Interspeech 2018
2-6 September 2018, Hyderabad

Detecting Media Sound Presence in Acoustic Scenes

Constantinos Papayiannis[1,2], *Justice Amoh*[1,3], *Viktor Rozgic*[1], *Shiva Sundaram*[1] *and Chao Wang*[1]

[1]Alexa Machine Learning, Amazon.com, Cambridge, MA, USA
[2]Department of Electrical and Electronic Engineering, Imperial College London, UK
[3]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

papayiannis@imperial.ac.uk, justice.amoh.jr.th@dartmouth.edu, rozgicv@amazon.com, sssundar@lab126.com, wngcha@amazon.com

Abstract

Using speech to interact with electronic devices and access services is becoming increasingly common. Using such applications in our households poses new challenges for speech and audio processing algorithms as these applications should perform robustly in a number of scenarios. Media devices are very commonly present in such scenarios and can interfere with the user-device communication by contributing to the noise or simply by being mistaken as user issued voice commands. Detecting the presence of media sounds in the environment can help avoid such issues. In this work we propose a method for this task based on a parallel CNN-GRU-FC classifier architecture which relies on multi-channel information to discriminate between media and live sources. Experiments performed using 378 hours of in-house audio recordings collected by volunteers show an F1 score of 71% with a recall of 72% in detecting active media sources. The use of information from multiple channels gave a relative improvement of 16% to the F1 score when compared to using information from only a single channel.

Index Terms: audio classification, media sound detection, acoustic scene analysis

1. Introduction

Analysis of acoustic scenes has been an active area of research for a number of years. The need for this understanding has increased significantly as a number of products allow us to use speech in our household acoustic environments to control home automation, entertainment systems and much more. Reliable speech communication with these devices greatly enhances the user-experience. It is very common for media sound activity, such as a TV set playing, to be active for long durations of time. Detecting the presence of the media device can help discriminate between user speech and media sounds and separate them for better interaction with the user.

The task of detecting the presence of media sources presents a new challenge to the field of analyzing acoustic environments. It overlaps with the concepts of Scene and Content Classification and Audio Event Detection (AED) however it cannot be defined as either. The main goal is to determine whether the received sounds have been reproduced by a media device present in the acoustic environment. The overlap with the fields discussed above is still significant as types of sound events and content type can be used as context to improve the accuracy of detecting media presence. In the case of simple acoustic scenes, such as speech, discriminating live and media sources is a great challenge which hasn't been directly addressed in the literature of the aforementioned fields. To address this we aim to characterize the channel between the emitted sound and the receiver.

In early work, scene analysis was broadly referring to the task of understanding the acoustic environment. In [1] the author describes it as the steps taken by humans to solve the "cocktail party" problem. Practical work to perform this was shown in [2], inspired by the human auditory system to do that. The concept has received significant research interest over the years. Recently, through the DCASE challenge [3] many interesting new approaches to the task of Scene Classification and AED have been proposed. Early work on AED relied on simple classification techniques [4] in order to categorize sound events and audio inputs in general. With recent advancements in machine learning algorithms, more complex models are trained for the tasks, using a larger amount of data. In [5, 6] Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were combined for the task. Detection through use of multi-channel information has shown promising results in [7]. Multi-channel information was also used for the task of Scene Classification in [8], with the audio further separated to harmonic and percussive components. Long Short-Term Memory (LSTM) networks and CNNs were used in parallel for the task in [9].

A growing issue with the development of new methods for the task of Scene Classification and AED is the lack of large scale databases with appropriate annotations. Many novel approaches have been previously proposed to address this. Learning through visual information and inferring only from audio cues has been proposed in [10] and the generation of further audio examples from available recordings has been proposed in [11].

In this work, we formulate the task of detecting media presence in everyday household acoustic environments. Inspired by the success of previously proposed methods for AED and Automatic Speech Recognition (ASR) [7] we investigate the use of multi-channel information for the task. Our motivation is that information from multiple sensors will be useful in discriminating between non-stationary live sources and stationary media sources in space. Spectral features are also used with the motivation that audio content can be inferred from spectral representations which can reinforce the classification. The proposed model is based on a parallel CNN-GRU-FC architecture [12], with separate convolutional layers and different frame rates for spectral inputs from a sensor-pair.

The rest of the paper is organized as follows: Section 2 formulates the problem of media presence detection, describes the features and the proposed model used for the task. Experimental results are presented in Section 3 and further discussed in Section 4. A conclusion is given in Section 5.

10.21437/Interspeech.2018-2559

2. Method

2.1. Signal Model

For a microphone array with Q sensors, the received signal at sensor q at sample with index n is denoted as $x_q(n)$. Assuming J number of live sources and L media sources with $K = L+J$, producing signals s_k, the following expression can be formed

$$x_q(n) = \sum_{k=1}^{K} s_k(n) * \mathbf{h}_{q,k}(n) + \theta(n), \tag{1}$$

where $\mathbf{h}_{q,k}(n)$ the Acoustic Impulse Response (AIR) of the system from source k to sensor q [13] and $\theta(n)$ the additive noise signal.

For live sources such as humans present in the room, the signal $s_j(n)$ can be speech, with their contribution to the observed signals being reverberant speech. For the case of media sources however their contribution to the observed signals will be more complex. Denoting the speech signal produced by the media source as $s_l(n)$, this signal is not guaranteed to be anechoic. It will also be affected by the response of the media device. For L active media devices, the responses $\mathbf{d}_l$ describe the effect of the recording equipment and environment along with the effect of the media device to the original signal played back by the device. Temporarily assuming that the additive noise is negligible, (1) can be rewritten as

$$x_q(n) = \sum_{j=1}^{J} s_j(n) * \mathbf{h}_{q,j}(n) + \sum_{l=1}^{L} s_k(n) * \mathbf{h}_{q,l}(n) * \mathbf{d}_l(n). \tag{2}$$

Spanning the above for $n \in \{0, \dots, N\}$, forms the vector $\mathbf{x}_q$ and subsequently the array $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_Q]$. Columns of the array correspond to different sensors. The objective of this work is, given observation $\mathbf{X}$ to determine whether any media sources are active at any instance throughout the recording. We wish therefore to design a model which performs the following task

$$g(\mathbf{X}) = \begin{cases} 1, & \text{if } L \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

2.2. Feature Extraction

There are two main discriminative aspects that we wish to exploit in this work. The signals s_j and s_k, relating to live and media sources respectively, are expected to have different characteristics. Media sources are active for prolonged periods of time and relate to complex and fast changing scenes. Live sources on the other hand can have sparse activities and usually refer to simpler acoustic scenes. Examples which illustrate these differences are music playback or the broadcast of an action movie from a TV set versus a human present in a kitchen preparing a meal. This first discriminative aspect between media and live sources is expected to be present when observing the spectrum of the received signals and their temporal progression.

The above example illustrates another important difference between media and live sources. Although media sources are complex with regards to their content, they are very predictable in terms of their Direction of Arrival (DoA). Media sources which are active for prolonged periods of time are expected to be stationary in space. Humans on the other hand can be mobile in space and also are free to speak in different directions, altering the DoA of sound to the sensors. This forms the second discriminative aspect we wish to exploit by observing spatial information extracted from the observed signal at the array.

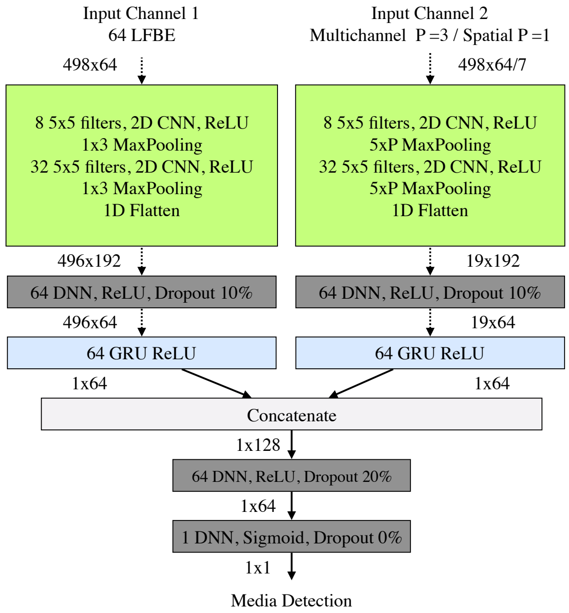
Input Channel 1
64 LFBE
498x64
8 5x5 filters, 2D CNN, ReLU
1x3 MaxPooling
32 5x5 filters, 2D CNN, ReLU
1x3 MaxPooling
1D Flatten
496x192
64 DNN, ReLU, Dropout 10%
496x64
64 GRU ReLU
1x64
Input Channel 2
Multichannel P =3 / Spatial P =1
498x64/7
8 5x5 filters, 2D CNN, ReLU
5xP MaxPooling
32 5x5 filters, 2D CNN, ReLU
5xP MaxPooling
1D Flatten
19x192
64 DNN, ReLU, Dropout 10%
19x64
64 GRU ReLU
1x64
Concatenate
1x128
64 DNN, ReLU, Dropout 20%
1x64
1 DNN, Sigmoid, Dropout 0%
1x1
Media Detection

Figure 1: *Split Long- Short-Term (SLST) Model Architecture*

Given the above discussion, we investigate the use of Log Filterbank Energies (LFBE's) as spectral features, which can be extracted from a single or multiple channels independently. For band $b \in \{1, \dots, B\}$ defined by the filter response $F_b(v) \in \mathbb{R}^{N/2}$, this is defined as

$$X_q(b) = \log \sum_{v=0}^{N/2} F_b(v) \left[\sum_{n=0}^{N} x_q(n) e^{\frac{i2\pi vn}{N}} \right]^2 . \tag{4}$$

Using LFBE's from a number of sensors can provide models with information to infer spatial properties of the scene which will help discriminate between media and live sources.

The time-domain cross-correlation between frames recorded at different sensors is also considered as spatial information and it is estimated as

$$c_{m_1,m_2}(n_c) = \sum_{n=0}^{N} x_{m_1}(n) x_{m_2}(n + n_c). \tag{5}$$

The motivation is that cross-correlation values across different sensor pairs will provide spatial information which will help discriminate between media and live sources.

2.3. Media Sound Detection Model

The model we proposed for this task is based on a parallel CNN-GRU-FC architecture. The model accepts two inputs extracted from one sensor pair. The first input consists of LFBE's vectors extracted from the first sensor. The second input is derived from the second sensor only, or a combination of the two sensors. We investigate the use of LFBE's and channel cross-correlations. The model is named Split Long- Short-Term (SLST) for reasons to be discussed in this section and it's outlined in Figure 1.

The SLST model separates the two inputs directly into two branches. The motivation of our work is shared to a wide extent by the work done in [7] for AED, where spatial and mutlichannel features along with spectral features were used. It is proposed to use separate convolutional layers branches for each feature set. The same approach is followed by the SLST model. The convolutional layers extract useful representations of the input and are followed by different max pooling layers in each branch. The first branch which processes spectral information from the first sensor performs pooling only in the frequency domain in order to reduce variations in frequency [12]; however, for the multi-channel or spatial information, pooling is done also in the time domain. Pooling in frequency is not suitable for spatial inputs, therefore the size of pooling filter size along the frequency axis is P=1. For for multichannel inputs we use P=3, matching the frequency pooling applied in the first branch. With the timescale reduced by 25 before the fully-connected (FC) layer for the second branch, the fine time domain information is lost and only the salient spectral-energy or spatial features are retained, hence the name Split Long- Short-Term (SLST). This approach was taken in order to guide the first branch to learn the contents of inputs and the left branch to learn a compact and spatially aware representation of the acoustic scene. A similar concept is discussed in [14] but with the two branches sharing the same input. The convolutional layers are followed by FC layers which have a twofold purpose, to reduce the dimensionality of the activations [12] and allow for more effective use of dropout which improves generalization [15]. Gated Recurrent Unit (GRU) layers are used, with the first branch layer processes 25 frames per every 1 frame the second branch does. The dimensionality is then halved using a FC layer before the final FC layer leads to the output neuron.

2.4. Model Training

The model is optimized using “Adam” [16] with cross-entropy loss with early-stopping. When the validation loss does not decrease for 15 epochs, training stops and the model with the lowest validation loss is kept.

Training data are split into train, validation and test sets with ratios which consist of 85.0%, 7.5% and 7.5% of the data respectively. Stratified partitioning is used for the splits with regards to class labels. The training data are split into positive and negative samples before training. At the beginning of each epoch, the two sets are shuffled [17]. Batches of 128 samples are generated during training which consist of 64 positive and 64 negative samples. This accounts for any imbalance in the data and avoids class-wise weighting of losses.

2.5. Data Augmentation

As in the case of AED, the availability of annotated data for the task is limited [11]. In our case, we wish for the available data to be in a multi-channel format. This limits the data availability even further. In order to acquire more data for training, we employ a method which spatializes monaural recordings to an arbitrary number of channels. The method relies on the availability of a number of monaural recordings which involve one stationary source and a stationary array, which is used to beamform towards the direction of the source. We assume that the received signal at this point is an estimate of the anechoic signal emitted by the source. Labeling the beamformed signal as $x_{bf}(n)$, the augmented data samples forming $\tilde{\mathbf{X}}$ are generated using

$$\tilde{x}_q(n) = x_{bf}(n) * \tilde{\mathbf{h}}_q(n), \qquad (6)$$

for $q \in \{1, \ldots, 7\}$. The AIR $\tilde{\mathbf{h}}_m(n)$ is generated using the model described in [18]. The direct path sound Time Difference of Arrival (TDoA) for the Q=7 receivers is contained based on the array architecture. The rest of the reflections are placed at random time intervals, with their amplitude scaled using Polack’s model [13].

3. Experiments

The performance of the proposed model is evaluated using 311 hours of audio data collected from household acoustic environments of volunteers participating in a data collection program. The 7-channel data was collected using Amazon Echo Dot devices [19]. The recordings are segmented into 5 second utterances which are individually labeled. The proportion of all utterances for each label prior to the addition of the augmentation data is the following: Media 12.1%, Human 33.5%, Other 73.0% and Silence 15.9%. With the exception of silence, the labels refer to sound sources. Labels are not exclusive and multiple source can be active in each utterance. An additional 67 hours of beamformed audio was augmented using the method described in Section 2.5 and used for training. All the augmented data samples are positive for media presence.

The model performance is compared against a baseline model. The baseline model uses information from a single channel and it is identical to the model described in Figure 1 however without the rightmost branch, which refers to the use of multi-channel data. The comparison between this baseline and the SLST model provides insight into the benefit of using multi-channel information as proposed in this work.

The 64 LFBE’s are extracted between 0.1 and 7.2 kHz and the filters $F_b(v)$ have their geometric centers an equal distance apart on the mel-scale. Extraction is done after the signal is segmented into frames of 25 ms with a step of 10 ms. Cross-correlations are extracted for the same frames using (5). One sensor pair is used for the extraction and the cross-correlation is estimated for frames corresponding to the same timestep. The lag is varied between -3 and $+3$ samples, which gives 7 coefficients per frame. All features are extracted from a maximally spaced sensor pair on the ring of the Echo Dot device array.

Training both the SLST and baseline models using the method described in Section 2.4 gives the results of Table 1.

Model $g(\cdot)$	F1	Precision	Recall	Accuracy
Baseline 1-LFBE	0.61	0.55	0.68	0.86
SLST LFBE + xCorr (P=1)	0.67	0.63	0.71	0.89
SLST LFBE + LFBE (P=3)	0.71	0.70	0.72	0.91

Table 1: *Baseline and SLST Results for Media Detection*

4. Discussion

Using spatial features provides a relative improvement of 10% to the F1 score of the classification over the baseline which relies only on monaural spectral features. Using the LFBE’s from a second channel offered a relative improvement to the F1 of 16%. This result illustrates that using information from a multiple channels can lead to improvements in the detection of active media sound sources. Spatial information in the form of time-domain cross-correlations offers a significant benefit in the classification. Using spectral information from two channels offers the best performance, and subsequent analysis we present concentrates on this model. The statistical significance of the improvement with respect to the baseline is evaluated using the method described in [20]. We split the test population into 30 splits, from which the error rate [20] is evaluated. With a 99% confidence interval the difference in performance between the proposed model and the baseline is significant.

Media Speech	Media Singing	Media Other	Baseline Accuracy	SLST Accuracy
			0.85	0.85
		✓	0.62	0.75
	✓		0.67	0.92
	✓	✓	0.97	1.00
✓			0.68	0.76
✓		✓	0.90	0.97
✓	✓		1.00	1.00
✓	✓	✓	1.00	1.00

Table 2: *Baseline and SLST model accuracy for media types with operating points set for EER. Summary of the accuracy in detecting the presence of media sound sources in scenes which contain the types indicated in the corresponding column per row.*

In Table 2, the accuracy of detecting specific types of media sounds is listed. For these results, the operating point for each network was set to provided an Equal Error Rate (EER). This setting balances the false-accept rate and the false-reject rate. The results summarize the accuracy of media utterances which contain the types indicated by ✓ in the corresponding column per row. The two rightmost columns compare the performance of the baseline and proposed SLST model (using two channel LFBE's) for each category. From the results, we can deduce that the proposed SLST model is more accurate for the majority of cases. It performs better at both detecting and rejecting media and non-media sources respectively. The most challenging task appears to be the distinction between media and live speech. Both models perform poorly for this subtask, with the SLST model offering marginal improvements.

The content proves to be an important factor for distinction of media and live sources. Samples containing music are reliably classified; however, speech signals are a source of confusion to the classifier. The multi-channel information improved performance on speech samples. Increased mobility of live sources makes multi-channel and spatial information more useful. Limited bandwidth of speech signals makes their classification more challenging as it provides less information about $\mathbf{d}_l(n)$ as given by equation (2).

5. Conclusion

A method for the detection of active media sound sources in acoustic scenes has been proposed. The proposed method relies on a CNN-GRU-FC model architecture, named Split Long-Short-Term (SLST), and the use of multi-channel information. The performance of the model is compared to a baseline model which relies on a single-channel input. Experiment results on dataset containing 378 hours of in-house audio recordings collected by volunteers show that the proposed model outperforms the baseline with an F1 score of 71%, a 16% relative improvement. The proposed model performs better both at detecting and rejecting media and non-media sources respectively.

6. References

[1] Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[2] M. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, 1993.

[3] 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 *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, Nov. 2017.

[4] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, vol. 3, no. 3, pp. 27–36.

[5] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[6] H. Lim, J. Park, and Y. Han, "Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks," in *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, Nov. 2017.

[7] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *arXiv Preprint arXiv:1706.02291*, 2017.

[8] Y. Han, J. Park, and K. Lee, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," in *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, Nov. 2017.

[9] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. DCASE 2016 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Budapest, Hungary, Sep. 2016.

[10] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," *arXiv:1610.09001 [cs]*, Oct. 2016.

[11] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," in *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, Nov. 2017.

[12] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Queensland, Australia: IEEE, Apr. 2015, pp. 4580–4584.

[13] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer, 2010.

[14] A. Schindler, T. Lidy, and A. Rauber, "Multi-Temporal Resolution Convolutional Neural Networks for Acoustic Scene Classification," in *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, Nov. 2017.

[15] Y. Gal and Z. Ghahramani, "A Theoretically Grounded Application of Dropout in Recurrent Neural Networks," *arXiv:1512.05287 [stat]*, Dec. 2015.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014.

[17] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2012, pp. 437–478.

[18] C. Papayiannis, C. Evers, and P. A. Naylor, "Sparse parametric modeling of the early part of acoustic impulse responses," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug. 2017, pp. 678–682.

[19] Amazon, "Amazon Echo Dot," https://www.amazon.com/Amazon-Echo-Dot-Portable-Bluetooth-Speaker-with-Alexa-Black/dp/B01DFKC2SO/.

[20] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.