Sound event detection using spatial features and convolutional recurrent neural network

Published in
2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI
10.1109/ICASSP.2017.7952260

SOUND EVENT DETECTION USING SPATIAL FEATURES AND CONVOLUTIONAL RECURRENT NEURAL NETWORK

Sharath Adavanne, Pasi Pertilä, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology

ABSTRACT

This paper proposes to use low-level spatial features extracted from multichannel audio for sound event detection. We extend the convolutional recurrent neural network to handle more than one type of these multichannel features by learning from each of them separately in the initial stages. We show that instead of concatenating the features of each channel into a single feature vector the network learns sound events in multichannel audio better when they are presented as separate layers of a volume. Using the proposed spatial features over monaural features on the same network gives an absolute F-score improvement of 6.1% on the publicly available TUT-SED 2016 dataset and 2.7% on the TUT-SED 2009 dataset that is fifteen times larger.

Index Terms— Sound event detection, multichannel audio, spatial features, convolutional recurrent neural network

1. INTRODUCTION

Sound event detection (SED) task involves recognizing the onset and offset of a sound event in an acoustic scene and further labeling the sound event. The world we live in offers a rich variety of sound events. For example, recognizing environmental sounds [1][2] will give an idea about the local biodiversity. Detecting sound events such as glass breaking and alarm detection can be used for surveillance [3][4]. Furthermore, the detected sound events can be used as a mid-level representation to help retrieval of content based query [5].

Traditionally SED systems have been using monaural audio. Temko et al. [6] proposed to use multichannel audio, and combined classification likelihoods across channels. While the multichannel audio was used, the actual potential of multichannel features was not exploited. Features like time difference of arrival (TDOA) and mel-band energies from the multichannel audio can potentially help the system differentiate the overlapping sound events. Similar multichannel features have been proposed in automatic speech recognition (ASR) [7] and source separation [8]. Just like humans have evolved to exploit the spatial data available at their ears (multichannel) to identify both isolated and polyphonic sound events [9], we can potentially train the SED systems to learn similar spatial information with multichannel data. Recently, such spatial features motivated by the binaural hearing of humans were proposed and shown to be promising for SED task in [10]. Although the features showed improvement over monoaural features, the dataset was too small (around one hour) to conclusively prove the superiority of binaural spatial features (referred as binaural features in future).

In this paper, we propose to use low-level features and compare it with using high-level features. For example, we compare using generalized cross-correlation with phase based weighting ($GCC\text{-}PHAT$) instead of the high-level TDOA feature which is extracted from $GCC\text{-}PHAT$, and show that the network learns powerful representation from just the low-level features. We show that arranging features from each channel as different layers of a multi-layered input volume enables the network to learn the sound events in multichannel audio better than a simple concatenation of the features. We propose to extend the convolutional recurrent neural network (CRNN) to handle more than one feature type and use a bi-directional LSTM. Finally, we evaluate the improvement of using binaural over monaural features on the 19 hours large TUT-SED 2016 dataset.

We present the binaural features used for SED in Section 2, the extended CRNN architecture in Section 3, the experimental set-up and results on two different real-life datasets in Section 4 and our conclusions in Section 5.

2. BINAURAL FEATURES FOR POLYPHONIC SED

Polyphonic SED is the task of recognizing overlapped sound events along with the isolated sound events. The proposed polyphonic SED system has two parts, feature extraction, and a neural network. The neural network described in Section 3 outputs a vector for every sound event class, where each entry in the vector indicates if the sound event was active or not. The feature extraction part extracts the following binaural features at a constant hop length of 20 ms.

2.1. Binaural mel-band energies

Sound sources which have different spatial locations have different intensities in the binaural channels. Furthermore, most overlapping sound events have different frequency spread in

The research leading to these results has received funding from the European Research Council under the European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

the spectrum. The combination of this intensity difference in different bands of frequencies can be exploited to differentiate overlapping sound events. This idea is motivated from the interaural intensity difference (IID) used by humans [9].

Log mel-band energies (referred as mel in future) extracted from both of the binaural channels using 40 mel-bands in 40 ms Hamming window are used as the features. A neural network which is capable of performing linear operations, which includes the difference, can learn to obtain the IID information from these channel-wise energies. By using the channel-wise energies instead of the multichannel energy difference directly, we allow the network to learn other potentially more informative features.

2.2. Time difference of arrival vs cross-correlation

Based on how the sound sources are spatially located with respect to the binaural microphones, they might have different $TDOA$ values. Furthermore, sound events which are overlapping do not always have the same frequency spread in the spectrum. The combination of this $TDOA$ difference in different frequency bands can be exploited by a network to differentiate overlapping sound events. We implemented it by dividing the spectral frame into five mel-bands and calculating the $TDOA$ values in each of the bands. The $TDOA$ is estimated using the $GCC\text{-}PHAT$ [11]. The $GCC\text{-}PHAT$ for each mel-band b is extracted separately:

$$R_b(\Delta_{12}, t) = \sum_{k=0}^{N-1} H_b(k) \frac{X_1(k,t) \cdot X_2^*(k,t)}{|X_1(k,t)||X_2(k,t)|} e^{\frac{i2\pi k \Delta_{12}}{N}}, \quad (1)$$

where, X_1 and X_2 are the FFT coefficients of the two binaural channels. $X_1(k,t)$ specifies the coefficient at time frame t and kth frequency bin, of the total N bins. $H_b(k)$ is the magnitude response of the bth band in B mel-bands and $\Delta_{12} \in [-\tau_{\max}, \tau_{\max}]$, where $\tau_{\max} = 30$ is the maximum sample delay for a sound wave to travel between binaural microphones. Finally, the peak magnitude for each mel-band and time frame is picked in the $GCC\text{-}PHAT$ by $\tau(b,t) = \underset{\Delta_{12}}{\operatorname{argmax}} \{|R_b(\Delta_{12}, t)|\}$.

$TDOA$'s for each band are extracted using multi-resolution windows of 120 ms, 240 ms, and 480 ms to accommodate sound events of variable length. Five $TDOA$ values picked from five bands, for each of the three resolutions, results in 15 $TDOA$ values per time frame.

Neural networks have the potential to learn powerful representations from the raw data. We investigate this by using low-level $GCC\text{-}PHAT$ and comparing it with high-level $TDOA$ feature (which are picked from the $GCC\text{-}PHAT$). $GCC\text{-}PHAT$'s are extracted using Eq. 1 with B set to one. To have a factorizable feature length for max pooling, 60 $GCC\text{-}PHAT$ values are picked in -29 to +30 lag for each of the three multi-resolution (same as $TDOA$), amounting to 180 $GCC\text{-}PHAT$ values per time frame. By using $GCC\text{-}PHAT$ instead of $TDOA$, we take the data-oriented approach and get rid of empirical limitations and let the network learn the representation best suited for the problem.

2.3. Dominant frequencies vs auto-correlation

In [10], it was shown that the three most dominant frequencies and their magnitudes (referred as $dom\text{-}freq$ in future) helped in the SED task. This was motivated by the idea that overlapping sound events do not always have the same dominant frequencies, and the network can learn to differentiate these overlapped events using the dominant frequencies. The $dom\text{-}freq$ values were picked from thresholded parabolically-interpolated STFT [12] in the 100 to 4000 Hz range from each of the binaural channels in frames of 40 ms. We continue to use this feature in this paper.

The pitch is a perceptual feature which human listeners have been using to recognize overlapping sound events [13]. One of the prominent way to estimate pitch values are from the auto-correlation (ACR). In the presented work, ACR is calculated on the binaural channels by time domain auto-correlation in 40 ms windows and choosing 400 correlation values in the range of 107.5 Hz to 4410 Hz. This was selected to be close to the $dom\text{-}freq$ extraction range and the number of correlation values easily factorizable during max pooling.

3. CONVOLUTIONAL RECURRENT NEURAL NETWORK

The best results to date in polyphonic SED was reported in [14], where an architecture exploiting the combined modeling capacities of a convolutional neural network (CNN), recurrent neural network (RNN) and fully connected (FC) layer termed as the convolutional recurrent neural network (CRNN) was proposed. We use this CRNN network and extend it for multichannel audio features.

Features from each channel of the multichannel features are layered one over the other to form a volume. More concretely, M frames of a feature, each of length L, from two channels are layered into a $M \times L \times 2$ volume. On slicing such a volume along a particular time frame, we get all the multichannel features corresponding to that time frame. The two-dimensional CNN's by design are built to learn on such volumes, i.e., it initially learns channel-wise filter weights, and further builds an activation map that is obtained as a combination of these channel-wise filter weights, which serves as the inter-channel information. This way we enable the CNN layers in the initial stages of the CRNN network to learn inter-channel information from multichannel features. We report the improvement in performance of using such a volume input over simple multichannel feature concatenation ($M \times 2L$) in Section 4.4.

Separate volumes of each of the multichannel features are created. T time frames of 40 mel features from the two binaural channels are layered into one volume of size $T \times 40 \times 2$. When using $dom\text{-}freq$, dominant frequencies and their magnitudes are treated as different features, and since their feature lengths are the same (3) we layer them in $T \times 3 \times 4$. For ACR we layer the 400 correlation values of each channel into a $T \times 400 \times 2$ volume. Similarly, the three multi-resolution

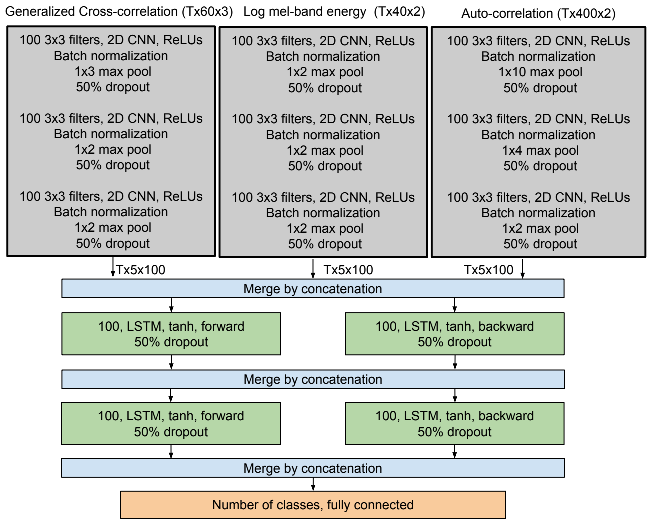

Fig. 1. Convolutional bi-directional recurrent neural network (CBRNN) architecture for multichannel audio features

$TDOA$ features are layered to $T \times 5 \times 3$ and the 60 values of GCC-$PHAT$ are layered to $T \times 60 \times 3$.

Separate CNN's are used to learn local shift-invariant features in each of these volumes as shown in Figure 1. Since the dimensions of mel, GCC-$PHAT$, and ACR are high, we use three CNN layers followed by max pooling to reduce the final feature map dimension to $T \times 5 \times 100$. When using $TDOA$ and dom-$freq$ features, a single 100-filter CNN layer is used without max pooling. To keep the time information intact for final sound event onset and offset detection, we do not apply max pooling in time (T) axis. Post CNNs, the feature maps are merged using concatenation and fed to two consecutive bi-directional long short term memory (LSTM). The output layer is a fully-connected time distributed layer which has as many units as the number of classes in the dataset. A sigmoid activation function is used at the output layer to allow several classes to be predicted as active simultaneously. We refer to this as the CBRNN system in future.

Batch normalization [15] is used in all the CNN layers. A 50% dropout [16] is utilized in all CNNs and LSTMs to avoid over-fitting of the network. The combined architecture was trained by backpropagation through time [17] using Adam optimizer [18] and binary cross-entropy objective. Early stopping was used to reduce overfitting if the F-score (Section 4.2) did not change for 50 epochs. A sequence length of 100 frames (2 seconds) and a batch size of 32 was chosen after calibrating. At test time the sigmoid layer outputs are thresholded with a fixed value of 0.5.

4. EVALUATION AND RESULTS

4.1. Datasets

The proposed SED system is evaluated on two real-life datasets -TUT Sound Events 2009 (TUT-SED 2009) [19] and TUT Sound Events 2016 Development set (TUT-SED 2016) [20]. Both datasets have been recorded using in-ear microphones. TUT-SED 2009 has been used for SED in monaural context [14], but no previous work has reported using the binaural recordings on this dataset. TUT-SED 2016 was published as part of the DCASE 2016 challenge [21], to allow public benchmarking. TUT-SED 2009 is fifteen times larger than TUT-SED 2016, by showing considerable improvement on TUT-SED 2009 we can conclusively say the proposed system is learning and exploiting spatial information.

All the work proposed in this paper is done in a context-independent manner, i.e., we train a single system to learn sound event classes across contexts.

The first dataset - TUT-SED 2009 consists of 103 binaural recordings from 10 different contexts (listed in Table 2). Each context consists of 8 to 14 recordings which vary from 10 to 30 minutes, amounting to an overall length of 1133 minutes. The recordings have been manually annotated, and the annotated events have been grouped into 61 event classes [19]. Each context has 9-16 event classes, while some events occur in multiple contexts, some are context specific. The dataset defines five-folds for training, validation, and testing.

The second dataset - TUT-SED 2016 consists of 22 binaural recordings for two contexts - home and residential area, amounting to 78 minutes. The home context has ten recordings with 11 sound event classes, and the residential area has 12 recordings with seven sound event classes [20]. The dataset defines four-folds for training and testing. We use 20% of the training data for validation, and the same validation is used for all our evaluations.

4.2. Metrics

The SED system output is evaluated with the reference in fixed length intervals, also called as segment-based evaluation [22]. For each segment k, the following are calculated (i) true positive ($TP(k)$): total number of events active in both reference and system output segment. (ii) False positive ($FP(k)$): total number of events active in system output segment but not in reference. (iii) False negative ($FN(k)$): total number of events active in reference segment but not in system output. The first metric, F-score is then calculated as,

$$F = \frac{2 \cdot \sum_{k=1}^{K} TP(k)}{2 \cdot \sum_{k=1}^{K} TP(k) + \sum_{k=1}^{K} FP(k) + \sum_{k=1}^{K} FN(k)} \quad (2)$$

The second metric, error rate (ER) evaluates the system output based on the number of insertions (I), deletions (D) and substitutions (S).

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \quad (3)$$

Where $N(k)$ is the number of sound events marked as active in the reference segment k, and

$$S(k) = \min(FN(k), FP(k)) \quad (4)$$

$$D(k) = \max(0, FN(k) - FP(k)) \quad (5)$$

$$I(k) = \max(0, FP(k) - FN(k)) \quad (6)$$

We use a segment length of one second for ER and F-score estimation. The evaluation metrics are calculated for each context separately and averaged result is presented.

4.3. Baseline

The proposed CBRNN architecture with binaural features is compared with the state of the art monaural SED system introduced in [14]. The system used 40 monaural log mel-band energies ($mel\text{-}monaural$) as features. The network had three CNN's each of 96 filters, followed by max pooling in frequency axis reducing the dimension to one. The feature map from CNN was then fed to three LSTMs with 256 units each. The output was a fully-connected layer with units equal to the number of classes in the dataset.

4.4. Results

Table 1 shows the metrics for multi-layered input of the binaural log mel-band energy features (mel) and concatenating it ($mel\text{-}concat$) for TUT-SED 2009 dataset. Using a multi-layered input is seen to perform relatively better than a simple concatenation. Similar improvement was observed using multi-layered input of $TDOA$, $dom\text{-}freq$, $GCC\text{-}PHAT$ and ACR (not tabulated).

From Table 1 we see that using binaural features improves both the ER and F-scores over monaural features ($mel\text{-}monaural$) across datasets. While the $dom\text{-}freq$ and mel feature combination gave the best performance in TUT-SED 2009, $TDOA$ and mel performed the best for TUT-SED 2016. In numbers, using binaural over monaural features on the same network gives an absolute F-score improvement of 2.7% for TUT-SED 2009 and 6.1% for TUT-SED 2016. By showing this improvement on a larger dataset like TUT-SED 2009, we can more confidently say that the network is truly learning the binaural information.

From the metrics in Table 1 and 2 we see that the performance of using $GCC\text{-}PHAT$ instead of $TDOA$ or ACR instead of $dom\text{-}freq$, is comparable. This is a significant result, showing that the network can learn equivalent information of powerful high-level features from just the low-level features. Thereby making the features dataset independent and relieving the tuning of parameters like the number of $dom\text{-}freq$ and $TDOA$ values.

Most of the sound event classes were seen to be recognized better with the binaural features. Since we cannot present all the 79 classes of the two datasets in this paper, we show the context based F-scores for TUT-SED 2009 dataset in Table 2. A general observation is that the $dom\text{-}freq$ / ACR and mel are useful for indoor and sound intense environment (bus, hallway, office, and basketball), while $TDOA$ / $GCC\text{-}PHAT$ and mel are seen to help in outdoor contexts (beach and street). This also explains why $dom\text{-}freq$ and mel gave better results for TUT-SED 2009. While TUT-SED 2016 had one each of indoor and outdoor contexts, TUT-SED 2009 had more indoor contexts than outdoor.

The proposed CBRNN architecture using the same $mel\text{-}monaural$ feature used in CRNN-baseline achieved an F-score of 68.0% for TUT-SED 2009 and 29.7% for TUT-SED 2016 (Table 1). The difference in the scores with respect to CRNN-baseline can be associated with using a higher dimensional input to LSTM's in the proposed CBRNN.

Feature combination	TUT-SED 2009		TUT-SED 2016	
	ER	F	ER	F
CRNN baseline [14]	0.49	68.8	**0.93**	31.3
$mel\text{-}monaural$	0.49	68.0	1.03	29.7
$mel\text{-}concat$	0.44	70.3		
mel	**0.43**	71.1	0.99	32.3
$mel + TDOA$	0.45	70.9	0.95	**35.8**
$mel + GCC\text{-}PHAT$	0.44	71.1	0.95	34.6
$mel + dom\text{-}freq$	**0.43**	**71.7**	0.98	32.8
$mel + ACR$	0.44	71.2	0.98	33.8
$mel + TDOA + dom\text{-}freq$	0.44	71.0	1.01	33.3
$mel + GCC\text{-}PHAT + ACR$	0.45	70.9	0.99	33.6

Table 1. Error rate (ER) and F-score achieved using binaural features and CBRNN on TUT-SED 2009 and 2016 datasets.

5. CONCLUSION

In this paper, we extended convolutional recurrent neural networks to handle multiple feature classes and process feature-maps using bi-directional LSTM's. A multi-layered input of multichannel features which enables the network to learn sound events in a multichannel audio better was proposed. Low-level features were used in place of high-level features, and the network was shown to learn high-level equivalent information from simple low-level features. The performance of the system was evaluated on two datasets - a larger dataset for proving that the binaural features truly help in improving the sound event detection, and a public dataset, to allow other researchers to benchmark. The proposed network using binaural spatial features was shown to recognize sound events better than using just the monaural features.

Feature combination	Indoor							Outdoor		
	Basketball	Bus	Hallway	Office	Car	Restaurant	Shop	Beach	Street	Track and Field
$mel\text{-}monaural$	79.7	52.6	59.1	81.8	78.2	80.7	62.4	56.5	60.3	70.1
mel	82.2	56.5	66.6	83.3	81.5	83.1	63.3	59.5	66.0	70.2
$mel + TDOA$	82.8	58.7	66.0	80.8	79.2	81.2	64.7	**60.9**	**66.9**	68.8
$mel + GCC\text{-}PHAT$	81.9	58.9	65.3	80.0	81.2	81.3	65.3	60.4	66.3	**72.6**
$mel + dom\text{-}freq$	**83.7**	**60.5**	**67.8**	**84.6**	80.8	81.8	64.6	60.7	66.6	67.6
$mel + ACR$	82.9	58.6	63.8	83.6	**83.4**	**82.3**	**65.5**	60.4	65.8	69.0
$mel + TDOA + dom\text{-}freq$	83.0	59.4	67.5	83.9	78.6	79.9	65.1	60.5	65.0	70.0
$mel + GCC\text{-}PHAT + ACR$	82.8	59.1	66.8	82.2	79.4	80.4	64.8	60.4	66.1	68.5

Table 2. Context wise F-scores for TUT-SED 2009 dataset.

6. REFERENCES

[1] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015.

[2] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," in *ACM Computing Surveys (CSUR)*, 2016.

[4] 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.

[5] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2008.

[6] A. Temko, C. Nadeu, and J. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in *Springer-Verlag, Berlin*, 2008.

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

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[9] J. W. Strutt, "On our perception of sound direction," in *Philosophical Magazine*, 1907.

[10] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[11] C. Knapp and C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.

[12] J. O. Smith, *Sinusoidal Peak Interpolation, in Spectral Audio Signal Processing*, accessed 23.06.2016, online book, 2011 edition. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/Sinusoidal_Peak_Interpolation.htm

[13] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," in *MIT Press*, 1990.

[14] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM TASLP Special Issue on Sound Scene and Event Analysis*, 2017, accepted for publication.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," in *Journal of Machine Learning Research (JMLR)*, 2014.

[17] P. J. Werbos, "Backpropagation through time: what it does and how to do it," in *Proceedings of the IEEE*, 1990.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv:1412.6980 [cs.LG]*, 2014.

[19] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[20] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *European Signal Processing Conference (EUSIPCO)*, 2016.

[21] "Detection and classification of acoustic scenes and events (DCASE)," 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

[22] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, 2016.