]. The SED task recognizes the sound class, the onsets and offset of a detected sound event. Polyphonic SED refers to cases where there are multiple sound events overlapping in time. The sound sources can be spatially static or dynamic. A well-trained SED model is expected to be robust to the source movement in space.

*This research was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060.

In the past decade, deep learning has achieved great success in classifying, tagging, and detecting sound events [4]. The state-of-the-art SED models are often built from convolutional neural networks (CNN) [1], recurrent neural networks (RNN) [5], and convolutional recurrent neural networks (CRNN) [6]. The polyphonic SED is formulated as multi-label multi-class classification where several sound classes can be simultaneously active at a given time step.

For single-channel SED task, log-mel spectrogram is the most commonly used feature [6, 7, 8] thanks to its compactness and simplicity to extract. Other single-channel input features for SED are raw audio signal [9], mel-frequency cepstral coefficient (MFCC) [10], spectrogram image feature [10]. For multi-channel SED task, log-mel spectrograms of all the input channels are often stacked together along the channel dimension to form a 3 dimensional (3D) input feature [11, 12]. In addition, many spatial features have been proposed for polyphonic SED. Adavanne *et al.* use multi-channel log-mel spectrogram, generalized cross-correlation with phase transform (GCC-PHAT), and auto-correlation to detect sound events using binaural audio [11]. Cao *et al.* employ multi-channel log-mel spectrograms together with GCC-PHAT, and intensity vector for SED [13]. Multi-channel features have also been applied in related domain such as automatic speech recognition [14] and source separation [15]. The motivation of using spatial features for SED is to train a model to learn spatial information to recognize isolated and overlapping sound events. However, the underlying mechanism of how a model trained with spatial features are able to detect multiple overlapping sound events is still unclear. In addition, to the best of our knowledge, multi-channel and spatial features have not been studied extensively for moving sources. To bridge the gap, we investigate the effectiveness of spatial features in dynamic-source scenarios. We experiment with a state-of-the-art CRNN model for SED and several input features such as single-channel and multi-channel log-mel spectrograms, GCC-PHAT, and intensity vector. We use DCASE2019 and DCASE2020 sound event detection and localization (SELD) datasets [16, 17],

which are simulated with static and dynamic sources, respectively. Experimental results on these two datasets show that multi-channel log-mel spectrograms and spatial features hinder the SED performance in dynamic-source cases. The best use of multi-channel audio inputs for polyphonic SED in both static and dynamic scenarios is build a model that inputs single-channel log-mel spectrograms. During training, single-channel log-mel spectrograms of different the channels of the same audio input are treated as different training samples. During inference, the trained model makes predictions for all of the single-channel log-mel spectrograms of different channels and the final prediction is obtained by taking the arithmetic mean of these output predictions.

The rest of our paper is organized as follows. Section II describes several input features for SED. Section III presents a state-of-the-art SED network. Section IV shows the experimental results and discussions. Finally, we conclude the paper in Section V.

2. INPUT FEATURES FOR SOUND EVENT DETECTION

In this section, we briefly describe several common multi-channel and spatial features for SED.

2.1. Multi-channel log-mel spectrogram

To extract log-mel spectrograms, time-domain audio input signals are first transformed into short-time Fourier Transform (STFT) domain. The complex spectrum is then converted to power spectrogram by applying absolute operator followed by power-of-2 operator. The power spectrogram is multiplied with a mel filter bank, and converted to a logarithmic scale. The same procedure is done for all the audio input channels. The multi-channel log-mel spectrograms are stacked along the channel dimension to form 3D input features. The dimensions of the multi-channel log-mel spectrogram are $M \times n\_frames \times n\_mels$, where M is the number of input channels, n_frames is the number of time frames, and n_mels is the number of mel filter bands. The dimensions of the single-channel log-mel spectrogram are $1 \times n\_frames \times n\_mels$.

2.2. GCC-PHAT

GCC-PHAT is shown to improve the performance of the SED task when used in conjunction with multi-channel log-mel spectrogram [11, 12]. GCC-PHAT is computed for each audio frame for all the microphone pairs. The maximum time lag of the GCC-PHAT spectrum is $f_s d_{max}/c$, where f_s is the sampling rate, d_{max} is the largest distances between two microphones, and c is the speed of sound.

The number of time lags to be included in the GCC-PHAT spectrum is selected to be equal to the number of mel filters so that the GCC-PHAT features can be stacked with the multi-channel log-mel spectrograms along the channel dimension. The dimensions of the GCC-PHAT feature are $M(M-1)/2 \times n\_frames \times n\_mels$.

2.3. Intensity vector

While GCC-PHAT features are extracted from microphone-array (mic-array) format of the multi-channel audio signals, intensity vector are extracted from first order ambisonic (FOA) format. The 4 channels of FOA format consist of omni-directional, x-directional, y-directional, and z-directional components, respectively. The magnitude differences between the x, y, z and the omni-directional components indicate the directions-of-arrival (DOAs) of sound sources [18]. The intensity vector expresses these magnitude differences and thus carries the DOA information. The intensity vector in each x, y, z direction can be computed in STFT domain as the real component of a product between the signal in each direction and the conjugate of the omni-directional signal. The intensity vector is normalized such that it has unit norm [13]. In order to combine intensity vector and multi-channel log-mel spectrograms, the intensity vectors are passed to the same set of mel filters that are used to compute log-mel spectrograms. The dimensions of the intensity vector feature are $3 \times n\_frames \times n\_mels$.

2.4. Feature normalization

All the single-channel, multi-channel log-mel spectrograms, GCC-PHAT and intensity vector features are normalized to have zero mean and unit variance along the mel-feature dimension for each channel separately.

3. SOUND EVENT DETECTION NETWORK

We use a state-of-the-art CRNN-based SED network that was proposed by Cao *et al* [12]. The SED network consists of 8 CNN layers, 1 bidirectional GRU layer, and 1 FC layer as shown in Table 1. Average pooling is used to reduce the feature map's size in both time and feature dimension after each convolutional block. Depending on the chosen input frame rate and the given label frame rate of each dataset, a upsampling layer is used after the classifier to match the label frame rate.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1. Datasets

We use two public datasets for our experiments. The DCASE2019 SELD dataset is simulated for static-source cases [16]. The dataset consists of 400 and 100 one-minute audio clips for development and evaluation, respectively.

Table 1: A CRNN-based SED network

Stage	Layer description
conv1	(conv2d 64 3x3, BN, ReLu) x 2, 2x2 average pooling
conv2	(conv2d 128 3x3, BN, ReLu) x 2, 2x2 average pooling
conv3	(conv2d 256 3x3, BN, ReLu) x 2, 2x2 average pooling
conv4	(conv2d 512 3x3, BN, ReLu) x 2, 2x2 average pooling
pooling	average pooling along mel-feature dimension
GRU	bidirectional GRU 256
FC	FC ($n\_classes$), sigmoid
upsample	$factor = label\_frame\_rate \times 16 / feature\_frame\_rate$

Table 2: Hyper-parameters for datasets

Parameters	2019 SELD dataset	2020 SELD dataset
fs	32000 Hz	24000 Hz
nfft	1024	1024
hopsize	320	300
window length	1024	1024
window	Hann	Hann
mel filters	96	96
label frame rate	50 Hz	10 Hz
upsampling factor	8	2

There are 11 sound classes. The azimuth and elevation ranges are $[-180°, 180°)$ and $[-40°, 40°]$, respectively with an angular resolution of $10°$. We use 400 development clips for training and 100 evaluation clips for testing.

The DCASE2020 SELD dataset [17] is simulated for both static and dynamic-source cases. The DCASE2020 development dataset consists of 400, 100, and 100 one-minutes audio clips for training, validation, and testing, respectively. There are 14 sound classes. The azimuth and elevation ranges are $[-180°, 180°)$ and $[-45°, 45°]$, respectively with an angular resolution of $1°$.

We chose these two datasets because they provide both mic-array and FOA format. The number of channels of both formats is 4. We ignore the DOA labels in both datasets and only use the SED labels.

4.2. Evaluation metrics

The SED task is often evaluated using the segment-based F1 score and error rate (ER) [19]. The commonly-used segment length is 1 second. The error rate computes the ratio between the number of the substitution, deletion, and insertion errors that a model makes and the total number of ground-truth events. The F1 score averages the precision and recall rate. Following the convention of DCASE SELD challenge, we combine error rate and F1 score into one single metric to better compare the performances of different models. The single SED error metric is defined as $SED\_error = (ER + (1 - F1))/2$. The lower the SED error is, the better the model's performance is.

4.3. Hyper-parameters and training procedure

Table 2 shows the hyper-parameters for the two datasets. We use Adam optimizer to train all the SED models for 50 epochs with a learning rate set to 0.001 for the first 30 epochs and reduced by 10% for each subsequent epoch until it reach 0.0001. For the 2019 dataset, we train all the models for a fixed 50 epochs. For the 2020 dataset, we use the validation set to select the best epoch.

4.4. Models for comparison

We use the same network architecture as shown in Section 3 to train all the SED models with different input features. Table 3 shows the models' names and their input descriptions. In the column *Input features for training*, *4 single-channel log-mel* input means all 4 single-channel log-mel spectrograms are used to train the model. For these models, since the amount of training data increases 4 folds, we reduce the number of training epoch by 4 from 50 to 12. In the column *Input features for testing*, *4 single-channel log-mel* input means the 4 single-channel log-mel spectrogram are passed to the model separately to produce 4 output predictions. The final output is the arithmetic mean of these 4 outputs. This is a form of output ensemble using one trained model.

4.5. SED experimental results

The SED experiment results using DCASE2019 and DCASE2020 SELD datasets are shown in Table 4 and Table 5, respectively. Overall, the SED performances of all the models using the DCASE2019 dataset are much higher than the SED performance of those using DCASE2020 dataset. The main reason is that the 2020 dataset is much more challenging than the 2019 dataset. The intra-class variance in the 2020 dataset is large and there are many difficult classes to distinguish such as *female-scream* and *male-scream*, *alarm* and *phone*. In addition, there are more variances in the event durations and background noises in the 2020 dataset than those in the 2019 dataset.

The experimental results show that the combination of multi-channel log-mel spectrograms and intensity vector does not improve the SED performance for both static and dynamic sound sources. One of the reasons might be that the DOA information are encoded as magnitude of omni, x, y, and z channels so the spectral contents are distorted in the x, y, and z channels.

The SED model that uses a combination of multi-channel log-mel spectrograms and GCC-PHAT *m-logmel-gcc* similar to [12] is the best model for mic-array format using the 2019 dataset. However, it is the worse model using the 2020 dataset. It seems that GCC-PHAT features reduces SED performance in dynamic-source scenarios.

Table 3: SED models with different types of input features

Model name	Number of input channels	Input features for training	Input features for testing	Post-processing
s-logmel-11	1	first-channel log-mel	first-channel log-mel	None
s-logmel-14	1	first-channel log-mel	4 single-channel log-mel	Average
s-logmel-41	1	4 single-channel log-mel	first-channel log-mel	None
s-logmel-44	1	4 single-channel log-mel	4 single-channel log-mel	Average
m-logmel	4	multi-channel log-mel	multi-channel log-mel	None
m-logmel-gcc	10	multi-channel logmel & GCC-PHAT	multi-channel logmel & GCC-PHAT	None
m-logmel-iv	7	multi-channel logmel & intensity vector	multi-channel logmel & intensity vector	None

Table 4: SED results using DCASE2019 dataset (static sources)

	FOA format			Mic-array format		
Model	ER	F1	SED error	ER	F1	SED error
s-logmel-11	0.117	0.935	0.091	0.136	0.926	0.105
s-logmel-14	0.124	0.933	0.096	0.128	**0.932**	**0.098**
s-logmel-41	0.113	0.939	0.087	0.133	0.927	0.103
s-logmel-44	**0.108**	**0.942**	**0.083**	0.129	0.930	0.100
m-logmel	0.110	0.940	0.085	**0.126**	0.930	**0.098**
m-logmel-gcc	-	-	-	**0.126**	0.930	**0.098**
m-logmel-iv	0.131	0.930	0.101	-	-	-

Table 5: SED results using DCASE2020 dataset (static and dynamic sources)

	FOA format			Mic-array format		
Model	ER	F1	SED error	ER	F1	SED error
s-logmel-11	0.373	0.745	0.314	0.373	0.749	0.312
s-logmel-14	0.355	0.761	0.297	0.352	**0.765**	**0.293**
s-logmel-41	0.362	0.755	0.303	0.365	0.744	0.311
s-logmel-44	**0.348**	**0.766**	**0.291**	**0.346**	0.756	0.295
m-logmel	0.395	0.733	0.331	0.380	0.728	0.326
m-logmel-gcc	-	-	-	0.428	0.697	0.365
m-logmel-iv	0.408	0.715	0.346	-	-	-

Among all the multi-channel features, multi-channel log-mel spectrograms achieve the best performance. Models using multi-channel log-mel spectrograms *m-logmel* ranks best and second best for mic-array and FOA format, respectively using the 2019 dataset. However, *m-logmel* is the second worst model for both audio formats using the 2020 dataset, just behind *m-logmel-iv* and *m-logmel-gcc*. We hypothesize that for moving sources, the multi-channel logmel and GCC-PHAT spatial features have high variation. Therefore these SED models over-fit the 2020 dataset which is relatively small.

Among all the single-channel models, the models that average the outputs of 4 channels (*s-logmel-14*, *s-logmel-44*) outperform the models that only process the first channel (*s-logmel-11*, *s-logmel-41*), respectively. This result is expected as output ensembles generally improve the final performance. Similarly, the models that train with all data from 4 input channels (*s-logmel-41*, *s-logmel-44*) often outperform the models that train with only the first channel (*s-logmel-11*, *s-logmel-14*), respectively, except for the case of mic-array format. Generally, audio signals of different microphone channels are slightly different, a model trained with all of the input channels benefits from this data-augmentation. However, there are more magnitude variance across different channels in FOA format compared to mic-array format, where the DOA information is encoded as the phase differences between different channels. As a result, FOA format gains more benefit from training with data from all microphone channels than mic-array format.

For both static and dynamic sources, the best model for FOA format is *s-logmel-44* and the best model of mic-array format is *s-logmel-14*. In static-source scenarios, *m-logmel* and *m-logmel-gcc* also achieves similar performance as *s-logmel-14* for mic-array format.

5. CONCLUSION

In conclusion, for small datasets, multi-channel log-mel spectrograms and spatial features such as GCC-PHAT and intensity vector are not as good as single-channel log-mel spectrogram for SED of dynamic sound sources. On the other hand, multi-channel log-mel spectrogram and GCC-PHAT are more useful for SED of static sound sources.

For both static and dynamic sources case, SED models that use single-channel log-mel spectrogram as input feature, train with data from all input channel, and average the model's outputs of all the channels during inference achieve the best SED performance for FOA format. SED models that use single-channel log-mel spectrogram as input feature, train with data from one input channel, and average the model's outputs of all the channels during inference achieve the best SED performance for mic array format.

These above results are limited for small datasets. However, they have a few important implications for SED applications. First, SELD joint models that use the same input features (multi-channel log-mel spectrograms and spatial features) for both SED and DOA optimization might observe lower SED performance due to these input features are not optimal for SED. Second, for SED applications that require quick inference time, it is sufficient to train a model that use single-channel log-mel spectrogram as input as the gain from multi-channel log-mel spectrograms and spatial features are not conclusive and is outweighed by the computational load required to process a large number of input channels.

6. REFERENCES

[1] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.

[2] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2016, pp. 1–6.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, Jan 2016.

[4] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[5] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[6] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[9] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.

[10] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1038–1047.

[11] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 771–775.

[12] Y. Cao, Q. Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.00268

[13] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," DCASE2019 Challenge, Tech. Rep., June 2019.

[14] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for dnn-based speech recognition in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4380–4384.

[15] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[16] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.08546

[17] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[18] S. Zhao, T. Saluev, and D. L. Jones, "Underdetermined direction of arrival estimation using acoustic vector sensor," *Signal Processing*, vol. 100, pp. 160–168, 2014.

[19] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162