Computational Modelling Auditory Awareness

Yu Su[1,2], Jingyu Wang[1], Ke Zhang[1], Kurosh Madani[2] and Xianyu Wang[1]
[1]*Northwestern Polytechnical University (NPU), 127 Youyi Xilu, Xi'an 710072, Shaanxi, China*
[2]*Universit Paris-Est, Signals, Images, and Intelligent Systems Laboratory (LISSI / EA 3956), University Paris-Est Creteil, Senart-FB Institute of Technology, 36-37 rue Charpak, 77127 Lieusaint, France*

Keywords: Auditory Awareness, Saliency Detection, Deep Learning.

Abstract: Research in the human voice and environment sound recognition has been well studied during the past decades. Nowadays, modeling auditory awareness has received more and more attention. Its basic concept is to imitate the human auditory system to give artificial intelligence the auditory perception ability. In order to successfully mimic human auditory mechanism, several models have been proposed in the past decades. In view of deep learning (DL) algorithms has better classification performance than conventional approaches (such as GMM and HMM), the latest research works mainly focused on building auditory awareness models based on deep architectures. In this survey, we will offer a quality and compendious survey on recent auditory awareness models and development trend. This article includes three parts: i) classical auditory saliency detection method and developments during the past decades, ii) the application of machine learning in ASD. Finally, summarizing comments and development trends in this filed will be given.

1 INTRODUCTION

Auditory cognition is an important part of the human perception system which helps human to perceive the surrounding environment accurately. Neurobiologist generally believes that the saliency-based selective attention mechanism could be the fastest way for humans to make responses to prominent stimulus received from surrounding environment. Hence, the bio-inspired saliency detection approaches could be regarded as a feasible way for computational modeling the human awareness for artificial intelligence to percept the surrounding environment precisely.

Auditory saliency detection (ASD) is one of the most important research fields of realizing machine awareness which aims at detecting the abnormal or conspicuous sound events in the surrounding environment. For example, when a rescue robot encounters an emergency, such as an explosion, tremendous amounts of salient stimulus are received simultaneously by the sensors of both visual and auditory channels. However, if the image of target need for rescue is blocked by some objects in the field of view or the image quality is not good, the related sound signals to this incident could play a pivotal role in the process of environmental perception for intelligent awareness.

A considerable amount of approaches has been presented to detect the auditory saliency property from sound signals over the past decades. Almost all the auditory saliency-driven awareness models are based on the idea of auditory saliency map (ASM). To be specific, ASM is derived from the concept of the visual salient map, which is first proposed by (Itti et al., 1998) where to reveal the saliency of a sound signal in a two-dimension image and make the saliency of sound clips more intuitive to researchers.

Recently, a growing number of researchers have begun to apply neural networks for sound events classification . At present, the deep architectures have conquered the field of image and speech recognition, but the use of ASD still falls behind. Researchers believe that in the next few years, deep neural networks based methods will provide a better solution for improving the performance of ASD and make it possible to put into practical applications.

In this work, we will present an elucidatory survey on recent developments and indicate the future trends in ASD. The rest of this paper is organized as follows. We will first introduce three classical ASD models and improved techniques in section 2. Then, a survey on the application of various deep models in acoustic classification will be presented in section 3. Finally, concludes the main trends of ASD and outlook for the future will be given in section 4.

Su, Y., Wang, J., Zhang, K., Madani, K. and Wang, X.
Computational Modelling Auditory Awareness.
DOI: 10.5220/0006925401600167
In *Proceedings of the 10th International Joint Conference on Computational Intelligence (IJCCI 2018)*, pages 160-167
ISBN: 978-989-758-327-8

2 AUDITORY SALIENCY MAP

Before discussing various auditory saliency detection models, we first depict three classical models which are the first proposed computational models for ASD.

2.1 Classical ASD Models

The reason why the sound saliency can be transformed into visual representation is that visual and auditory perception channels have perceptual correlations in high-level perceptual processing. (Li et al., 2008) conducted a theoretical and experimental research on the relevance of audio-visual perception information, point out that there are correlations between images and sounds in human perception system. Moreover, the perception of auditory saliency could be converted into the perception of saliency of the visual channel. This result provides a theoretical basis and a method to realize computational models of ASD.

Based on this result, several ASD models have been proposed for salient sound detection. Kayser first proposed an auditory saliency map (ASM) based on Ittis visual saliency map in (Kayser et al., 2005). Afterward, based on Kaysers work, two improved ASM approaches were proposed by (Kalinli and Narayanan, 2008) and (Duangudom and Anderson, 2007).

The auditory saliency model proposed by Kayser is based on the spectrogram of input sound signals and the auditory saliency maps were obtained by using the center-surround difference operator and linear combination to convert auditory saliency into image saliency for further analyzing. The center-surround mechanism was applied to compare each feature maps obtained at different scales. Then normalize them to facilitate those maps that contain highly prominent peaks. These maps are folded across different scales to yield the saliency maps for each feature. Finally, linear combined the saliency maps of each feature to acquire the auditory saliency map. The structure of ASM which is identical to the visual saliency map is shown in Figure 1.

In order to improve the detection accuracy of ASM, the model presented in (Kalinli and Narayanan, 2008) added the characteristics of orientation and pitch as new sound features. The information of orientation is extracted from the spectrum at angles of 45 degrees and 135 degrees. Orientation features simulate the auditory neuron's response to dynamic ripples in the primary auditory cortex. Since the pitch is the most basic element of sound, therefore, Kalinli also considered extracting the pitch as an auditory feature.

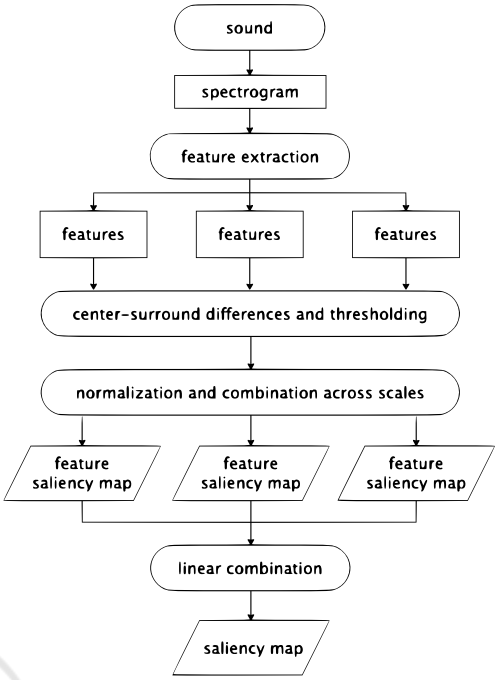
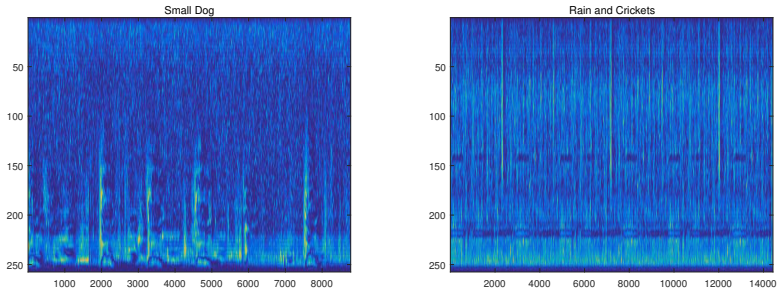


Figure 1: Kaysers Auditory Saliency Map.

(Duangudom and Anderson, 2007) proposed the third classical ASM in which the time-frequency receiver domain model and adaptive suppression were used to provide the final auditory saliency map. The model presented in this paper is basically the same as Kayser's auditory saliency map, but this model extracted three different kinds of features: global energy, time modulation, spectral modulation and high temporal-spectral modulation.

These models have been proved to be effective, however, they all use simple sounds or speeches to test their models. Kayser and Duangudom tested their models on simple sounds with noises. Results show that Kaysers model matches psychoacoustic experiment results of human saliency perception. Duangudom uses the same type of sounds but his model gives lower achievement than Kaysers result. The model proposed by Kalinli has been tested only on prominent syllables in speech. As the environment sound does not contain a definite pitch and its instability characteristics, the sound saliency detection accuracy of these models are doubtful. Figure 2 shows the real environment sound (dog barging and crickets with rain) saliency detection result of Kayser's model. The yellow part of the ASM represents the conspicuous part of a sound clip while the blue part represents the background noises. It is shown from the results that the accuracy and robustness of Kayser's approach will decrease sharply when the background noise is relatively strong and overlaps the salient sounds.



(a) (b)

Figure 2: The auditory saliency map of Kaysers model.

The other two classical models which based on the similar detection principle could also confront the same limitation.

2.2 Improved ASD Models

In order to solve the drawbacks of applying only auditory saliency map for extraction features, many researches have proposed several new ASD models during the past decades. Based on the theory that the auditory saliency of a sound event is obtained by measuring the difference in the time domain between the sound and its surrounding sounds, (Kaya and Elhilali, 2012) proposed a novel model which only defined over time. Unlike the previously mentioned three auditory saliency maps which transform the input sound events into the spectrogram at first, this auditory saliency map treats the input signals as a one-dimensional temporal input. The model uses rich high-dimensional feature space to define auditory events and each auditory dimension was processed across multiple scales but only considers the temporal saliency of the sound.

Research shows that the three classical models only performed well when detecting the saliency of short-term sound signals. For overcoming this drawback, (Botteldooren and De Coensel, 2009) proposed an auditory saliency map for detecting the saliency in long-term sound signals. This model first formed a sonic environment by 1/3 octave band spectrograms of different sound signals and implemented the method proposed in (Zwicker and Fastl, 2013) for calculating a simplified cochlea. Considering the energy masking effects, for one sound source, all the other sound sources can be considered as the background noise. Thus, the specific loudness versus time map contains only non-zero values for those time and space portions of each source, which are not obscured by the sum of all other sources. Then, the same approach for extracting the multi-scale feature maps and the process of forming the final ASM proposed in classical approach mentioned above is applied to acquire the final saliency map. To provide the essential higher-level cognitive information, while referring to the limited knowledge of the attention mechanism, a simple feedback mechanism is applied to simulate top-down attention mechanisms. In order to validate the efficiency of this model, it has been used to study the ability of typical urban parks to mask road traffic noise. Results showed that it can effectively mask the noise generated by traffics while this model showed how perceptual masking could work in addition to energetic or physiological masking to improve the mental image of a sonic environment.

In order to understand how does human divert our attention in different voices over time, (De Coensel and Botteldooren, 2010) proposed a model for mimicking human top-down and bottom-up attention mechanisms. The model consists of four parts. Each input sounds and their summation are first converted to spectrums through the Gammatone filterbank separately. Then, the spectrogram of signals summation is calculated by Kaysers ASM to obtain the saliency map S and Time-Frequency masks for the spectrograms M_i of each sound resources was calculated at the same time. Afterward, T-F masked spectrograms M_i and auditory saliency map S are combined to yield the saliency score of each acoustic signal via:

$$S_i(t) = \int M_i(t,f)S(t,f)\,df \qquad (1)$$

It should be noticed that this equation hypothesis that saliency combined across frequency channels. After saliency scores calculation, Coensel proposed an attention model which mainly based on a function model proposed by (Knudsen, 2007) while implement the winner-takes-all competition to identify the most salient sound. The model was tested with the sound of traffic and its surroundings while the results proved that this model can mask undesired sound signals.

Inspired by the research results of bird auditory system, a task-related sound locating method through interaural time difference and interaural level diffe-

rence was presented in (Mosadeghzad et al., 2015). After locating the input sounds, the Gammatone filterbank has been used to decompose the left and right inputs in the frequency domain. Then, a saliency score was acquired by multiplying the sum of the peaks with the number of peaks in spectrograms of all the frames. Finally, this saliency-based fusion framework was applied to the iCub robot and tested it in real time to identify the real speaker when two people were talking. Results showed that although the model is still inadequate, however, it is a feasible way to simulate the human cognitive characteristics to some extent.

Almost all the models mentioned above could achieve acceptable or even prominent experiment results, however, the sound data used in their experiment is human voices, simple sound clips (short recordings with no background noises) or A few syllables played by one musical instrument. Meanwhile, the previously introduced auditory saliency models are mainly based on the local spatiotemporal contrast and little global saliency information has been taken into account. Considering the unstable and non-linear characteristics of environment sound, it is difficult to prove that these models are effective enough in salient sound detection tasks when the input is complex environment recordings.

Therefore, some researchers start to consider other methods to successfully detect the auditory saliency in real envirnment. (Schauerte and Stiefelhagen, 2013) proposed a Bayesian Surprise Model-based auditory saliency detection model to lower the computation time. The surprising means the statistical abnormal values based on the signal which is observed before. First, the time-frequency analysis and Bayesian probability frame of the sound signals was analyzed by fixed discrete cosine transform. Then, used the Gamma model and based on the prior experience and the current signal to detect the frequency saliency. Meanwhile, a decay factor was applied to reduce the confidence of the prior experience to ensure the computes efficiency. The mean value of saliency of each frequency was regarded as the final saliency. Finally, the oriented evaluation method was used to quantitative estimate the acquired frequency saliency, to analyze whether the saliency of each frequency was real.

(Wang et al., 2015) proposed a bio-inspired model to detect the salient environment sounds for realizing intelligent perception. This approach first calculated the Short-term Shannon entropy to estimate the background noise level of the input signals over the entire time period. Meanwhile, aiming to reduce the impact from time length on the accuracy of saliency detection, Wang proposed an Inhibition of Return (IOR) based saliency select model. After calculating the Short-term Shannon entropy, the sound signal was divided into several significant sound clips and analyzed the temporal and frequency saliency of each clip. In the temporal domain, the saliency was obtained by analyzing the Mel Frequency Cepstral Coefficient (MFCC) curve. In the frequency domain, the model obtained the frequency saliency through the PSD curve of the sounds. The prominent features of the temporal domain and frequency domain were then filtered by the IOR calculation model. Meanwhile, the image saliency was acquired by calculating the red-green channel of opponent color space on the log scale spectrums of the input sound signals. Finally, each saliency map was combined through a heterogeneous information fusion method to produce the auditory saliency map. In the experiment, the model has been tested with environment sound, except background noise, which contains more than one conspicuous sound. Results showed that the accuracy of this model is much higher than Kaysers model.

To conclude, the conventional ASD models are based on the theory of saliency map while several improved models use the statistical method or bio-inspired approach to detect the prominent sounds. The conventional models which are based on local features have been proved to be effective to some extent, but it has to be noticed that the experimental data are simple recordings. The bio-inspired model presented in (Wang et al., 2015) validated its efficiency with real environment mixtures, however, the Shannon entropy-based approach will cost a lot of computational resources. Meanwhile, almost all the features mentioned in these models are manually selected which could not fully conform to the characteristics of human auditory system and will definitely lose some important information.

3 AUDITORY AWARENESS WITH DEEP LEARNING

Environment sounds always exist with the higher level noise than indoor sound signals and always contains more than one salient sounds. This characteristic increased the difficulty of auditory saliency detection or recognition tasks. Most of the conventional approaches proposed for sounds recognition or salient sounds detection tasks through predefined features such as MFCC and Gammatone Frequency Cepstral Coefficient (GFCC) which compress and reduce details (Chachada and Kuo, 2013), which means some needed details have been neglected. Compared with the conventional sound event detection methods

using manually selected features, neural network models can directly take spectrograms and even raw waveforms as inputs to train the networks while selecting the features itself. Therefore, a growing number of researchers believe that the deep neural architectures which simulate the learning mechanism of human could significantly improve the performance in acoustic detection and classification tasks.

Deep neural networks, also known as deep learning, is part of a broader family of machine learning methods based on learning data representations, it is an algorithm that attempts to abstract high-level data using multiple processing layers consisting of complex structures or multiple nonlinear transformations. Deep learning architectures such as deep neural networks, convolutional neural networks, and recurrent neural networks have been applied to fields including computer vision, speech recognition and audio recognition, which show superior performance than conventional classifiers. The benefit of deep learning is to use unsupervised or semi-supervised feature learning and hierarchical feature extraction algorithms instead of manually acquiring features. This characteristic of DL can effectively reduce the problems caused by predefined features in environment sound classification and recognition.

(Mohamed et al., 2012) applied the generative pre-trained input based deep belief networks (DBN) for acoustic modeling in phone recognition. (Gencoglu et al., 2014) proposed a novel feature-based acoustic events recognition method with the deep neural network (DNN) classifiers. The features consisted of Mel energy features and 4 more frames around it. The pre-trained DNN with 5 hidden layers performed well in the experiment when compared with several traditional approaches (Li et al., 2017). (Espi et al., 2014) proposed a deep learning (DL) based AED model. In this literature, a high-resolution spectrograms patch is treated as the feature. The patch is a window of sound spectrogram frames stacked together and used as the input instead of the predefined features for deep neural networks (DNN).

In recent years, the log-mel features and MFCC features of sounds which is represented by spectrograms are commonly used as inputs to train deep models for sound classification, hence, the convolutional neural networks (CNN), which able to extract higher-level features that are invariant to local spectral and temporal variations, based sound classification approaches have drawn a lot of attention in recent years. A deep model consisting of 2 convolutional layers with max-pooling layer and 2 fully connected layers is presented in (Piczak, 2015), which firstly provide the evidence that CNN can be effectively used in classifying environment sounds. (Salamon and Bello, 2017) proposed a model which comprised of 3 convolutional layers interleaved with 2 pooling operations, followed by 2 fully connected layers. Most works, including the above-mentioned networks, in auditory event recognition follow similar strategies, where signals lasting from tens to hundreds of ms are modeled first. (Takahashi et al., 2018) proposed a 9 layer CNN networks which inspired by VGG Net (Simonyan and Zisserman, 2014) with signals lasting multiple seconds handled as a single input. Furthermore, the labeled data is one of the main barriers to training deep neural networks, and available datasets of environmental sound recordings are still very limited. Therefore, the above-mentioned literature all use data augmentation to acquire more training database. In (Salamon and Bello, 2017), the experiment result also proved that the classifying accuracy of CNN will be better when data augmentation was applied.

Environment sound always exists as mixtures, and there can be multiple sound sources that belong to the same class. Furthermore, monophonic sound event detection systems (conventional auditory events detection approaches) handle the polyphonic data by detecting only the prominent event, resulting in a loss of information in realistic environments (Mesaros et al., 2010). These factors mainly represent the challenges over acoustic event detection in real-life situations. If deep neural models only have the ability to classify and recognize individual sounds, it will greatly limit the practical application of neural networks based artificial intelligence. Hence, polyphonic detection is an essential ability for artificial intelligence to perform well in the complex environment.

Compared with monophonic sound event detection which deals with a single sound event at a time instance, polyphonic sound events detection aims at detecting simultaneously happened and multiple overlapped sound events. (Cakir et al., 2015) proposed a multi-label DNNs for polyphonic sound event detection in realistic environments. The inputs are first annotated frame by frame to mark out the active sound events in each frame. Recurrent neural networks (RNN) could directly use their internal state to process sequences of inputs in audio recordings. RNN has the ability to remember past states which make it could avoid the need for post-processing or smoothing steps. The application of RNN have obtained excellent results on complex audio detection tasks, such as speech recognition (Graves et al., 2013) and polyphonic piano note transcription (Bock and Schedl, 2012). Motivated by this, (Parascandolo et al., 2016) proposed a bi-directional long short-term memory RNN based approach to polypho-

nic sound event detection in real life recordings.

According to the ability that RNN can learn the longer term temporal information in the audio signals, combined RNN with CNN could achieve better performance in complex sound event detection tasks. A combined CNN and RNN two complementary classification methods which is called CRNN are presented in (Cakir et al., 2017). The higher level features are first extracted by convolutional layers, then, after pooling in the frequency domain, these features are fed into recurrent layers to acquire sound event activity probabilities. Experiment result shows a clear improvement of the proposed architecture when compared with CNN and RNN in polyphonic sound event detection. (Adavanne et al., 2017) propose to use low-level spatial features extracted from multichannel audio for sound event detection with a CRNN. Furthermore, the author extended convolutional recurrent neural networks to handle multiple feature classes and process feature maps using bi-directional LSTMs. Experiment result showed that with binaural spatial features, the accuracy is higher than using monaural features.

The most popular feature to train DNNs, and their variants, are log-mel features and MFCC features. However, a filter bank that is designed from perceptual evidence is not always guaranteed to be the best filter bank in a statistical modeling framework (Sainath et al., 2015). Inspired by this, a growing number of researchers start to use raw waveforms as inputs to avoid information lost for training neural networks. (Sainath et al., 2015) proposed a Convolutional, Long Short-Term Memory Deep Neural Network (CLDNN) acoustic model which trained on over 2,000 hours of speech and achieve a competitive performance with log-mel based deep models. Research result shows that applying LSTM directly will cause excessive time steps, hence, the original recordings are first processed by a time convolutional layer (tConv). Then, the frame-level feature produced by tConv is passed to the CLDNNs. In the CLDNNs, convolution layers are used to reduce frequency domain changes, LSTM layers are used for long-time domain modeling and DNN is the classifier.

(Dai et al., 2017) found that most waveform-based models have generally used very few (less than two) convolutional layers, which might be insufficient for building high-level discriminative features. Hence, the author proposes a very deep convolutional neural networks with waveforms as inputs. In order to have a better representation of sounds, large receptive field in the first convolutional layer are applied to mimic bandpass filters, while very small receptive fields are used to control the model capacity. Meanwhile, fully connected layers and dropout function are abandoned as well. Experiment result showed that with the layers of model gained, the performance gained.

Computational modeling of auditory awareness plays a major role in a large variety of artificial intelligence applications. To be specific, a bottom-up attention mechanism specifically designed for efficient auditory surveillance demonstrated powerful detection of alarming sound events such as gunshots and screams in natural scenes (Hu et al., 2010). However, due to the insufficiency of the database, current deep model-based methods commonly use single sounds or artificially synthesized mixtures to train the networks rather than the real environmental sounds. Meanwhile, manually-selected has been proved that the loss of information cannot be avoided, but use raw waveforms as input need not only more layers but also large receptive fields in the first layer which demand a high quality of hardware. Furthermore, the conclusion from (Sainath et al., 2015; Dai et al., 2017) clearly showed that training deep architectures with waveforms could only match the performance of the model using log-mel features.

4 CONCLUSIONS

We conduct a survey on auditory saliency map based models and recent deep neural networks based approaches in the research field of modeling acoustic perception in this paper. The existing methods could be classified into two categories: convolution techniques based on ASM and other methods. The first kind of approaches was based on visual saliency map and they are easy to compute. However, depending on the experiment results we can conclude that they are not efficient enough in simple sound saliency detection, not to mention the environment sounds which has more complex composition and structures. Even the second kind of techniques performed better than the first one, experiment results showed that there is still vast room to improve the current techniques.

One major barrier to the application of acoustic perception models is the absence of a universally applicable and accurate human auditory perception model. Existing models are proposed for specific tasks and the performance will be greatly reduced when they are used to detect other kinds of sound. Another important problem is the lack of a universal environment sound database. Although there is some sound database, most of the units are just simple sounds. Furthermore, the current research works showed that the accuracy could be improved but the computational complexity will raise as well. This means that the

existing method could not detect the salient sounds in real time. Therefore, the application of acoustic perception models on the artificial intelligence for understanding the real environment is still severely constrained.

The novelty detection approaches show a possible way of applying DL in auditory saliency detection. Novelty detection aims at recognizing situations in which unusual events occur. Plethoric novelty based approaches have been proposed in medical diagnosis (Clifton et al., 2011), electronic IT security (Patcha and Park, 2007) and damage inspection (Surace and Worden, 2010). Only in recent years, novelty detection attracts the attention of researchers in the field of auditory perception. (Principi et al., 2015) presented a system to deal with acoustic novelty. When abnormal events occurred, this system could alert the users to help them take appropriate decisions. (Marchi et al., 2017) presented a broad and extensive evaluation of state-of-the-art methods focus on novelty detection and show evidence that RNN-based autoencoders significantly outperform other methods. In practical application, we assume that after training DNN or its variants by tremendous sound samples of different classes and different sound scenes, while training the model to judge what kind of acoustic events are possible to exist in the current scene. After all, the model could be used to detect whether there are some abnormal or salient sounds appeared in the current environment. This could be regarded as recognizing the novelty or saliency sound in the mixtures through deep architectures.

Deep neural networks clearly have the potential to improve the performance of environment sound perception and further help to build practical artificial awareness. However, no competitive research results have been achieved compared to its applications in visual awareness, great effort is still needed in the future.

FUNDING

This article was supported by the China Scholarship Council [Grant Nos. 201606290083] for 1.5-year study at the University Paris-Est. This research was also supported by the National Science Foundation of China under Grants 61502391 and by the Natural Science Basic Research Plan in Shaanxi Province of China under [Grant Nos. 2017JM6043].

REFERENCES

Adavanne, S., Pertila, P., and Virtanen, T. (2017). Sound event detection using spatial features and convolutional recurrent neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Bock, S. and Schedl, M. (2012). Polyphonic piano note transcription with recurrent neural networks. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Botteldooren, D. and De Coensel, B. (2009). Informational masking and attention focussing on environmental sound. In *Proceedings of the NAG/DAGA meeting*.

Cakir, E., Heittola, T., Huttunen, H., and Virtanen, T. (2015). Polyphonic sound event detection using multi label deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303.

Chachada, S. and Kuo, C.-C. J. (2013). Environmental sound recognition: A survey. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE.

Clifton, L., Clifton, D. A., Watkinson, P. J., and Tarassenko, L. (2011). Identification of patient deterioration in vital-sign data using one-class support vector machines. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 125–131. IEEE.

Dai, W., Dai, C., Qu, S., Li, J., and Das, S. (2017). Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

De Coensel, B. and Botteldooren, D. (2010). A model of saliency-based auditory attention to environmental sound. In *20th International Congress on Acoustics (ICA-2010)*, pages 1–8.

Duangudom, V. and Anderson, D. V. (2007). Using auditory saliency to interpret complex auditory scenes. *The Journal of the Acoustical Society of America*, 121(5):3119–3119.

Espi, M., Fujimoto, M., Kubo, Y., and Nakatani, T. (2014). Spectrogram patch based acoustic event detection and classification in speech overlapping conditions. In *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE.

Gencoglu, O., Virtanen, T., and Huttunen, H. (2014). Recognition of acoustic events using deep neural networks. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 506–510. IEEE.

Graves, A., rahman Mohamed, A., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Hu, R., Hang, B., Ma, Y., and Dong, S. (2010). A bottom-up audio attention model for surveillance. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Kalinli, O. and Narayanan, S. (2008). A top-down auditory attention model for learning task dependent influences on prominence detection in speech. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Kaya, E. M. and Elhilali, M. (2012). A temporal saliency map for modeling auditory attention. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*. IEEE.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947.

Knudsen, E. I. (2007). Fundamental components of attention. *Annual Review of Neuroscience*, 30(1):57–78.

Li, J., Dai, W., Metze, F., Qu, S., and Das, S. (2017). A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Li, X., Tao, D., Maybank, S. J., and Yuan, Y. (2008). Visual music and musical vision. *Neurocomputing*, 71(10-12):2023–2028.

Marchi, E., Vesperini, F., Squartini, S., and Schuller, B. (2017). Deep recurrent neural network-based autoencoders for acoustic novelty detection. *Computational Intelligence and Neuroscience*, 2017:1–14.

Mesaros, A., Heittola, T., Eronen, A., and Virtanen, T. (2010). Acoustic event detection in real life recordings. In *Signal Processing Conference, 2010 18th European*, pages 1267–1271. IEEE.

Mohamed, A.-R., Dahl, G. E., and Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22.

Mosadeghzad, M., Rea, F., Tata, M. S., Brayda, L., and Sandini, G. (2015). Saliency based sensor fusion of broadband sound localizer for humanoids. In *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.

Parascandolo, G., Huttunen, H., and Virtanen, T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12):3448–3470.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE.

Principi, E., Squartini, S., Bonfigli, R., Ferroni, G., and Piazza, F. (2015). An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications*, 42(13):5668–5683.

Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., and Vinyals, O. (2015). Learning the speech front-end with raw waveform cldnns. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283.

Schauerte, B. and Stiefelhagen, R. (2013). "wow!" bayesian surprise for salient acoustic event detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Surace, C. and Worden, K. (2010). Novelty detection in a changing environment: A negative selection approach. *Mechanical Systems and Signal Processing*, 24(4):1114–1128.

Takahashi, N., Gygli, M., and Gool, L. V. (2018). AENet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, 20(3):513–524.

Wang, J., Zhang, K., Madani, K., and Sabourin, C. (2015). Salient environmental sound detection framework for machine awareness. *Neurocomputing*, 152:444–454.

Zwicker, E. and Fastl, H. (2013). *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media.