SCENE-DEPENDENT ANOMALOUS ACOUSTIC-EVENT DETECTION BASED ON CONDITIONAL WAVENET AND I-VECTOR

Tatsuya Komatsu[⋆], *Tomoki Hayashi*[†], *Reishi Kondo*[⋆], *Tomoki Toda*[‡], *Kazuya Takeda*[†]

[⋆]Data Science Research Laboratories, NEC Corporation, Japan
[†]Department of Information Science, Nagoya University, Japan
[‡]Information Technology Center, Nagoya University, Japan

ABSTRACT

This paper proposes a scene-dependent anomalous acoustic-event detection based on conditional WaveNet and i-vector. The WaveNet builds normal acoustic event models by exhaustive learning of time-domain signals in the public space to provide scene-independent anomaly detection. I-vectors are used as additional features to describe acoustic scenes, where the input signals are observed, to complement the WaveNet. The proposed method can detect anomalous acoustic-events in environments whose acoustic scenes vary depending on time, location, and surrounding environment. Evaluations with data recorded from the real environment demonstrate that the proposed method achieved as much as 15 pt higher F-measure than LSTM and AE. The difference in F-measure by the WaveNet with and without i-vector turned out to be 1.5 pt.

Index Terms— WaveNet, Anomaly detection, i-vector, Anomalous sound event detection

1. INTRODUCTION

Anomalous acoustic-event detection methods based on a DNN (deep neural network) have been actively investigated in recent years [1, 2, 3]. Among others, a WaveNet based method by Hayashi *et al.* [4] achieved good detection performance. WaveNet [5] is a predictor which receives a fixed length time-domain signal and outputs an *a posteriori* probability distribution of the next sample amplitude. It precisely trains models of time-domain signals. Hayashi *et al.* builds normal acoustic event models with the WaveNet by exhaustive training of signal data obtained in the public space. An input is detected as an anomalous acoustic-event when the *a posteriori* probability distribution given by the WaveNet is not sufficiently sharp. It is because a flat *a posteriori* probability distribution means that there is no acoustic event model in the WaveNet to describe an anomalous acoustic-event which has not been encountered during the training process.

However, WaveNet cannot detect a scene-dependent anomalous acoustic-event in environments whose acoustic "scene" changes with time, location, or surrounding environment. WaveNet has no descriptor for a scene and considers same acoustic events in different scenes to be identical as illustrated in Fig.1. A laughter in the daytime is normal whereas one in the late night is an anomaly to be alarmed. It indicates that an acoustic event in an environment with a time-varying scene cannot be detected by WaveNet. In such an environment, dedicated acoustic event models for different times need to be learned by training, however, computational cost and the memory size will be problems in reality. Komatsu *et al.* proposed an anomaly detection assuming periodicity over an entire day [6]. Nevertheless, it cannot be applied to general changing scenes. Suppose

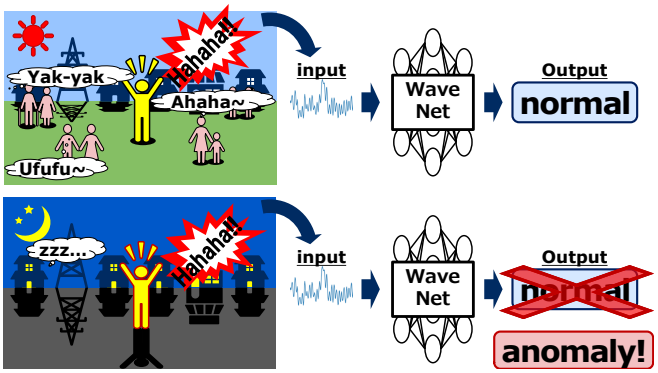

Fig. 1. Same acoustic-event in different scenes have different meaning.

a stadium. At the same time and location, the acoustic scene is significantly different if there is an American football match, a Taylor Swift [7] concert or nothing. An anomalous acoustic-event detection incorporating the acoustic scene is necessary.

For scene description, i-vectors are widely used in acoustic scene clasification [8, 9, 10, 11]. I-vectors, which are features originally developed for speaker verification, are obtained by factor analysis of acoustic feature distribution difference between a universal background model (UBM) and a model for each utterance. They provide high accuracy in speaker verification and are popular in applications such as music genre classification [12], and language classification [13]. I-vectors are promising for scene description in anomalous acoustic-event detection.

This paper proposes an anomalous acoustic-event detection based on conditional WaveNet and i-vector. WaveNet provides scene-independent anomaly detection and i-vector complements the lack of scene description. The next section presents details of the new method and in Section 3 evaluation results demonstrate the validity of the proposed method.

2. THE PROPOSED METHOD

An overview of the proposed method, separated into training and detection part, is shown in Fig. 2. In the training part, a signal waveform $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ is divided into 25-ms-blocks with 96% overlap to calculate a 40 dimensional log mel spectrum and 20 dimensional MFCCs (Mel-Frequency Cepstrum Coefficients). MFCCs are used to i-vector extractor training, as described in Section 2.2, and 60-dimentional i-vector is extracted from MFCCs

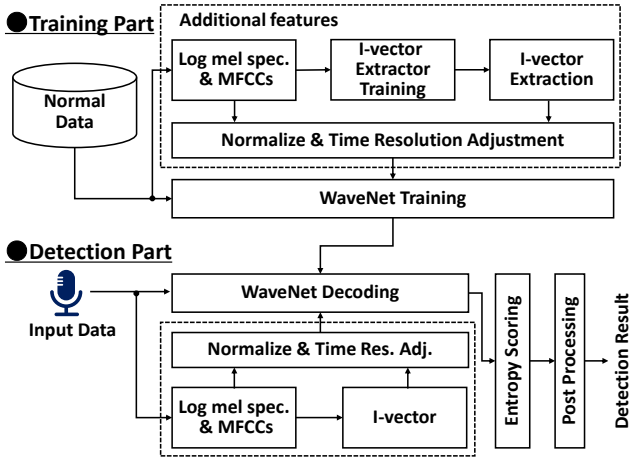

Fig. 2. Blockdiagram of the proposed method.

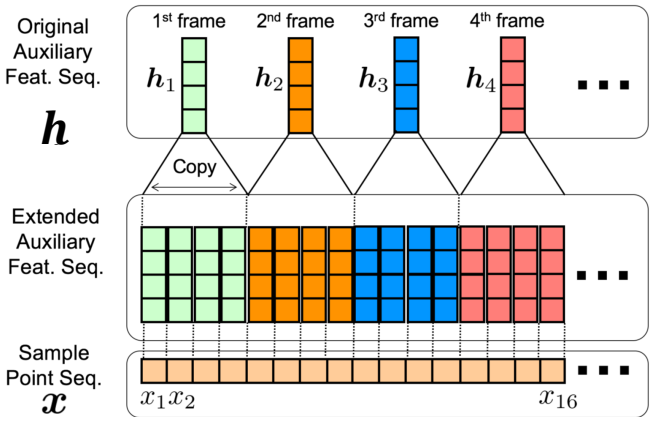

Fig. 3. Time-resolution adjustment

within a 30 second long segment. The proposed method employs an i-vector to describe a scene of a segment of each input signal waveform. An extracted 40 dimensional log mel spectrum and a 60 dimensional i-vector are concatenated and used as an auxiliary feature $\mathbf{h}$ for WaveNet. The statistics of $\mathbf{h}$ are calculated over training data to perform global normalization, making a mean and a variance of each dimension of the features 0 and 1, respectively. A time-resolution adjustment procedure shown in Fig. 3 is performed to ensure that the time resolution of the features is same as that of the waveform signal. The waveform signal $\mathbf{x}$ is quantized and then converted into a sequence of one-hot vectors. Finally, WaveNet is trained with the sequence and the features, as described in Section 2.1.

In the detection part, as in the training part, the log mel spectrum and the i-vector features are extracted from the input waveform signal and normalized using the statistics of the training data. The input waveform signal is also quantized and then converted into a sequence of one-hot vectors. WaveNet then calculates a posteriogram (a sequence of posterior distributions) with the sequence and the features. Note that, since WaveNet is used as a finite impulse response (FIR) filter as explained in Section 2.1, this process is much faster than the autoregressive generation process of the original WaveNet. Next, an entropy of each posterior distribution is calculated over the posteriogram. We then perform thresholding for a sequence of entropies to detect anomalies and three kinds of post-processing are performed to smooth the detection result, as described in Section 2.3.

2.1. WaveNet

To directly model acoustic patterns in the time domain, the proposed method uses WaveNet [5], which is a generative model based on a convolutional neural network. The conditional probability of a waveform $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ given the auxiliary features $\mathbf{h}$ is factorized as a product of conditional probabilities as follows:

$$p(\mathbf{x}|\mathbf{h}) = \prod_{n=1}^{N} p(x_n|x_1, x_2, \ldots, x_{n-1}, \mathbf{h}). \quad (1)$$

For the auxiliary feature $\mathbf{h}$, the log mel spectrum is used in the conventional method [4]. In addition to the log mel spectrum, the proposed method employs i-vector for $\mathbf{h}$ to describe the scene where the input waveform $\mathbf{x}$ is occurred. WaveNet approximates the above conditional probability by canceling the effect of past samples of a finite length as follows:

$$p(x_n|x_1,x_2,\ldots,x_{n-1},\mathbf{h}) \simeq p(x_n|x_{n-R-1},x_{n-R},\ldots,x_{n-1},\mathbf{h}), \quad (2)$$

where R is the number of past samples to take into account, which is known as the *"receptive field"*. In order to generate a waveform directly, it is necessary to secure a very large receptive field, which requires huge computational resources. WaveNet can achieve a large receptive field efficiently through the use of *"dilated causal convolutions"*, which are convolutions with holes, so that the output does not depend on future samples. This architecture not only secures very large receptive fields, but also significantly reduces computational cost and the number of model parameters. The overall structure of WaveNet is shown in Fig. 4.

WaveNet consists of many residual blocks, each of which consists of 2×1 dilated causal convolutions, a gated activation function and 1×1 convolutions. The gated activation function is formulated as follows:

$$\mathbf{z} = \tanh(\mathbf{W}_{f,k} * \mathbf{x} + \mathbf{V}_{f,k} * f(\mathbf{h})) \odot \sigma(\mathbf{W}_{g,k} * \mathbf{x} + \mathbf{V}_{g,k} * f(\mathbf{h})), \quad (3)$$

where W and V are trainable convolution filters, $W * \mathbf{x}$ represents a dilated causal convolution, $V * f(\mathbf{h})$ represents a 1×1 convolution, $\odot$ represents element-wise multiplication, σ represents a sigmoid activation function, subscript k is the layer index, subscripts f and g represent the "filter" and "gate", respectively, and $f(\cdot)$ represents the function which transforms features $\mathbf{h}$ to have the same time resolution as the input waveform. The waveform signal is quantized into 8 bits by μ-law algorithm [14] and converted into a sequence of 256 dimensional (= 8 bits) one-hot vectors.

Upon training, WaveNet is used as an FIR filter, i.e., it predicts a future sample x_t from observed samples $x_{t-R-1:t-1}$. WaveNet is optimized through back-propagation using the following cross-entropy objective function:

$$E(\mathbf{\Theta}) = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c} \log \hat{y}_{t,c} \quad (4)$$

where $\mathbf{y}_t = \{y_{t,1}, y_{t,2}, \ldots, y_{t,C}\}$ represents the one-hot vector of the target quantized waveform signal, $\hat{\mathbf{y}}_t = \{\hat{y}_{t,1}, \hat{y}_{t,2}, \ldots, \hat{y}_{t,C}\}$

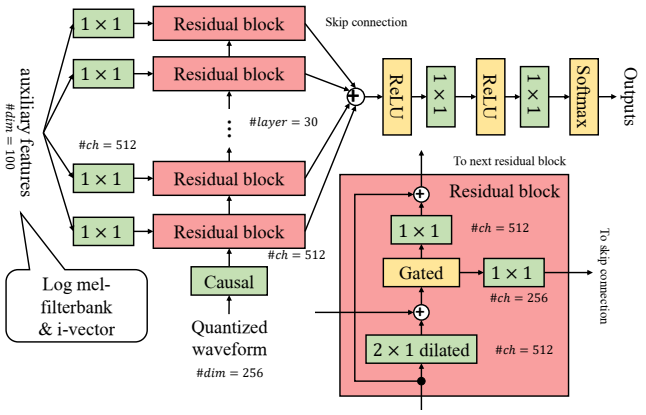

Fig. 4. WaveNet

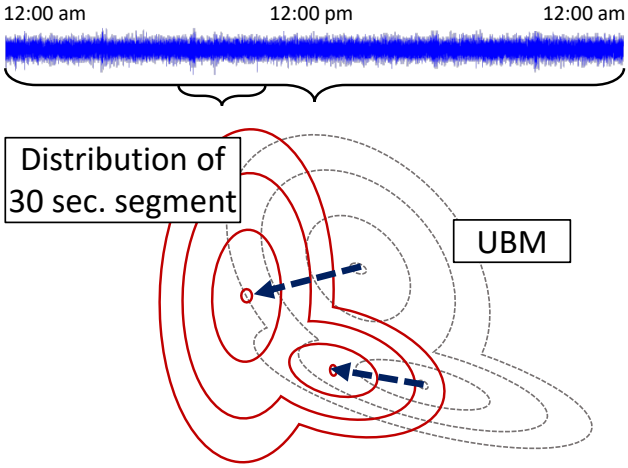

Fig. 5. An i-vector is a low-dimensional representation of two blue arrows which represent a difference between UBM and a segment-specific model.

represents the posterior distribution of the amplitude class, t and i represent the index of the waveform samples and their amplitude class, respectively, T and C represent the number of waveform samples and number of amplitude classes, respectively.

2.2. I-vector

An i-vector is employed as a new feature for scene description that is what kind of sounds occurred in the input-signal segment. The proposed method extracts the i-vector from a set of frame-level features, such as MFCCs, in a short segment. Fig.5 illustrates the i-vecotr, which is a low-dimensional representation of the distribution difference between a universal background model (UBM) and a specific model for each sound segment. The UBM is trained on MFCCs of an entire data set. The specific model for each segment is obtained by adapting UBM using MFCCs in the segment, whose length is set as 30 seconds in the proposed method. An i-vector is a low-dimensional representation of blue arrows in Fig. 5 obtained by factor analysis applied to the difference between the UBM and the segment-specific model.

As described in [15], the factor analysis is used to define a new low-dimensional space referred to as a total variability space. In this new space, a given sound segment is represented by a new vector named i-vector. Given a sound segment, its feature vector such as a GMM supervector $\mathbf{M}$ is written as follows,

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{5}$$

where $\mathbf{n}$ is a scene-independent supervector typically taken from a UBM. A total variability matrix $\mathbf{T}$ is a rectangular matrix and $\mathbf{w}$ is an i-vector. The i-vector for a given sound segment u can be obtained by

$$\mathbf{w}_u = \left(\mathbf{I} + \mathbf{T}^\top \mathbf{\Sigma}^{-1} \mathbf{N}(\mathbf{u}) \mathbf{T}\right)^{-1} \mathbf{T}^\top \mathbf{\Sigma}^{-1} \mathbf{F}(\mathbf{u}). \tag{6}$$

This equation loads two statistics $N(u)$ and $F(u)$ which have elements written as follows. When a Gaussian mixture model is used as a UBM, elements of UBM mixture component c are

$$\mathbf{N}_c(\mathbf{u}) = \sum_{t=1}^{L} p(c|\mathbf{u}_t), \tag{7}$$

$$\mathbf{F}_c(\mathbf{u}) = \sum_{t=1}^{L} p(c|\mathbf{u}_t)(\mathbf{u}_t - \mathbf{m}_c), \tag{8}$$

where u_t is a t-th frame of the sound segment u with L frames, and m_c is a mean of the component c. More information about the training procedure for T and i-vector extraction can be found in [15, 16].

2.3. Anomalous Acoustic-Event Detection

For speech synthesis purpose, WaveNet is usually used as an autoregressive filter, i.e., it predicts the future sample $\hat{x}_t$ from predicted samples $\hat{x}_{t-R-1:t-1}$ and repeats the procedure to randomly generate a waveform signal [5]. On the otherhand, in the case of anomaly detection, the observed waveform signals can be directly used for prediction. Therefore, WaveNet is used here as an FIR filter in the same manner as during training.

To detect anomalous acoustic-event, the proposed method estimates an uncertainty of the prediction from the shape of posterior distribution. The shape of posterior distribution of a known sound is sharp while that of an unknown sound is flat. Hence, it is expected that anomalous acoustic-event are identified based on the shape of the prediction. To quantify the uncertainty of prediction, an entropy e of the posterior distribution is calculated as follows:

$$e_t = -\sum_{c=1}^{C} \hat{y}_{t,c} \log_2 \hat{y}_{t,c}. \tag{9}$$

The entropy is calculated over the posteriogram, resulting in the entropy sequence $\mathbf{e} = \{e_1, e_2, \ldots, e_T\}$. Finally, thresholding over the sequence of entropies is performed using the following threshold value:

$$\theta = \mu + \beta\sigma, \tag{10}$$

where θ represents the threshold value, μ and σ represent the mean and the standard deviation of the entropy sequence, respectively, and β is a hyper parameter. The value of parameter β is decided through preliminary experiments.

To smooth the detection results, three kinds of post-processing are applied.

1. Apply a median filter with a predetermined filter span;
2. Fill gaps which are shorter than a predetermined length;

Table 1. *The detail of auxiliary features*

Auxiliary features for WaveNet	40-dim. log mel spectrum 60-dim. i-vector
Frame size for log-mel spectrum	25ms
Frame-lebel feature for i-vector	MFCCs, Δ, ΔΔ
Order of MFCC	20
Segment length for i-vector	30 seconds,
UBM components	256
Order of I-vector	400

3. Remove events whose duration is shorter than a predetermined length.

The parameters for post-processing are decided through preliminary experiments.

3. EVALUATIONS BY COMPUTER SIMULATION

Experimental evaluation using two-weeks of audio data recoded at a subway station was conducted. Data from the first week was used as training data, and the rest of the data are used as evaluation data. The continuous audio data was divided into 30 second pieces and added anomalous sounds to each piece of evaluation data. The added anomalous sounds included the sound of glass breaking, screaming, and growling, and are selected from the Sound Ideas Series 6000 General Sound Effects Library [17]. Each sound was added at random temporal positions with three signal-to-noise ratios (SNRs): 0 dB, 10 dB, and 20 dB. Evaluation was conducted in two regimes, event-based metric (onset only) and segment-based evaluation metric, where the F1-score was utilized as the evaluation criteria (see [18] for more details). The detail of the auxiliary features is shown in Table 1.

To compare the performance of our proposed method, we used the following methods:

1. **Auto-encoder (AE)**
2. **Auto-regressive LSTM (AR-LSTM)**
3. **Bidirectional LSTM auto-encoder (BLSTM-AE)**

   These networks consist of 3 hidden layers with 256 hidden units, and their inputs are 40 dimensional log mel spectrum, which are extracted with 25 ms window and a 10 ms shift. All of these networks were optimized using Adam [19] under the objective function based on the root mean squared error.
4. **WaveNet without i-vector** [4]

   This method is equivalent to the the proposed method without i-vector.

Thresholding and post-processing were the same as our proposed method. All networks were trained using the open source toolkit Keras [20] and TensorFlow [21] with a single GPU (Nvidia GTX 1080Ti).

The experimental results are shown in Figs 6 and 7. The results show that the proposed method outperforms the conventional methods for both event-based and segment-based metrics. Thus, we can confirm the effectiveness of the proposed method. The proposed method achieved as much as 15 pt higher F-measure than LSTM and AE in the event-based results. The difference in F-measure by methods with and without i-vector turned out to be 1.5 pt.

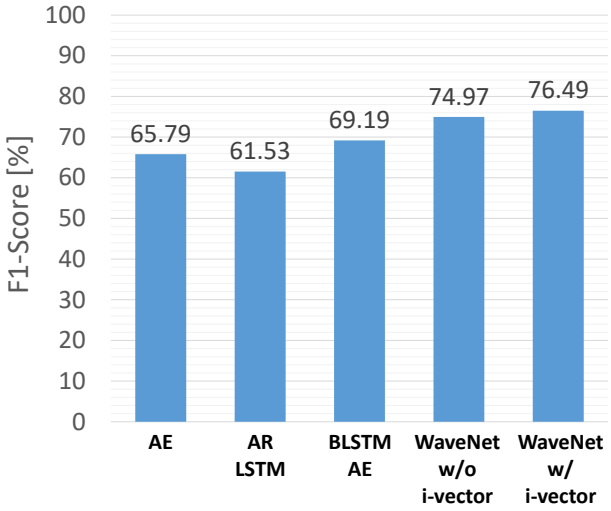

Fig. 6. Event-based experimental results.

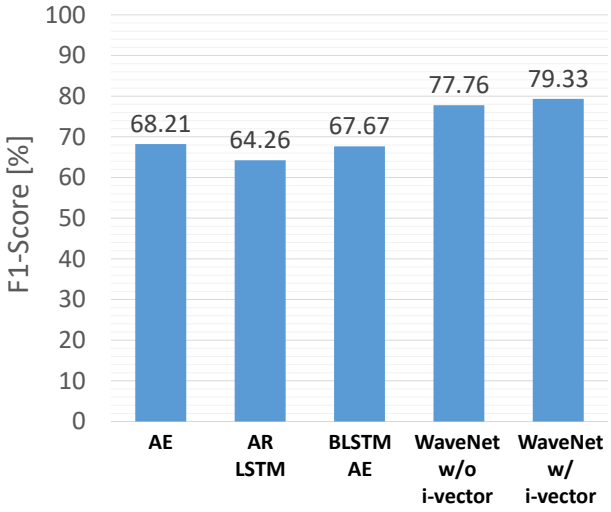

Fig. 7. Segment-based experimental results.

4. CONCLUSION

This paper has proposed a scene-dependent anomalous acoustic-event detection based on WaveNet and i-vector. The WaveNet builds normal acoustic event models by exhaustive learning of time-domain signals in the public space to provide scene-independent anomaly detection. I-vectors are used as additional features to describe acoustic scenes, where the input signals are observed, to complement the WaveNet. The proposed method can detect anomalous acoustic-events in environments whose acoustic scenes vary depending on time, location, and surrounding environment. Evaluations with data recorded from the real environment have demonstrated that the proposed method achieved as much as 15 pt higher F-measure than LSTM and AE. The difference in F-measure by the WaveNet with and without i-vector turned out to be 1.5 pt.

5. REFERENCES

[1] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM, 2014, p. 4.

[2] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 1996–2000.

[3] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc*. Presses universitaires de Louvain, 2015, p. 89.

[4] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on wavenet," *in Proc. EUSIPCO*, 2018, pp. 2508–2512.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[6] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," in *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 376–380.

[7] T. Swift, "The official website of Taylor Swift," `https://www.taylorswift.com/`, [Accessed: 1- Nov- 2018].

[8] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," in *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 126–130.

[9] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," *in Proc. EUSIPCO*, 2017, pp. 2749–2753.

[10] B. Elizalde, H. Lei, G. Friedland, and N. Peters, "An i-vector based approach for audio scene detection," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2013)*, 2013.

[11] B. Elizalde, H. Lei, and G. Friedland, "An i-vector representation of acoustic environments for audio-based video event detection on user generated content," in *Proc. IEEE International Symposium on Multimedia (ISM)*, IEEE, 2013, pp. 114–117.

[12] H. Eghbal-Zadeh, B. Lehner, M. Schedl, and G. Widmer, "I-vectors for timbre-based music similarity and music artist classification.," in *ISMIR*, 2015, pp. 554–560.

[13] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[14] G. Recommendation, "Pulse code modulation (PCM) of voice frequencies," *ITU*, 1988.

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, vol. 14, pp. 28–29, 2005.

[17] "Series 6000 general sound effects library," `http://www.sound-ideas.com/sound-effects/series-6000-sound-effects-library.html`, [Accessed: 1- Nov- 2018]

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.

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

[20] F. Chollet et al., "Keras," `https://github.com/fchollet/keras`, [Accessed: 1- Nov- 2018].

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "Tensorflow: A system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.