Anomalous Sound Event Detection Based on WaveNet

Tomoki Hayashi*, Tatsuya Komatsu†, Reishi Kondo†, Tomoki Toda‡, Kazuya Takeda*
*Department of Information Science
Nagoya University, Nagoya, Japan 492–8411
hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp,takeda@nagoya-u.ac.jp
†Data Science Research Laboratories
NEC Corporation, Kawasaki, Japan 211–8666
t-komatsu@ew.jp.nec.com,kondoh@ct.jp.nec.com
‡Information Technology Center
Nagoya University, Nagoya, Japan 492–8411
tomoki@icts.nagoya-u.ac.jp

***Abstract*—This paper proposes a new method of anomalous sound event detection for use in public spaces. The proposed method utilizes WaveNet, a generative model based on a convolutional neural network, to model in the time domain the various acoustic patterns which occur in public spaces. When the model detects unknown acoustic patterns, they are identified as anomalous sound events. WaveNet has been used to precisely model a waveform signal and to directly generate it using random sampling in generation tasks, such as speech synthesis. On the other hand, our proposed method uses WaveNet as a predictor rather than a generator to detect waveform segments causing large prediction errors as unknown acoustic patterns. Because WaveNet is capable of modeling detailed temporal structures, such as phase information, of the waveform signals, the proposed method is expected to detect anomalous sound events more accurately than conventional methods based on reconstruction errors of acoustic features. To evaluate the performance of the proposed method, we conduct an experimental evaluation using a real-world dataset recorded in a subway station. We compare the proposed method with the conventional feature-based methods such as an auto-encoder and a long short-term memory network. Experimental results demonstrate that the proposed method outperforms the conventional methods and that the prediction errors of WaveNet can be effectively used as a good metric for unsupervised anomalous detection.**

***Index Terms*—anomaly detection, anomalous sound event detection, WaveNet, neural network**

I. Introduction

In response to the rising number of terrorism incidents, demands for better public safety have been increasing all over the world. To meet these demands, video-based monitoring systems have been developed which make it possible to automatically detect suspicious people or objects [1], [2], as well as sound-based security systems which can automatically detect anomalous sound events such as glass breaking [3]–[5]. Video-based systems have proven to be effective, however due to blind spots and limited installation of cameras it is difficult for these systems to monitor an entire area. On the other hand, sound-based systems have been attracting increased attention because they have no blind spots, and microphones cheaper are easier to install than cameras. Therefore, sound-based systems can complement video-based systems by covering camera blind spots. Furthermore, a combination of sound-based and video-based systems is likely to result in more intelligent monitoring systems.

The key technology of sound-based monitoring system can be divided into two types; supervised and unsupervised approaches. Supervised approaches use manually labeled data, and include acoustic scene classification (ASC) [6], [7] and acoustic event detection (AED) [8]–[10] methods. Here, scenes represent the environment which the audio segments are recorded, and sound events represent sounds related to human behaviors or moving of objects. The task of ASC is to classify long-term audio segments into pre-defined scenes, while the task of AED is to identify the start and end times of pre-defined sound events to label them. These technologies make it possible to understand an environment and detect various types of sounds, but they require the pre-definition of all of the possible scenes and sound events, and it is difficult to collect and label so much of this type sound data.

Unsupervised approaches, on the other hand, do not require manually labeled data, so they are less costly. One unsupervised approach is change point detection [11]–[13], which compares a model of the current time with that of a previous time to calculate a dissimilarity score, and then identifies highly dissimilar comparison results as anomalies. However, in the case of public spaces, the sounds which can occur are highly variable and non-stationary, and therefore, the detected change points are not always related to anomalies that are of concern (e.g., the sound of the departure of the train). Another unsupervised approach is outlier detection [14]–[16], which models an environment's "*normal*" sound patterns, and then detects patterns which do not correspond to the normal model and identifies them as anomalies. Note that the "*normal*" patterns are patterns which appeared in the training data. Typically, a Gaussian mixture model (GMM) or one-class support vector machine (SVM) with acoustic features such as mel-frequency cepstrum coefficients (MFCCs) is used [17], [18]. With the recent advances in deep learning, neural net-

ISBN 978-90-827970-1-5

work based methods have been attracting attention [19]–[21]. These methods train an auto-encoder (AE) or a long short-term memory recurrent neural network (LSTM-RNN) with only normal scene data. While the AE encodes the inputs as latent features and then decodes them as the original inputs, LSTM-RNN predicts the next input from the previous input sequence. Using a trained model, reconstruction errors between observations and the predictions are calculated and the high error patterns are identified as anomalies. Although these methods have achieved high performance, it is difficult to directly model the acoustic patterns in the time domain due to their highly non-stationary nature and their high sampling rate so they typically use feature vectors extracted from audio signals.

In this study, we propose a new method of anomalous sound event detection method in public spaces which utilize WaveNet [22]–[24], a generative model based on a convolutional neural network, to directly model the various acoustic patterns occurring in public spaces in the time domain. Based on this model, unknown acoustic patterns are identified as anomalous sound events. WaveNet has been used to precisely model a waveform signal and to directly generate it using random sampling in generation tasks, such as speech synthesis. On the other hand, our proposed method uses WaveNet as a predictor rather than a generator to detect waveform segments causing large prediction errors as unknown acoustic patterns. Because WaveNet is capable of modeling detailed temporal structures, such as phase information, of the waveform signals, the proposed method is expected to detect anomalous sound events more accurately than conventional methods based on reconstruction errors of acoustic features. To evaluate the performance of the proposed method, we conduct an experimental evaluation using a real-world dataset recorded in a subway station. We compare the proposed method with the conventional feature-based methods such as an auto-encoder and a long short-term memory network. Experimental results demonstrate that the proposed method outperforms the conventional methods and that the prediction errors of WaveNet can be effectively utilized as a good metric for unsupervised anomalous detection.

II. Anomalous Sound Event Detection System Based on WaveNet

A. System overview

An overview of our proposed system, separated into training and test phases, is shown in Fig. 1. In the training phase, the waveform signal is divided into 25 ms windows with 4 % overlap to calculate a 40 dimensional log mel filter bank. Note that we use a large amount of overlap in order to directly model the waveform signal in the time domain. The statistics of the extracted features are calculated over training data to perform global normalization, making the mean and variance of each dimension of the features 0 and 1, respectively. The time resolution adjustment procedure shown in Fig. 2 is performed to make the time resolution of the features the same as the waveform signal. The waveform signal

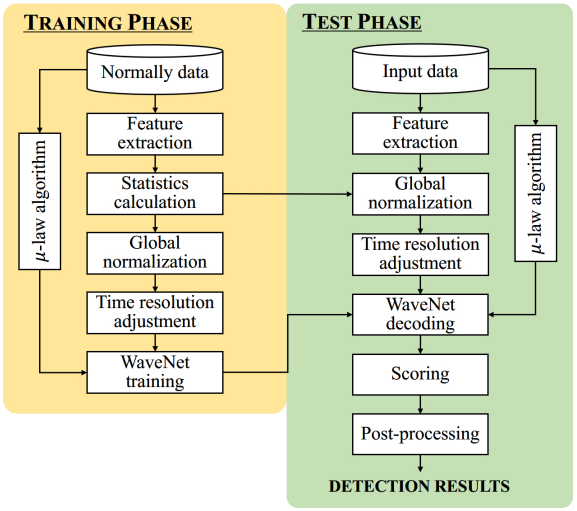


Fig. 1: *System overview.*

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 II-B.

In the test phase, as in the training phase, features are calculated 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 posterigram (a sequence of posteriors) with the sequence and the features. Note that since WaveNet is used as a finite impulse response (FIR) filter, as explained in Section II-B, decoding is much faster than when using the original WaveNet decoding process. Next, the entropy of each posterior is calculated over the posteriogram. We then perform thresholding for the sequence of entropies to detect anomalies, as described in Section II-C. Finally, three kinds of post-processing are performed to smooth the detection result, as described in Section II-D.

B. WaveNet

To directly model acoustic patterns in the time domain, we use WaveNet [22], 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

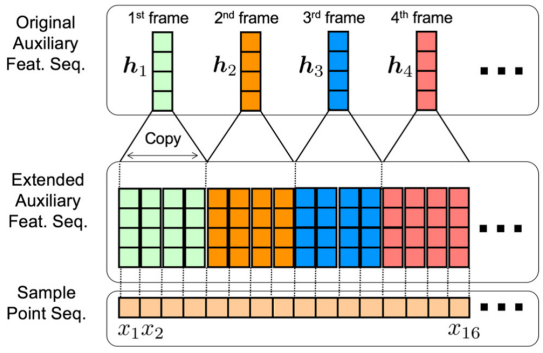


Fig. 2: *The procedure of time resolution adjustment.*

ISBN 978-90-827970-1-5 © EURASIP 2018

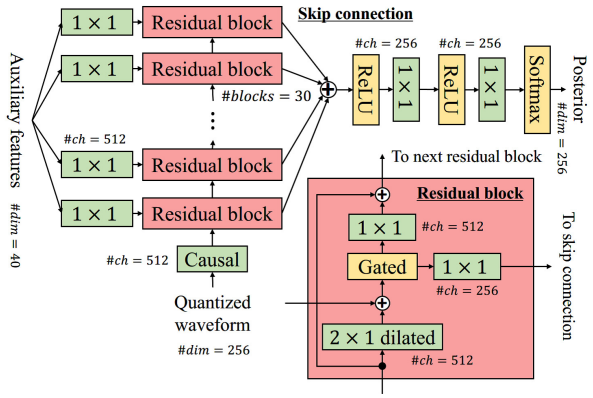


Fig. 3: *Overview of WaveNet's structure.*

$\mathbf{h}$ is factorized as a product of conditional probabilities as follows:

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

WaveNet approximates the conditional probability above by canceling the effect of past samples of a finite length as follows:

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

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 this task more 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 can not only secure 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. 3. 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:

$$\begin{aligned}\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * f(\mathbf{h})) \odot \\ \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * f(\mathbf{h})),\end{aligned} \quad (3)$$

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 [25] and converted into a sequence of 256 dimensional (= 8 bits) one-hot vectors.

During 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, $\hat{\mathbf{y}}_t = \{\hat{y}_{t,1}, \hat{y}_{t,2}, \ldots, \hat{y}_{t,C}\}$ represents the posterior 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.

When decoding, 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 [22]. However, in the case of anomaly detection, we can use all of the observed waveform signals for decoding. Therefore, WaveNet is used here as an FIR filter in the same manner as during training.

C. Scoring

To detect anomalous sound events, we focus on an uncertainty of the prediction. Examples of the posteriors of known and unknown sounds are shown in Fig. 4. These figures indicate that the shape of posterior of a known sound is sharp while that of an unknown sound is flat. Hence, it is expected that we can identify unknown sounds as anomalous sound events based on an uncertainty of the prediction.

To quantify the uncertainty of the prediction, an entropy e of the posterior is calculated as follows:

$$e_t = -\sum_{c=1}^{C} \hat{y}_{t,c} \log_2 \hat{y}_{t,c}. \quad (5)$$

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

$$\theta = \mu + \beta\sigma, \quad (6)$$

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 hyperparameter. The value of parameter β is decided through preliminary experiments.

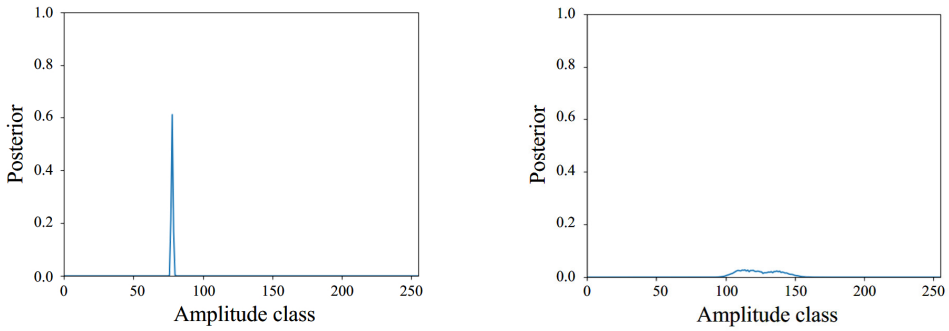
An example of a sequence of entropies is shown in Fig. 5. We can see that entropy increases in the parts of the sequence corresponding to the unknown sound.

D. Post-processing

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

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) *Posterior of a known sound* (b) *Posterior of a unknown sound*

Fig. 4: *Examples of the WaveNet posterior for known and unknown sounds.*

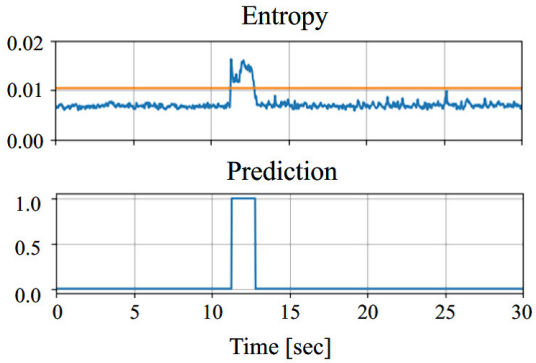
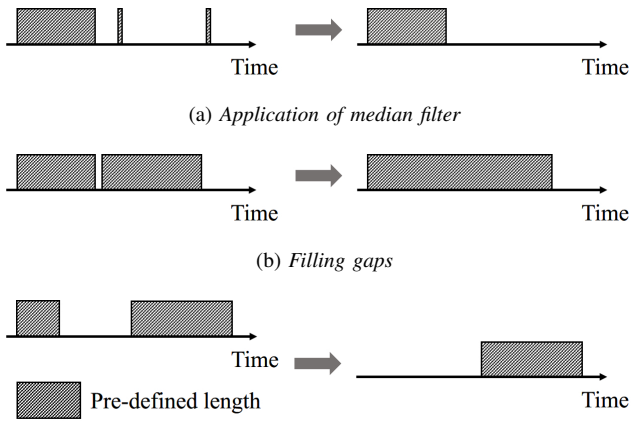


Fig. 5: *Example of a sequence of entropies. The top figure represents a sequence of entropies and their threshold values, and bottom figure represents the binarized detection results.*



(a) *Application of median filter*

(b) *Filling gaps*

(c) *Removal of short duration events*

Fig. 6: *Outline of each post-processing step.*

An outline of each post-processing step is illustrated in Fig. 6. The parameters for post-processing are decided through preliminary experiments.

III. Experimental Evaluation

We evaluated our proposed methods using two-weeks of audio data recoded at a subway station. Data from the first week was used as training data, and the rest of the data are used as evaluation data. We divided the continuous audio data into 30 seconds 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 [26]. 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 is utilized as the evaluation criteria (see [27] for more details).

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

- Auto-encoder (AE),
- Auto-regressive LSTM (AR-LSTM),
- Bidirectional LSTM auto-encoder (BLSTM-AE).

These networks consisted of 3 hidden layers with 256 hidden units, and the inputs were 40 dimensional log mel filter bank features, which were extracted with 25 ms window and a 10 ms shift. All of these networks were optimized using Adam [28] under the objective function based on the root mean squared error. Thresholding and post-processing were the same as our proposed method. All networks were trained using the open source toolkit Keras [29] and TensorFlow [30] with a single GPU (Nvidia GTX 1080Ti).

Our experimental results are shown in Table I, where EB and SB represent event-based metric and segment-based metric, respectively. The result shows our proposed method outperforms the conventional methods for both event-based and segment-based metrics, thus we can confirm the effectiveness of our proposed method. An example of the detection results are shown in Fig. 7. We can see that our proposed method can detect anomalous sound events even if they are difficult to distinguish through the spectrogram.

TABLE I: *Experimental results.*

Method	EB F1-score [%]	SB F1-score [%]
AE	65.8	68.2
AR-LSTM	61.5	64.3
BLSTM-AE	69.2	67.7
WaveNet	75.0	77.8

ISBN 978-90-827970-1-5 © EURASIP 2018

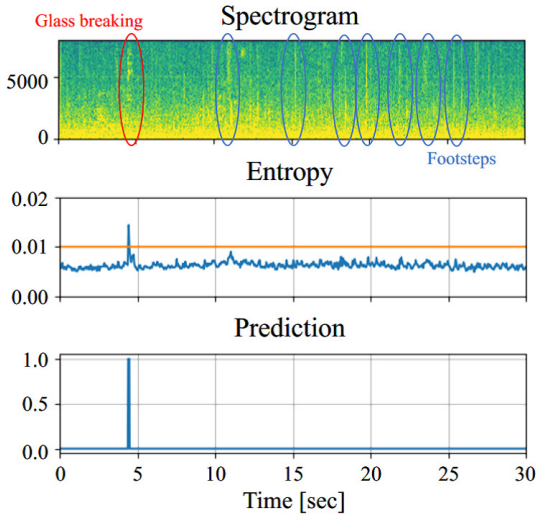


Fig. 7: *Example of detection results. In the spectrogram, first pulse represents a glass breaking sound, while the others represent footsteps of a woman wearing high heels.*

IV. Conclusion

In this paper, we proposed a new method of anomalous sound event detection in public spaces which utilized WaveNet to model in the time domain the various acoustic patterns which occur in public spaces. Based on our model, unknown acoustic patterns are identified as anomaly sounds. The use of WaveNet allows the modeling of detailed temporal structures of acoustic patterns, such as phase information, that occur in public spaces. Our experimental results, when using a real-world dataset in a public space, demonstrated that the proposed method outperformed the conventional methods and that the prediction errors of WaveNet can be effectively used as a good metric for unsupervised anomalous detection.

In future works, we will investigate the effect of auxiliary features, improve the thresholding process and apply our method to another dataset.

References

[1] S. C. Lee and R. Nevatia, "Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system," *Machine vision and applications*, vol. 25, no. 1, pp. 133–143, 2014.

[2] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," *Expert systems with Applications*, vol. 42, no. 21, pp. 7991–8005, 2015.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International conference on*. IEEE, 2005, pp. 1306–1309.

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

[5] Y. Chung, S. Oh, J. Lee, D. Park, H.-H. Chang, and S. Kim, "Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems," *Sensors*, vol. 13, no. 10, pp. 12 929–12 942, 2013.

[6] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[7] Y. Han and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE2017 Challenge, Tech. Rep., September 2017.

[8] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, September 2016, pp. 45–49.

[9] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2059–2070, 2017.

[10] S. Adavanne, A. Politis, and T. Virtanen, "Multichannel sound event detection using 3d convolutional neural networks for learning inter-channel features," *arXiv preprint arXiv:1801.09522*, 2018.

[11] M. Basseville, I. V. Nikiforov *et al.*, *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[12] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," *Neural Networks*, vol. 43, pp. 72–83, 2013.

[13] Y. Ono, Y. Onishi, T. Koshinaka, S. Takata, and O. Hoshuyama, "Anomaly detection of motors with feature emphasis using only normal sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2800–2804.

[14] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," 1995.

[15] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 1. IEEE, 2000, pp. 452–455.

[16] M. Markou and S. Singh, "Novelty detection: a reviewpart 1: statistical approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[17] D. W. Scott, "Outlier detection and clustering by partial mixture modeling," in *COMPSTAT 2004Proceedings in Computational Statistics*. Springer, 2004, pp. 453–464.

[18] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 3. IEEE, 2003, pp. 1741–1745.

[19] 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.

[20] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1996–2000.

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

[22] 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.

[23] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *INTERSPEECH*, 2017.

[24] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

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

[26] "Series 6000 general sound effects library," http://www.sound-ideas.com/sound-effects/series-6000-sound-effects-library.html.

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

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

[29] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[30] 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*, vol. 16, 2016, pp. 265–283.

ISBN 978-90-827970-1-5 © EURASIP 2018