SOUND EVENT DETECTION IN MULTICHANNEL AUDIO USING SPATIAL AND HARMONIC FEATURES

Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology

ABSTRACT

In this paper, we propose the use of spatial and harmonic features in combination with long short term memory (LSTM) recurrent neural network (RNN) for automatic sound event detection (SED) task. Real life sound recordings typically have many overlapping sound events, making it hard to recognize with just mono channel audio. Human listeners have been successfully recognizing the mixture of overlapping sound events using pitch cues and exploiting the stereo (multichannel) audio signal available at their ears to spatially localize these events. Traditionally SED systems have only been using mono channel audio, motivated by the human listener we propose to extend them to use multichannel audio. The proposed SED system is compared against the state of the art mono channel method on the development subset of TUT sound events detection 2016 database [1]. The usage of spatial and harmonic features are shown to improve the performance of SED.

Index Terms— Sound event detection, multichannel, time difference of arrival, pitch, recurrent neural networks, long short term memory

1. INTRODUCTION

A sound event is a segment of audio that a human listener can consistently label and distinguish in an acoustic environment. The applications of such automatic sound event detection (SED) are numerous; embedded systems with listening capability can become more aware of its environment [2][3]. Industrial and environmental surveillance systems and smart homes can start automatically detecting events of interest [4]. Automatic annotation of multimedia can enable better retrieval for content based query methods [5][6] .

The task of automatic SED is to recognize the sound events in a continuous audio signal. Sound event detection systems built so far can be broadly classified to monophonic and polyphonic. Monophonic systems are trained to recognize the most dominant of the sound events in the audio signal [7]. While polyphonic systems go beyond the most dominant sound event and recognize all the overlapping sound events in a segment [7][8][9][10]. We propose to tackle such polyphonic soundscape which replicates real life scenario in this paper.

Some SED systems have tackled polyphonic detection using mel-frequency cepstral coefficients (MFCC) and hidden Markov models (HMMs) as classifiers with consecutive passes of the Viterbi algorithm [7]. In [11], a non-negative matrix factorization was used as a pre-processing step, and the most prominent event in each of the stream was detected. However, it still had a hard constraint of estimating the number of overlapping events. This was overcome by using coupled NMF in [12]. Dennis et al [8] took an entirely different path from the traditional frame-based features by combining generalized Hough transform (GHT) with local spectral features.

More recently, the state of the art SED systems have used log mel-band energy features in DNN [9], and RNN-LSTM [10] networks trained for multi-label classification. Motivated by the good performance of RNN-LSTM over DNN as shown in [10], we continue to use the multi-label RNN-LSTM network.

The present state of the art polyphonic SED systems have been using a single channel of audio for sound event detection. Polyphonic events can potentially be tackled better if we had multichannel data. Just like humans use their two ears (two channels) to recognize and localize the sound events around them [13], we can also potentially train machines to learn sound events from multichannel of audio. Recently, Xiao et al [14] have successfully used spatial features from multichannel audio for far field automatic speech recognition (ASR) and shown considerable improvements over just using mono channel audio. This further motivates us to use spatial features for SED tasks. In this paper, we propose a spatial feature along with harmonic feature and prove its superiority over mono channel feature even with a small dataset of around 60 minutes.

The remaining of the paper is structured as follows. We describe in Section 2 the features used and the proposed approach. Section 3 presents a short introduction to RNNs and long short-term memory (LSTM) blocks. Section 4 presents the experimental set-up and results on a database of real life recordings. Finally, we present our conclusions in Section 5.

2. SOUND EVENT DETECTION

The sound event detection task involves identifying temporally the locations of sound event and assigning them to one among the known set of labels. Sound events in real life have no fixed pattern. Different contexts, for example, forest, city, and home have a different variety of sound events. They can be of different sparsity based on the context, and can occur in isolation or be completely overlapped with other sound events. While recognizing isolated sounds have been done with an appreciable accuracy [15], detecting the mixture of labels in an overlapped sound event is a challenging task, where still a considerable amount of improvements can be made. Figure 2 shows a snippet of sound event annotation, where three sound events - speech, car, and dog bark happen to occur. At time frame *t*, two events - speech and car are overlapping. An ideal SED system should be able to handle such overlapping events.

The human auditory system has been successfully exploiting the stereo (multichannel) audio information it receives at its ears to

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, and Google Faculty Research Award project “Acoustic Event Detection and Classification Using Deep Recurrent Neural Networks”. The authors also wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

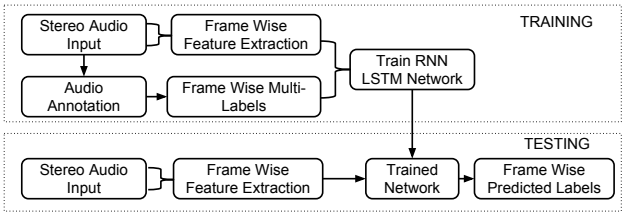

Figure 1: Framework of the training and testing procedure for the proposed system.

isolate, localize and classify the sound events. A similar set up is envisioned and implemented, where the sound event detection system gets a stereo input and suitable spatial features are implemented to localize and classify sound events.

The proposed sound event detection system, shown in Figure 1, works on real life multichannel audio recordings and aims at detecting and classifying isolated and overlapping sound events.

Three sets of features -log mel-band energies, pitch frequency, and its periodicity, and time difference of arrival (TDOA) in subbands, are extracted from the stereo audio. All features are extracted at a hop length of 20 ms to have consistency across features.

2.1. Log mel-band Energy

Log mel-band energies have been used for mono channel sound event detection extensively [9][10][16] and have proven to be good features. In the proposed system we continue to use log mel-band energies, and extract it for both the stereo channels. This is motivated from the idea that human auditory system exploits the interaural intensity difference (IID) for spatial localization of sound source [13]. Neural networks are capable of performing linear operations, which includes the difference. Therefore, when trained on the stereo log mel-band energy data, it will learn to obtain information similar to IID.

Each channel of the audio is divided into 40 ms frames with 50% overlap using hamming window. Log mel-band energies are then extracted for each of the frames (mel in Table 1). We use 40 mel-bands spread across the entire spectrum.

2.2. Harmonic features

The pitch is an important perceptual feature of sound. Human listeners have evolved to identify different sounds using the pitch cues, and can make efficient use of pitch to acoustically separate each of the mixture in an overlapping sound event [17]. Uzkent et al [18] have shown improvement in accuracy of non speech environmental sound detection used pitch range along with MFCC's. Here we propose using the absolute pitch and its periodicity as the features ($pitch$ in Table 1).

The librosa implementation of pitch tracking [19] on thresholded parabolically-interpolated STFT [20] was used to estimate the pitch and periodicity.

Since we are handling multi-label classification it is intuitive to identify as many dominant fundamental frequencies as possible and use them to identify the sound events. The periodicity feature gives the confidence measure for the extracted pitch value and helps the classifier to make better decisions based on pitch.

The overlapping sound events in the training data (Section 4.1) did not have more than three events overlapping at a time, hence we have limited ourselves to using the top three dominant pitch values per frame. So, for each of the channels, top three pitch values, and its respective periodicity values are extracted at every frame in 100-4000 Hz frequency range ($pitch3$ in Table 1).

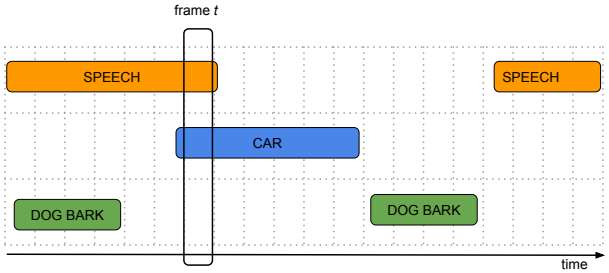

Figure 2: Sound events in a real life scenario can occur in isolation or overlapped. We see that at frame t, speech and car events are overlapping.

2.3. Time difference of arrival (TDOA) features

Overlapping sound events have forever troubled classification systems. This is mainly because the feature vector for the overlapped frame is a combination of different sound events. But, human listeners have been able to successfully identify each of the overlapping sound events by isolating and localizing the source spatially. This has been possible due to the interaural time delay (ITD) [13]

Each sound event has its own frequency band, some occur in low frequencies, some in high, and some occur all across the frequency band. If we can divide the frequency spectrum into different bands, and identify the spatial location of the sound source in each of these bands, then this is an extra dimension of the feature, which the classifier can learn to estimate the number of possible sources in each frame, and their orientation in the space. We implement this by dividing the spectral frame into five mel-bands and calculating the time difference of arrival (TDOA) at each of these bands.

For example, if a non-overlapping isolated sound event is spread across the entire frequency range, and we are calculating the TDOA in five mel-bands. We should have the same TDOA values for each of the bands. However, if we have two overlapping sounds S_1 and S_2, where S_1 is spread in the first two bands and S_2 is spread in the last two bands. The feature vector will have different TDOA values for each of the sounds, which the classifier can learn to isolate and identify them as separate sound events.

The TDOA can be estimated using the generalized cross-correlation with phase-based weighting (GCC-PHAT) [21]. Here, we extract the correlation for each mel-band separately:

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

where N is the number of frequency bands, $X(k,t)$ is the FFT coefficient of the kth frequency band at time frame t and the subscript specifies the channel number, $H_b(k)$ is the magnitude response of the bth mel-band of total of B bands and Δ_{12} is the sample delay value between channels. The TDOA is extracted as the location of correlation peak magnitude for each mel-band and time frame.

$$\tau(b,t) = \underset{\Delta_{12}}{\text{argmax}} \{R_b(\Delta_{12}, t)\} \quad (2)$$

The maximum and minimum TDOA values are truncated between values $-2\tau_{\text{max}}, 2\tau_{\text{max}}$, where τ_{max} is the maximum sample delay between a sound wave traveling between microphones.

Feature Name	Length	Description
mel	40	Log mel-band energy extracted on a single channel of audio
$pitch$	2	Most dominant pitch value and periodicity extracted on a single channel
$pitch3$	6	Top three dominant pitch and periodicity values extracted on a single channel
$tdoa$	5	Median of multi-window TDOA's extracted from stereo audio
$tdoa3$	15	Concatenated multi-window TDOA's extracted from stereo audio

Table 1: Definitions of acoustic features proposed for sound event detection.

The sound events in the training set were seen to be varying from 50 ms to a few seconds. In order to accommodate such variable length sound events, TDOA was calculated in three different window lengths — 120, 240 and 480 ms, with a constant hop length of 20 ms. The TDOA values of these three windows were concatenated for each mel-band to form one set of TDOA features. So, TDOA values extracted in five mel-band, and for three window lengths, on concatenation gives 15 TDOA values per frame ($tdoa3$ in Table 1).

TDOA values in small windows are generally very noisy and unreliable. To overcome this, the median of the TDOA values from the above three different window lengths for each sub-band of the frame was used as the second set of TDOA features ($tdoa$ in Table 1). Post filtering across window lengths, the TDOA values in each mel-band were also median filtered temporally using a kernel of length three to remove outliers.

3. MULTI-LABEL RECURRENT NEURAL NETWORK BASED SOUND EVENT DETECTION

Deep neural networks have shown to perform well on complex pattern recognition tasks, such as speech recognition [22], image recognition [23] and machine translation [24]. A deep neural network typically computes a map from an input to an output space through several subsequent matrix multiplications and non-linear activation functions. The parameters of the model, i.e. its weights and biases, are iteratively adjusted using a form of optimization such as gradient descent.

When the network is a directed acyclic graph, i.e. information is only propagated forward, it is known as a feedforward neural network (FNN). When there are feedback connections the model is called a recurrent neural network (RNN). An RNN can incorporate information from previous timesteps in its hidden layers, thus providing context information for tasks based on sequential data, such as temporal context in audio tasks. Complex RNN architectures — such as long short-term memory (LSTM) [25] — have been proposed in recent years in order to attenuate the vanishing gradient problem [26]. LSTM is currently the most widely used form of RNN, and the one used in this work as well.

In SED, RNNs can be used to predict probabilities for each class to be active in a given frame at timestep t. The input to the network is a sequence of feature vectors $\mathbf{x}(\mathrm{t})$; the network computes hidden activations for each hidden layer, and at the output layer a vector of predictions for each class $\mathbf{y}(\mathrm{t})$. A sigmoid activation function is used at the output layer in order to allow several classes to be predicted as active simultaneously. By thresholding the predictions at the output layer it is possible to obtain a binary activity matrix.

3.1. Neural network configurations

For each recording, we obtain a sequence of feature vectors, which is normalized to zero mean and unit variance, and the scaling parameters are saved for normalizing the test feature vectors. The sequences are further split into non-overlapping sequences of length 25 frames. Each of these frames has a target binary vector, indicating which classes are present in the feature vector.

We use a multi-label RNN-LSTM with two hidden layers each having 32 LSTM units. The number of units in the input layer depends on the length of the feature being used. The output layer has one neuron for each class. The network is trained by back propagation through time (BPTT) [27] using binary cross-entropy as loss function, Adam optimizer [28] and block mixing [10] data augmentation. Early stopping is used to reduce over-fitting, the training is halted if the segment based error rate (ER) (see Section 4.2) on the validation set does not decrease for 100 epochs.

At test time we use scaling parameters estimated on training data to scale the feature vectors and present them in non-overlapping sequences of 25 frames, and threshold the outputs with a fixed threshold of 0.5, i.e., we mark an event is active if the posterior in the output layer of network is greater than 0.5 and otherwise inactive.

4. EVALUATION AND RESULTS

4.1. Dataset

We evaluate the proposed SED system on the development subset of TUT sound events detection 2016 database [1]. This database has stereo recordings which were collected using binaural Soundman OKM II Klassik/studio A3 electret in-ear microphones and Roland Edirol R09 wave recorder using 44.1 kHz sampling rate and 24-bit resolution. It contains two contexts - home and residential area. Home context has 10 recordings with 11 sound event classes and the residential area context has 12 recordings with 7 classes. The length of these recordings is between 3-5 minutes.

In the development subset provided, each of the context data is already partitioned into four folds of training and test data. The test data was collected such that each recording is used exactly once as the test, and the classes in it are always a subset of the classes in the training data. Also, 20% of the training data recordings in each fold were selected randomly to be used as validation data. The same validation data was used across all our evaluations.

4.2. Metrics

We perform the evaluation of our system in a similar fashion as [1] which uses the established metrics for sound event detection defined in [30]. The error rate (ER) and F-scores are calculated on one second long segments. The results from all the folds are combined to produce a single evaluation. This is done to avoid biases caused due to data imbalance between folds as discussed in [31].

4.3. Results

The baseline system for the dataset [1] uses 20 static (excluding the 0th coefficient), 20 delta and 20 acceleration MFCC coefficients

	Feature combination	Home		Residential area		Average	
		ER	F (%)	ER	F (%)	ER	F (%)
Baseline system using GMM classifier in DCASE 2016 [1][29]	$mfcc; delta; acc$	0.96	15.9	0.86	31.5	0.91	23.7
Mono channel feature With RNN-LSTM network	mel_1	**0.94**	**27.4**	0.88	38.3	0.91	32.9
Hybrid (mono and stereo) features with RNN-LSTM network	$mel_1; pitch_1$	0.97	25.4	0.85	43.4	0.91	34.4
	$mel_1; pitch3_1$	0.96	27.6	0.88	43.9	0.92	35.7
	$mel_1; tdoa$	1.02	19.4	0.89	40.2	0.96	29.8
	$mel_1; tdoa3$	0.98	25.9	0.87	40.5	0.92	33.2
Stereo features with RNN-LSTM network	mel_2	1.03	25.4	0.84	45.9	0.93	35.6
	$mel_2; pitch_2$	1.03	24.9	0.93	40.9	0.98	32.9
	$mel_2; pitch3_2$	0.97	26.6	0.88	41.7	0.92	34.2
	$mel_2; tdoa$	1.01	24.4	**0.82**	**46.4**	**0.91**	**35.4**
	$mel_2; tdoa3$	0.96	24.9	0.86	38.5	0.91	31.7
	$mel_2; tdoa3; pitch_2$	0.97	25.7	0.85	43.1	0.91	34.4
	$mel_2; tdoa3; pitch3_2$	0.99	26.5	0.91	35.2	0.95	30.9
	$mel_2; tdoa; pitch_2$	0.98	24.7	0.87	43.8	0.92	34.2
	$mel_2; tdoa; pitch3_2$	0.94	26.3	0.89	40.5	0.91	33.4

Table 2: Segment based error rate (ER) and F-score achieved for different feature combinations in home and residential area contexts for the development set. The features listed in Table 1 are used in different combinations with the proposed RNN-LSTM network. The subscripts '$_1$' and '$_2$' in the feature combinations column represent how many channels the features were extracted on. For example, feature combination $mel_2; tdoa; pitch_2$ means that the final feature vector has log mel-band energies, most dominant pitch and periodicity values extracted on both the stereo channels, and the time difference of arrival (TDOA) calculated between the stereo channels. The highlighted ER and F-score pair for each context is the best ER score achieved.

extracted on mono audio with 40 ms frames and 20 ms hop length. A Gaussian mixture model (GMM) consisting of 16 Gaussians is then trained for each of the positive and negative values of the class. This baseline system gives a context average ER of 0.91 and F-score of 23.%. An ideal system should have an ER of 0 and an F-score of 100%.

In Table 2 we compare the segment based ER and F-score for different combinations of proposed spatial and harmonic features. In all these evaluations, only the size of the input layer changes based on the feature set, with the rest of the configurations in the RNN-LSTM network remaining unchanged.

Mono channel audio was created by averaging the stereo channels in order to compare the performance of the proposed spatial and harmonic features for multichannel audio. One of the present state of the art SED system for mono channel is proposed in [10]. An RNN-LSTM network is trained in a similar fashion with log mel-band energy feature (Section 2.1) and evaluated. Across contexts, the F-score was seen to be better than the GMM baseline system with comparable ER. Here onwards we use this mono-channel log mel-band feature and RNN-LSTM network configuration result as a baseline for comparisons.

A set of hybrid combinations were tried as shown in Table 2. All combinations other than $mel_1; tdoa$ performed better than the baseline across contexts in F-score.

Finally, the full spectrum of proposed spatial and harmonic features were evaluated in different combinations with RNN-LSTM network. With a couple of exceptions - $mel_2; pitch_2$ and $mel_2; tdoa3; pitch3_2$, all the combinations of features performed equal to or better than the baseline in average F-scores, with marginally similar average ER as baseline. Given the dataset size of around 60 minutes, it is difficult to conclusively say that the binaural features are far superior to monaural features; but they surely look promising.

Binaural features - mel_2 and $mel_2; tdoa; pitch_2$ in Table 3 were submitted to the DCASE 2016 challenge [29], where they were evaluated as the top performing systems. Monaural feature mel_1 was submitted unofficially to compare the performance with binaural features. The hyper-parameters of the network were tuned before the submission, and hence the development set results in Table 3 are different from Table 2. Three hidden layers with 16 LSTM units each were used for mel_2, while mel_1 and $mel_2; tdoa; pitch_2$ were trained with two layers each having 16 LSTM units.

Feature combination	Evaluation dataset		Development dataset	
	ER	F (%)	ER	F (%)
mel_1	0.79	46.6	0.90	35.3
mel_2	0.80	47.8	0.88	34.7
$mel_2; tdoa; pitch_2$	0.88	37.9	0.87	34.8

Table 3: Comparison of segment based error rate (ER) and F-score for development and evaluation dataset. The evaluation dataset scores are the result of DCASE 2016 challenge [29].

5. CONCLUSION

In this paper, we proposed to use spatial and harmonic features for multi-label sound event detection along with RNN-LSTM networks. The evaluation was done on a limited dataset size of 60 mins, which included four cross validation data for two contexts — home and residential area. The proposed multi-channel features were seen to be performing substantially better than the baseline system using mono-channel features.

Future work will concentrate on finding novel data augmentation techniques. Augmenting spatial features is an unexplored space, and will be a challenge worth looking into. Concerning the model, further studies can be done on different configurations of RNN like extending them to bidirectional RNN's and coupling with convolutional neural networks.

6. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[2] S. Chu, S. Narayanan, and C. J. Kuo, "Environmental sound recognition with timefrequency audio features," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, 2009, p. 1142.

[3] S. Chu, S. Narayanan, C. C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *IEEE Int. Conf. Multimedia and Expo (ICME)*, 2006, p. 885.

[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)*, vol. 4, no. 2, 2008, p. 11.

[6] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," in *Dept. Electronic Eng., Columbia Univ., New York*, 2001.

[7] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013, p. 1.

[8] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," in *Pattern Recognition Letters*, vol. 34, no. 9, 2013, p. 1085.

[9] E. Cakir, 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.

[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] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada*, 2013,, p. 8677.

[12] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013,, p. 1.

[13] J. W. Strutt, "On our perception of sound direction," in *Philosophical Magazine*, vol. 13, 1907,, p. 214.

[14] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and DongYu, "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*, 2016,.

[15] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *European Signal Processing Conference (EUSIPCO 2014)*, 2014,.

[16] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

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

[18] B. Uzkent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using svms with a new set of features," in *International Journal of Innovative Computing, Information and Control*, 2012, p. 3511.

[19] B. McFee, M. McVicar, C. Raffel, D. Liang, O. Nieto, E. Battenberg, J. Moore, D. Ellis, R. YAMAMOTO, R. Bittner, D. Repetto, P. Viktorin, J. F. Santos, and A. Holovaty, "librosa: 0.4.1," Oct. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.32193

[20] 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

[21] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[22] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[27] P. J. Werbos, "Backpropagation through time: what it does and how to do it," in *Proceedings of the IEEE,*, vol. 78 no. 10, 1990, p. 15501560.

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

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

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6(6):162, 2016.

[31] G. Forman and M. Scholz, "Apples-to-apples in cross validation studies: Pitfalls in classifier performance measurement," in *SIGKDD Explor. Newsl.*, vol. 12, no. 1, Nov. 2010, p. 49.