A REPORT ON SOUND EVENT DETECTION WITH DIFFERENT BINAURAL FEATURES

Sharath Adavanne, Tuomas Virtanen

Department of Signal Processing , Tampere University of Technology

ABSTRACT

In this paper, we compare the performance of using binaural audio features in place of single channel features for sound event detection. Three different binaural features are studied and evaluated on the publicly available TUT Sound Events 2017 dataset of length 70 minutes. Sound event detection is performed separately with single channel and binaural features using stacked convolutional and recurrent neural network and the evaluation is reported using standard metrics of error rate and F-score. The studied binaural features are seen to consistently perform equal to or better than the single-channel features with respect to error rate metric.

Index Terms—

Polyphonic sound event detection, binaural, monochannel, convolutional recurrent neural network

1. INTRODUCTION

Sound event detection (SED) is the task of recognizing the sound events and their respective temporal start and end time in a recording. Sound events in real life do not always occur in isolation, but tend to considerably overlap with each other. Recognizing such overlapping sound events is referred as polyphonic SED. Applications of such polyphonic SED are numerous. Recognizing sound events like alarm and glass breaking can be used for surveillance [1, 2]. Environmental sound event detection can be used for monitoring biodiversity studies [3, 4, 5]. Further, SED can be used for automatically annotating audio datasets, and the sound events recognized can be used as a query for retrieval.

Polyphonic SED using monochannel audio has been studied extensively. Different approaches have been proposed using supervised classifiers like Gaussian mixture model - hidden Markov model [6], fully-connected networks [7], convolutional neural networks (CNN) [8, 9] and recurrent neural networks (RNN) [10, 11]. More recently, the state of the art method for polyphonic SED was proposed in [12] and evaluated on multiple private and publicly available datasets. They used log mel-band energies along with a convolutional recurrent neural network (CRNN) architecture as their method.

Recognizing overlapping sound events using monochannel audio is a difficult task. These overlapping sound events can potentially be recognized better with multichannel audio. One of the first methods to use multichannel audio for SED was [13]. They performed SED on each of the monochannel audio and the combined likelihoods across channels was used for the final prediction. More recently [14] extended the state of the art CRNN network of [12] for multichannel features and multiple feature classes and showed that using binaural instead of monochannel recordings of the same datasets used in [12] improved the SED performance. Binaural features exploiting the inter-aural intensity and time differences were used in this method. These initial results on binaural audio motivates us to further explore polyphonic SED using binaural audio.

In this paper, we explore and study the performance of three different binaural features - a) log mel band energy, b) log mel band energy extracted in three different resolution windows and c) magnitude and phase component of short-term Fourier transform, all extracted in both the channels of the binaural audio. While a) has been used in [11, 14], b) and c) has not been used in polyphonic SED task previously. We compare the performance of SED amongst the binaural features and also compare with the single channel log mel-band energy feature. We separately train the multichannel network method [14] with features extracted from the publicly available TUT Sound Events 2017 dataset and present the results.

The feature extraction and neural network used is described in section 2. The dataset creation, evaluation metrics and procedure are explained in section 3. Finally, the results and discussion are presented in section 4.

2. METHOD

The input to the method is an audio signal. Features are extracted in consecutive time windows from each channel of the audio. These audio features are fed to a multichannel convolutional and recurrent neural network architecture, which maps the audio features to the sound event labels in the dataset. The output of the neural network is in the range of [0, 1] for each of the sound event label, where one refers to the sound event being active, and zero for absence. The detailed description of feature extraction and the neural network is presented below.

2.1. Feature extraction

In this paper, we study the performance of three binaural audio features and compare it with single channel audio feature. All features are extracted in hop length of 20 ms to keep the number of frames same.

2.1.1. Single channel feature

Log mel-band energy (mbe) has been used extensively for the SED task [10, 12, 11, 14] we continue to use the feature in this paper. mbe is extracted in Hamming window of length 40 ms. We use 40 mel-bands in the frequency range of 0-22500 Hz. For a given audio input of F frames, this feature extraction block results in a $F \times 40$ output.

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.

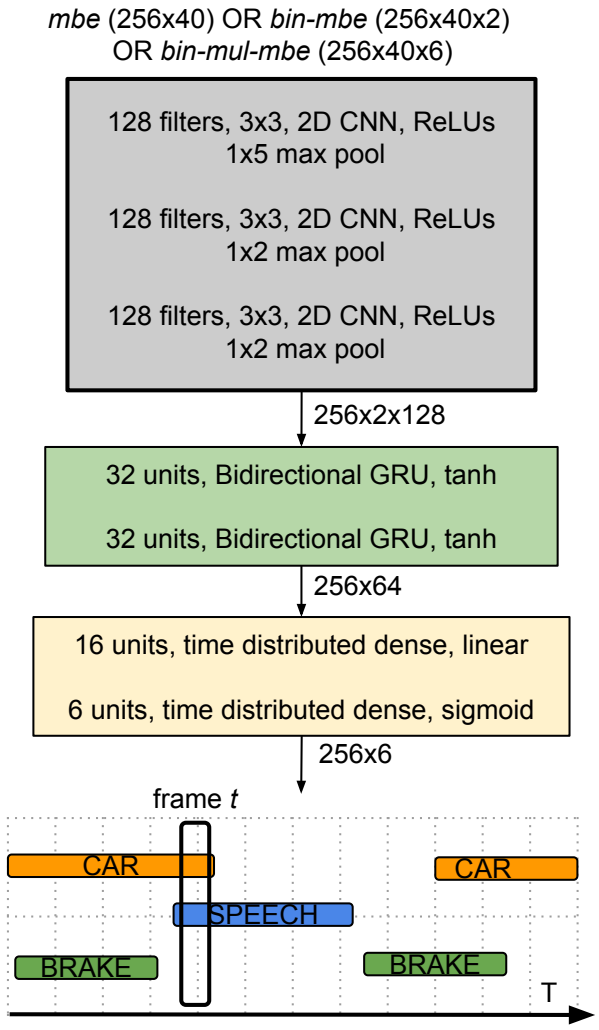

Figure 1: Stacked convolutional and recurrent neural network for binaural polyphonic sound event detection.

2.1.2. *Binaural features*

The first binaural feature we study is the binaural log mel-band energy (bin-mbe) from the works of [11], where it was shown to perform better than the mbe. We extract bin-mbe in a similar fashion as mbe on each of the binaural channels resulting in a $F \times 80$ (40*2 = 80) output.

Su et al. in [15] reported that mbe extracted in multi-resolution windows give considerable improvement for SED over using just the single resolution mbe. Motivated by this we extend it to binaural scenario, and extract it in both the channels of audio (bin-mul-mbe). Specifically, we use three different window sizes 1024, 4096, and 16384 as in the paper [15] and extract mbe feature in each of the windows and each of the binaural channels. This feature extraction block results in a $F \times 240$ (40*3*2 = 240) output.

Recently, it was shown that the neural networks can estimate the direction of arrival from just the phase components of the multi-channel short-term Fourier transform (STFT) coefficients [16]. Motivated by this, we extend it to binaural channels by extracting STFT in each of the binaural channels and propose to also use the magnitude component along with the phase component (bin-fft). We extract STFT in windows of 40 ms using 2048 points, post which we calculate the magnitude and the phase component resulting in a $F \times 4096$ (1024*2*2 = 4096) output.

2.2. Neural network

The input to the neural network in the case of single channel audio is T consecutive time frames of mbe, with a dimension of $T \times 40$ as shown in Figure 1. In the case of binaural audio features, we stack each of the channel features separately. Specifically, for bin-mbe we stack the $T \times 80$ output of feature extraction block to a volume of dimension $T \times 40 \times 2$. In case of bin-mul-mbe and bin-fft along with the channels, we also stack the multi-resolution windows, phase and magnitude components separately, resulting in volumes of dimension $T \times 40 \times 6$ and $T \times 1024 \times 4$ respectively. Based on the task of mono or binaural SED, the network is fed with the respective feature sequence. In this paper, we use a sequence length of T = 256 for all the features.

We use convolutional neural network (CNN) as our initial layers to learn local shift-invariant patterns from audio feature. The receptive filters of these CNNs are of the size 3×3 size. The output activation from the CNN layers are padded with zeros to keep the dimension of the output the same as input. Batch normalization [17] and max-pooling is performed after every layer of CNN to reduce the final dimension to $T \times 2 \times N$, where N is the number of filters in the final layer of CNN. We perform max-pooling in the frequency axis only, this is done to preserve the time resolution of the input. The CNN layer activation is further fed to layers of bi-directional gated recurrent units (GRU), to learn long term temporal activity patterns. This is followed by layers of time-distributed fully-connected (dense) layers. The time resolution remains the same as input feature in both GRU and dense layers. The final prediction layer has an output dimension of $T \times C$, where C is the number of classes in the dataset. The prediction layer has sigmoid activation in order to be able to produce multi-label output.

The training is performed for 500 epochs using binary cross-entropy loss function and Adam [18] optimizer with a learning rate of 0.0001. Dropout [19] is used as a regularizer after every layer of neural network to make it robust to unseen data. Early stopping is used to stop over-fitting the network to training data. Training is stopped if the error rate (see Section 3.2) on the test split does not improve for 100 epochs. The neural network implementation was done using Keras [20] framework with Theano [21] as backend.

Sound events	Length
brakes squeaking	67.6
car	2541.5
children	346.1
large vehicle	727.0
people speaking	630.6
people walking	1079.2
Total	5391.9

Table 1: Distribution of sound event in the dataset. The length is given in seconds.

3. EVALUATION

3.1. Dataset

We study the performance of the binaural features on the development set of TUT sound events 2017 dataset organized as part of Detection and Classification of Acoustic Scenes and Events [22].

This dataset consists of about 70 minutes of audio data collected in street scenario with annotations of six sound event classes. Table 1 presents the sound event classes and their distribution in the dataset. There are 24 recordings in total, each of about 3-5 minutes, recorded using Soundman OKM II Klassik/studio A3 electret in-ear microphone and a Roland Edirol R-09 wave recorder. The recordings are sampled at 44.1 kHz and 24 bit resolution. The single channel audio for single channel feature study is created by taking the average of the binaural channel audio. The dataset provides four cross validation splits for the above data, with train, validation and test splits.

3.2. Metric

The SED method is evaluated using the polyphonic SED metrics proposed in [23]. Particularly we use segment wise error rate (ER) and F-score calculated in segments of one second length. According to which the F-score is calculated as,

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

where for each one second segment k, $TP(k)$ is the true positives, the number of sound event labels active in both predictions and groundtruth. $FP(k)$ is the false positives, the number of sound event labels active in predictions but inactive in groundtruth. $FN(k)$ is the false negatives, the number of sound event labels active in the ground truth but inactive in the predictions.

The error rate is measured as,

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)}, \quad (2)$$

where, $N(k)$ is the total number of active sound events in the ground truth of segment k. The substitutions ($S(k)$), deletions ($D(k)$) and insertions($I(k)$) are measured using the following equations for each of the K one second segments.

$$S(k) = \min(FN(k), FP(k)) \quad (3)$$

$$D(k) = \max(0, FN(k) - FP(k)) \quad (4)$$

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

For an ideal SED method, ER is zero and F-score is 100.

3.3. Baseline

The baseline method for the dataset used is provided in [22]. This method uses single channel mbe as the audio feature. The network consists of two fully-connected layers with 50 units in each followed by a dropout layer with 0.2 dropout rate. The prediction layer has number of sigmoid units equal to the number of classes in the dataset. The method uses a context of 5 frames resulting in a feature length of 200 (40*5). The network is trained with cross-entropy loss and Adam optimizer for 200 epochs. The evaluation metric scores for this method is reported in Table 2.

3.4. Evaluation procedure

A random hyper-parameter search [24] is performed by varying the number of layers and units of CNN, GRU and dense layers, and the dropout in the set of $\{0.05, 0.25, 0.5, 0.75\}$ for each of the feature. The hyper-parameter tuning was done to achieve the best ER on the test split. The best configuration found was the same for all the mel based features (mbe, bin-mbe, bin-mul-mbe), while the bin-fft stereo was seen to give good results with the same network but larger max-pooling (1×8 after each layer of CNN). The network for mel based features and its configuration is as shown in Figure 1. The dropout rate for the above configuration was 0.5 for mbe and bin-mbe, 0.25 for bin-mul-mbe and 0.05 for bin-fft

We perform SED on the above dataset individually with all the single channel and binaural features and report the average ER and F-scores of five separate runs of four cross-validation provided in the dataset.

	Development		Challenge	
Audio features	ER	F	ER	F
baseline [22]	0.69	56.7	0.94	42.8
mbe	0.55	69.3	**0.79**	41.7
bin-mbe	0.52	69.1	0.80	**42.9**
bin-mul-mbe	**0.50**	**70.3**	0.85	41.4
bin-fft	0.55	66.9	0.87	36.2

Table 2: Best evaluation metric scores achieved with different audio features on the development dataset and the evaluation dataset of DCASE 2017 challenge [22].

4. RESULTS AND DISCUSSION

The evaluation results for SED using single channel and binaural features are presented in Table 2. We see that the stacked convolutional and recurrent neural network with the single channel audio feature (mbe) outperforms the baseline method [22].

Binaural features in general have similar performance as single channel features on the evaluated dataset. In particular the ER of binaural features is seen to be equal or better than the single channel feature. The noteworthy performance is of bin-mul-mbe which is seen to improve the ER considerably over mbe.

The validation and training loss of the network with bin-fft was considerably higher than the other features. Suggesting that the size of the data was possibly less for this feature to find the best weights.

4.1. DCASE 2017 challenge results

The stacked convolutional and recurrent neural network trained respectively with the three binaural and the single channel feature was submitted in the real life sound event detection task of DCASE 2017 challenge [22]. The results obtained on the evaluation data of the challenge is presented in Table 2. All the submitted systems fared well on the evaluation data and resulted in challenging scores. In particular, the network trained on mbe fared as the best method in the challenge followed by bin-mbe in close second among the 34 submitted methods from 14 different teams.

5. CONCLUSION

In this paper, we proposed to study the performance of using different binaural audio features for polyphonic sound event detection. In this regard three binaural features were studied and compared with a baseline single channel audio feature. We performed SED separately for each of the feature using a stacked convolutional and recurrent neural network. The evaluation was carried out on the publicly available TUT Sound Events 2017 dataset. It was observed that using binaural features gave similar or better error rate than single channel features. In particular log mel-band energy feature extracted in different resolution windows was seen to produce the best results for the given dataset.

6. REFERENCES

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

[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] 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.

[4] T. A. Marques *et al.*, "Estimating animal population density using passive acoustics," in *Biological reviews of the Cambridge Philosophical Society*, vol. 88, no. 2, 2012, pp. 287–309.

[5] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," in *Journal of Wildlife Management*, vol. 79, no. 2, 2014, p. 325337.

[6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, 2010.

[7] E. Çakır, 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.

[8] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[9] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *Interspeech*, 2016.

[10] 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.

[11] 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.

[12] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," in *IEEE/ACM TASLP*, volume 25, issue 6, 2017.

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

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

[15] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[16] E. A. P. H. Soumitro Chakrabarty, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *arXiv:1705.00919*, 2017.

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

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

[19] 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.

[20] F. Chollet, "Keras v1.1.2," https://github.com/fchollet/keras, 2015.

[21] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

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

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

[24] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," in *Journal of Machine Learning Research*, 2012.