CLASS WISE DISTANCE BASED ACOUSTIC EVENT DETECTION

Xianjun Xia[1], *Roberto Togneri*[1], *Ferdous Sohel*[2], *David Huang*[1]

[1] School of Electrical, Electronic and Computer Engineering, The University of Western Australia
Xianjun.Xia@research.uwa.edu.au, {Roberto.Togneri, David.Huang}@uwa.edu.au
[2] School of Engineering and Information Technology, Murdoch University
F.Sohel@murdoch.edu.au

ABSTRACT

In this paper, we propose a class wise distance based approach in a neural network based acoustic event detection system. The neural network output probabilities are updated by calculating the distance between the acoustic features of each frame and the class wise distance of each event class. The detected acoustic segments are re-evaluated segmentally using the class wise distances. Cross-validation detection results on the development set of DCASE2017 show the efficiency of the proposed method by achieving a 4% absolute reduction in segment-based error rate compared to the baseline system.

Index Terms— acoustic event detection, class wise distance, re-evaluation

1. INTRODUCTION

Acoustic event detection (AED) has been widely applied in many real world applications, such as surveillance systems [1], siren detection systems [2], chew event detection systems [3] and human-computer interaction [4]. Intra-class variations and the spectral-temporal properties across classes pose great challenges to AED. Due to the varied real world applications of AED and the challenges being faced, some campaigns, such as DCASE [5][6][7] have attempted to capture a wide range of variations in the design of the AED database [8].

Many approaches have been proposed to deal with the challenges that AED systems now face. Gaussian Mixture Models (GMM) based methods are presented in [9] and Hidden Markov Models (HMM) based AED methods have been proposed in [10]. Some other ideas, such as random forest based approaches have been applied in [11][12][13]. The DCASE challenge series has been running since 2013 and many novel ideas have emerged. One popular approach for the AED in DCASE challenge is the neural network based methods. Among the neural network based AED systems, the detection task is considered as a multi-label classification problem. Deep neural network (DNN) based AED systems were developed in [14][15]. In [16], convolutional neural network (CNN) was adopted and a recurrent neural network (RNN) based AED was presented in [17][18]. A global threshold is applied to the neural network output probabilities to determine the active acoustic events at each time index and the post processing strategy is adopted to generate the final acoustic event segments. However, a global threshold applied to all the event classes cannot capture the variable polyphonic levels [19] and cannot represent the characteristics of all the event classes. The insertion error increases and recall accuracy decreases if the global threshold is set too low or too high.

To deal with the different polyphonic levels across time and the different characteristics of the audio signals across classes, this paper presents a method to utilize the class wise distance to make the output probabilities of each frame belonging to different event classes more discriminative. The class wise distance is used to update the neural network output probabilities and to re-evaluate the detected acoustic events segmentally. The neural network output probabilities are then updated by adopting the class wise distance based probability. The class wise distance based probability is represented by calculating the distance between the acoustic features of each frame and the mean acoustic features of each acoustic event class. There are two advantages in adopting the class wise distance based approach. To begin with, this will automatically update the probabilities that each frame belongs to different acoustic event classes using the class wise information, which makes the probabilities more discriminative. Secondly, the class wise distance based approach can be utilized as a strategy to re-evaluate the detected acoustic event segments from an even longer duration by re-evaluating the acoustic events segmentally rather than by frame.

This paper is organized as follows. In Section 2, we briefly introduce the neural network based AED system. Our proposed approaches and algorithms are presented in Section 3. In Section 4, we provided the experimental results and analysis followed by conclusion and future work in Section 5.

2. THE NEURAL NETWORK BASED AED SYSTEM

The neural network based AED system considers the polyphonic acoustic event detection as a multi-label classification problem. Fig. 1 illustrates the task of the polyphonic acoustic event detection. As shown in Fig. 1, each frame may correspond to more than one event label (“car” overlaps with “people speaking”, “large vehicle” and “brakes squeaking” at different time index). The system is composed of three parts: the feature extraction, the neural network training and the post processing.

2.1. Feature extraction

The feature representations adopted by the baseline system are the log mel-band energies. A short-time Fourier transform (STFT) is applied over a 40ms window of the audio and a 50% hop size is used. To capture the time dependency, 5 consecutive frame representations are used to form a 200 dimensional input feature vector. The acoustic features are normalized to a zero mean and unit variance space across the whole database. The implementation

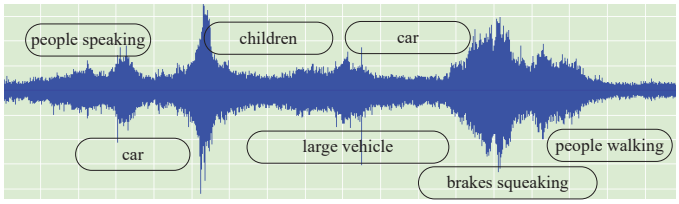

Figure 1: The polyphonic acoustic event detection task.

of the feature extraction is based on speech analysis via librosa[1].

2.2. Neural network training

Two neural network models are used in this paper. One is the multi-layer percepton (MLP) and the other is the CNN. The DCASE2017 challenge organizers provided the MLP based AED system. The multi-layer percepton network includes a fully connected neural network with 2 hidden layers. The number of hidden unit nodes for each layer is 50. The Relu [20] activation function and the Adam [21] optimizer are used to optimize the weights between different layers. The sigmoid output and binary cross entropy loss function are used to train the multi-layer percepton network. Dropout [22] strategy with a value of 0.2 and early stopping criteria which starts after 100 epochs are adopted to overcome the over-fitting problem. The learning rate is set to 0.001.

In this paper, we also extended the MLP based baseline system by substituting the multi-layer percepton neural network with a convolutional neural network. The convolutional neural network includes two convolutional layers, two max-pooling layers, two batch normalization layers, a flattening layer and a sigmoid output layer. The first layer performs a convolution over the input acoustic features with 16 kernels characterized by 3 by 3. The second convolutional layer is the same as the first one except that the number of kernels is set to 32 in order to obtain a higher level representation.

2.3. Post processing

For testing, the trained neural network outputs the probabilities that each frame belongs to each acoustic event class. Then a threshold is adopted to determine whether an acoustic event is active at that frame index.

The median filtering strategy is applied to post process the event activity. The length of the filtering window is set to 0.54s. Afterwards, the minimum event length and the minimum event gap are set to 100ms to determine the final beginning and end times of the detected acoustic events.

3. CLASS WISE DISTANCE BASED AED SYSTEM

To make the output probabilities of each frame belonging to the different event classes more discriminative and the detected events to be re-evaluated from a longer duration level, we propose in this work to incorporate a class wise distance measure to the baseline system. Fig. 2 is the flowchart of the proposed AED system. The class wise distance based probability is used to update the neural network output probabilities and the class wise distance based re-evaluation is to re-evaluate the detected acoustic events segmentally. How the class wise distances, class wise distance based probabilities are calculated and the class wise distance based re-evaluation strategy is applied will be elaborated below.

3.1. Class wise distance calculation

The class wise distances are represented by the normalized mean log-mel energies for each acoustic event class, which can be expressed as:

$$M_d(c) = \frac{\sum_{t=1}^{t=I_c} X_d(t)}{I_c} \quad (1)$$

where c, d, t mean the event class label, the dimension index and the time index respectively. The I_c denotes the number of training samples for the cth event class and $X_d(t)$ is the d-dimensional acoustic representation at the frame index t. The acoustic representations are all normalized by subtracting the global mean and dividing the global standard deviation across the whole database.

Fig. 3 displays the class wise distance between every two

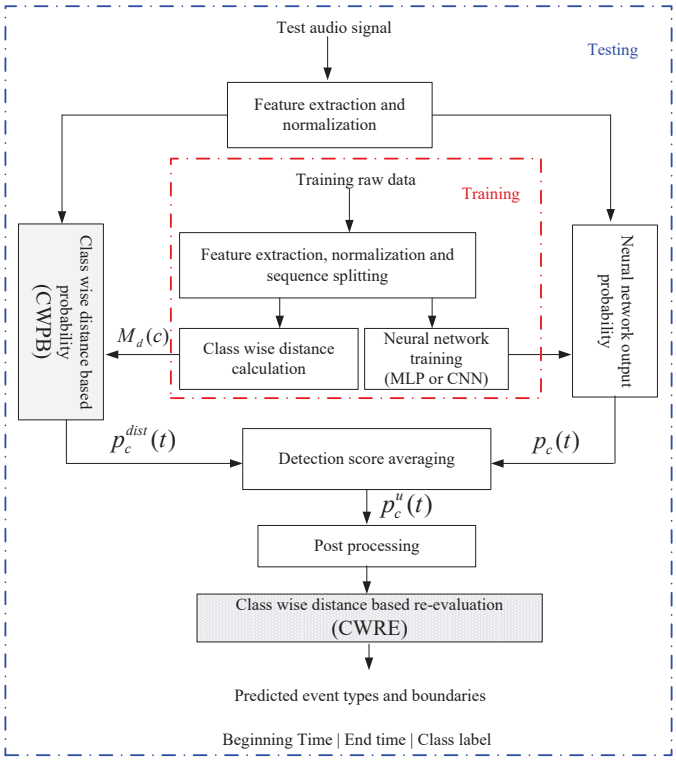

Figure 2: The flowchart of the proposed AED system.

acoustic events. As can be seen in Fig. 3, the class wise mean distance for most of the acoustic events can be discriminated between each other except that the mean contours for the event "car", "brakes squeaking" and "large vehicle" are close and difficult to be differentiated, which is consistent with the real-world human being perceptions. Discriminative class wise distances between the event classes makes it possible to utilize the class wise information to improve the performance of the AED system.

[1] https://github.com/librosa

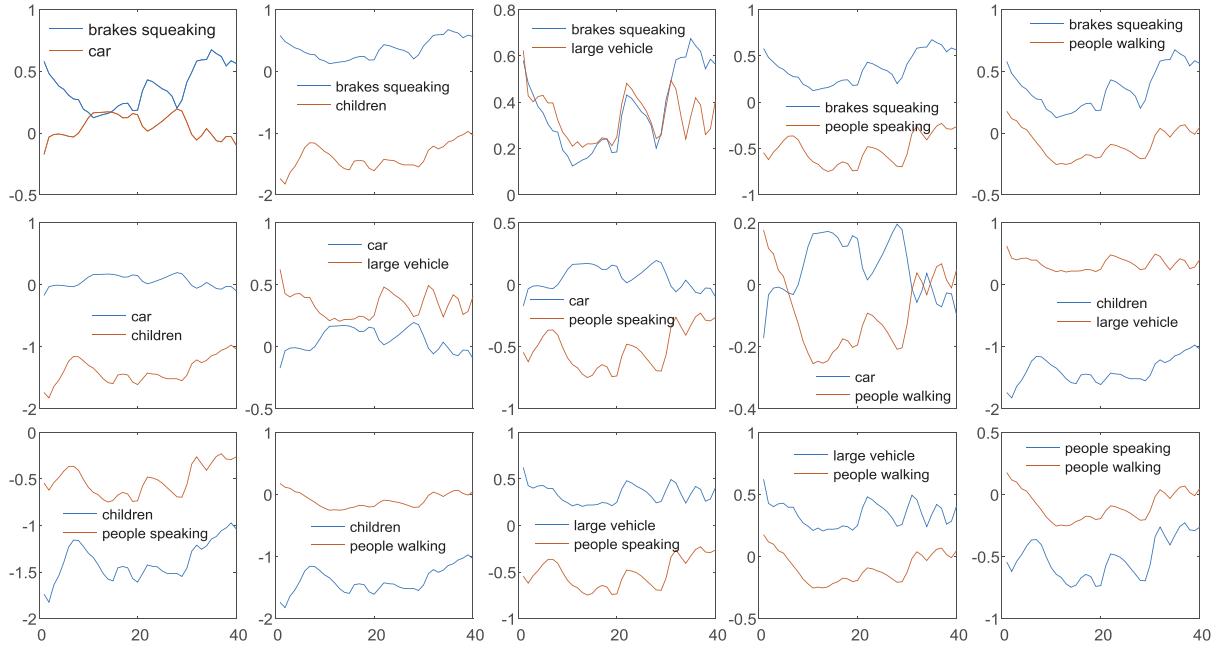

Figure 3: The class wise distances between every two acoustic event classes, where the x-axis is the feature dimension d and y-axis is the normalized mean value $M_d(c)$ for the event class c.

3.2. Class wise distance based probability (CWPB)

After calculating the class wise distance for each event class, the class wise distance based probability $p_c^{dist}(t)$ at frame index t is calculated as follows:

$$p_c^{dist}(t) = e^{-(dist(t,c))^2} \tag{2}$$

$$dist(t,c) = \frac{\sum_{d=1}^{d=D} f(X_d(t), M_d(c))}{D} \tag{3}$$

where $p_c^{dist}(t)$ is the class wise distance based probability at the frame index t for the cth class. The D is the total dimension for the log-mel energies and the f function denotes the Euclidean distance operation.

3.3. Detection score averaging

The class wise information is utilized by averaging the neural network output probabilities and the class wise distance based probabilities:

$$p_c^u(t) = \alpha * p_c(t) + (1-\alpha) * p_c^{dist}(t) \tag{4}$$

The $p_c^u(t)$ is the updated probability to be used to determine the active events and the α is a coefficient which is experimentally set to 0.8 in the proposed AED system. The preset global threshold 0.5 will be applied during post processing to the updated probability $p_c^u(t)$ to determine the active event classes by frame.

3.4. Class wise distance based re-evaluation (CWRE)

After post processing, the detected acoustic events are in segmental format with a beginning and end time (e.g from 1s to 3s, where the detected acoustic event is the event "car"). To further utilize the class-wise distance information, the segmental acoustic events are re-evaluated over a longer duration by calculating the re-evaluation distance $d_{seg,\hat{c}}$ for each test segment:

$$d_{seg,\hat{c}} = -\frac{\sum_{d=1}^{d=D} f(A_{seg,d}, M_d(\hat{c}))}{D} \tag{5}$$

Here, $\hat{c}$, t_b and t_e are the predicted acoustic event class label, detected the beginning and end time respectively. The $A_{seg,d}$ denotes the average log-mel energies for the detected acoustic event segment, which can be expressed as:

$$A_{seg,d} = \frac{\sum_{t=t_b}^{t=t_e} X_d(t)}{t_e - t_b} \tag{6}$$

Properly detected frames of event $\hat{c}$ should exhibit average energies $A_{seg,d}$ close to the training data energies $M_d(c)$ and hence in the proposed system, if the distance $d_{seg,\hat{c}}$ for a detected segment does not rank the top k within all the classes, the detected segment would be discarded. In this paper, the k is experimentally set to 3.

4. EXPERIMENTS

4.1. Database

The TUT sound event 2017 database [23] is used in this paper to evaluate the performance of different systems. The TUT sound event database is partitioned into development and evaluation set. The development set is used to evaluate the performance of different AED systems presented here and the evaluation subset is used for the DCASE2017 challenge (the released evaluation database is without any ground truth). For the development set, a cross-validation setup is provided to uniform the reported results from participants. The detailed description of the data recording and

annotation procedure can be found in [23].

For the acoustic event detection task in DCASE2017 challenge, the selected 6 target acoustic event classes are: “brake squeaking”, “car”, “children”, “large vehicle”, “people speaking” and “people walking”. The total duration for the training and the test subset for each acoustic event class are shown in Table 1 and Table 2.

Table 1: Time duration (seconds) for the training subset for each event class.

	fold1	fold2	fold3	fold4
brake squeaking	70.08	86.80	79.02	66.20
car	1883.16	1405.46	1695.70	1729.60
children	303.52	328.66	267.80	107.54
large vehicle	682.46	665.66	617.02	629.20
people speaking	653.42	677.16	608.70	378.90
people walking	1069.54	1029.00	789.46	963.14

Table 2: Time duration (seconds) for the test subset for each event class.

	fold1	fold2	fold3	fold4
brake squeaking	30.62	13.90	21.68	34.50
car	348.82	832.52	542.24	508.40
children	32.32	7.18	68.04	228.30
large vehicle	182.32	199.12	247.76	235.58
people speaking	119.32	95.56	164.02	393.82
people walking	214.18	254.70	494.26	320.58

4.2. Evaluation metrics

The segment-based and event-based F-scores and error rates are used to evaluate the different AED systems. The segment-based F-score and error rate are calculated with respect to a segment. In this paper, the duration for the evaluation segment is set to 100ms. The event-based F-score and error rate are calculated with respect to the event instances. A higher F-score or a lower error rate indicates a better AED system. Detailed definitions about the F-score and error rate are described in [24]

4.3. Experimental results and analysis

To demonstrate the efficiency of our proposed system, several multi-layer percepton based AED systems are developed as follows:
1) MLP: This system is a multi-layer percepton with median filtering as the post processing technique as described in Section 2.
2) MLP-CWPB: This system uses a multi-layer percepton to train the acoustic models together with the class wise distance based probability.
3) MLP-CWRE: This system uses a multi-layer percepton to train the acoustic models and the class wise distance based re-evaluation strategy.
4) MLP-CW: This system uses a multi-layer percepton to train the acoustic models and adopts both the class wise distance based probability technique (CWPB) and the class wise distance based re-evaluation strategy (CWRE).

Table 3 presents the performance details of each defined AED system. As can be seen in Table 3, the system MLP-CW achieves the best overall performance among the MLP based AED systems, which indicates that the class wise distance based probability technique and post processing strategy benefit the AED system.

To further demonstrate the effectiveness of the class wise distance based probability technique and the re-evaluation strategy, the multi-layer percepton neural network is substituted by the convolutional neural network, the structure of which is described in Section 2.

Table 3: F-scores and error rates for different AED systems.

Techniques \ Metrics	segment F-score	segment ER	event F-score	event ER
MLP	56.15	0.69	5.10	3.35
MLP-CWPB	56.71	0.69	5.51	3.34
MLP-CWRE	56.14	0.67	5.65	2.97
MLP-CW	57.11	0.67	6.42	3.17
CNN	56.17	0.67	8.87	3.05
CNN-CWPB	56.02	0.65	**9.00**	2.94
CNN-CWRE	56.18	0.65	8.80	2.83
CNN-CW	**57.25**	**0.65**	8.92	**2.76**
DCASE2017 baseline[2]	56.70	0.69	-	-

From Table 3 it is evident that the CNN provides superior performance to MLP across most measures. Table 3 also shows the comparison between the development set baseline results from the challenge organisers and our proposed system. Our proposed system (CNN-CW) achieved a 0.55% absolute segment-based F-score improvement and 4% absolute segment-based error rate reduction.

5. CONCLUSION

This paper presents an approach to utilize the class wise information to improve the performance of the AED system. The class wise information makes the output probabilities more discriminative and the class distance based re-evaluation strategy can evaluate the acoustic events segmentally from a even longer duration. Experimental results demonstrate the efficiency of the proposed method by achieving a 4% absolute segment-based error rate reduction over the DCASE2017 challenge baseline. However, class wise distances for some events are too close to be discriminated (e.g the event “brakes squeaking” and the event “large vehicle”) and how to utilize the class wise distance when training the neural network models to ease this effect will be our next research direction.

6. ACKNOWLEDGMENT

This work was supported by the International Postgraduate Research Scholarship (IPRS) from the University of Western Australia. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

[2]http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio

7. REFERENCES

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2007, pp. 21–26.

[2] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models." in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 493–497.

[3] S. Päßler and W. J. Fischer, "Food intake monitoring: Automated chew event detection in chewing sounds," *IEEE journal of biomedical and health informatics*, vol. 18, no. 1, pp. 278–289, 2014.

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

[5] D. Giannoulis, S. Dan, B. Emmanouil, R. Mathias, L. Mathieu, and D. P. Mark, "A database and challenge for acoustic scene classification and event detection," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2013, pp. 1–5.

[6] T. Virtanen, A. Mesaros, T. Heittola, M. D. Plumbley, P. Foster, E. Benetos, M. Lagrange, E. Cakir, T. Heittola, T. Virtanen, *et al.*, "Proceedings of the detection and classification of acoustic scenes and events 2016 workshop (dcase2016)," vol. 6, 2016, p. 162.

[7] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 24th European*. IEEE, 2016, pp. 1128–1132.

[9] Z. Xiaodan, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using gaussian mixture models and gmm supervectors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 69–72.

[10] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. Workshop Applicat. Signal Process. Audio Acoust.(WASPAA)*. IEEE, 2013.

[11] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[12] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest regression based acoustic event detection with bottleneck features," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, accepted on 27th Feb.

[13] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest classification based acoustic event detection," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, accepted on 27th Feb.

[14] W. Dai, L. Juncheng, P. Pham, S. Das, S. Qu, and F. Metze, "Sound event detection for real life audio dcase challenge," *Detection and Classification of Acoustic Scenes and Events*, 2016.

[15] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for dcase challenge," 2016.

[16] A. Gorin, N. Makhazhanov, and N. Shmyrev, "Dcase 2016 sound event detection system based on convolutional neural network," *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*, 2016.

[17] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," *arXiv preprint arXiv:1706.02293*, 2017.

[18] T. Vu and J. Wang, "Acoustic scene and event recognition using recurrent neural networks," *Detection and Classification of Acoustic Scenes and Events*, 2016.

[19] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Frame-wise dynamic threshold based polyphonic acoustic event detection," in *Interspeech*. IEEE, 2017, accepted on 22nd May.

[20] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning*, 2010, pp. 807–814.

[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[24] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, no. 1, pp. 154–154, 2007.