A bio-inspired smart perception system based on human's cognitive auditory skills

Yu Su

Thesis

Presented to obtain the title of
DOCTEUR DE L'UNIVERSITÉ PARIS-EST
DOCTOR OF ENGINEERING

Specialization: Signal and Image Processing

By: **Yu SU**

A bio-inspired smart perception system based on human's cognitive auditory skills

Defended on 11 November 2019 in presence of commission composed by

Prof.	Vladimir	Golovko	Reviewer / Brest State Technical University
Prof.	Yikang	Yang	Reviewer / Xi'an Jiaotong University
Prof.	Ying	Liu	Examiner / Xi'an University of Posts and Telecommunications
Dr.	Jingyu	Wang	Examiner / Northwestern Polytechnical University
Prof.	Ke	ZHANG	Co-supervisor / Northwestern Polytechnical University
Prof.	Kurosh	MADANI	Co-supervisor / LISSI - University PARIS-EST Créteil

Thèse

Présentée pour l'obtention des titres de
DOCTEUR DE L'UNIVERSITÉ PARIS-EST
DOCTOR OF ENGINEERING

Spécialité: Traitement du Signal et des Images

Par: **Yu SU**

Un système Ingénieux de perception bio-inspiré basé sur les capacités auditives cognitives humaines

Soutenue publiquement le 11 novembre 2019 devant la commission d'examen composée de

Prof.	Vladimir	Golovko	Rapporteur / Brest State Technical University
Prof.	Yikang	Yang	Rapporteur / Xi'an Jiaotong University
Prof.	Ying	Liu	Examinateur / Xi'an University of Posts and Telecommunications
Dr.	Jingyu	Wang	Examinateur / Northwestern Polytechnical University
Prof.	Ke	ZHANG	Codirecteur de thèse / Northwestern Polytechnical University
Prof.	Kurosh	MADANI	Codirecteur de thèse / LISSI - Université PARIS-EST Créteil

Acknowledgement

First of all, I am very grateful to my mentor, from Université Paris-Est Créteil, Professor Kurosh Madani, for his professional instruction and invaluable support to my doctoral research. His guidance has greatly improved my understanding of academic writing and taught me a lot of specific research skills. It has been a great honor for me to conduct my work under his supervision and to be one of his PhD students.

I would like to give my sincerest appreciation to my advisor from Northwestern Polytechnical University, Professor Ke Zhang, for the continuous support and help of my PhD research. I am grateful to him wholeheartedly, not only for his tremendous academic advice, but also for providing me with so many precious opportunities and unconditional trust.

I would like to express my gratitude to Prof. Vladimir Golovko and Prof. Yikang Yang, referees (“rapporteurs”) of my thesis, for their kind acceptation of assessing my doctoral thesis. In the same way, I address special thanks to Prof. Ying Liu for her kind acceptance to be examiner and a member of my defence committee.

I would like to extend my special thanks to my tutor from Université Paris-Est Créteil, Dr. Christophe Sabourin, for his generous support and great patience. I am very grateful to him for his scientific experience and especially his guidance which helped me in all the time of research.

My sincere thanks also should be given to Dr. Jingyu Wang, for his scientific suggestions, professional advices, helps and encouragements on my PhD research work.

I am very grateful to Dr. Minghu Tan for proof reading several chapters.

Finally, I would like to thank my family and particularly my wife and daughter for their constant support and encouragement

Table of Contents

List of Figures

List of Tables

List of Symbols

f	Frequency
$ERB(f)$	ERB scale
f_c	Center frequency
ϕ	Phase of the carrier
α	Amplitude
n	Order of the filter
b	Bandwidth of the filter
t	Time
$g_m(i)$	Matrix which represents a variant of cochleagram
$GFCC$	Gammatone filterbank cepstral coefficient
$GFCC_{curve}$	GFCC curve
$s_{i,j}$	The j_{th} frame of signal S
$E(S_j)$	Short-term Shannon entropy of j_{th} frame
$E(S)$	The short-term Shannon entropy of the entire signal
r	Tolerance
m	Number of points
N	Time series of a signal
$A^m(r)$	Probability that two sequences will match for $m+1$ points
$B^m(r)$	Probability that two sequences will match for m points
$SampEn(m,r,N)$	Sample Entropy

E_i	Value of each point in entropy domain
$\max(E(S))$	Max value of entropy
D_i	The rest deviancy point in entropy domain
$P_{peak}(i)$	Maximum point of PSD curve
P_{avg}	Mean of PSD curve
P_i	Maximum point greater than P_{avg}
P_{i-1}	Point before P_i
P_{i+1}	Point after P_i
Px_i	The location in frequency axis of P_i
Px_{i-1}	The location in frequency axis of P_{i-1}
Px_{i+1}	The location in frequency axis of P_{i+1}
$S_{s,i}$	Salient points
I_c	log scale gammatone spectrogram of a sound signal
$I_c(R)$	Red color values of pixels in RGB color space
$I_c(G)$	Green color values of pixels in RGB color space
$I_c(B)$	Blue color values of pixels in RGB color space
S_I	Image indicator
TP	True positives
FP	False positives
FN	False negatives
P	Precision
R	Recall

F_{score}	F-score
$C \& C^{\#}$	Chroma
t_n	Tonal centroid vector
L_1	L_1 norm
c_n	Chroma vector
$mel(f)$	Mel-scale frequency
X_m	Log energy in m^{th} log mel spectrogram
c	Index of the cepstral coefficient
X_k	X_k is the input patch centered at location (i, j)
W_k	weight vector of k_{th} filter
b_k	bias vector of k_{th} filter
$s(i, j)$	The corresponding position element of the output matrix
$f(x)$	ReLU function
Θ	Discernment
A_i	Element of the power set $P(\Theta)$
$P(\Theta)$	Power set
$M(x)$	Mass function
$M(A)$	Probability in subset A
$M(\varnothing)$	Empty set of mass function
$\oplus$	Dempster's rule
$\cap$	Intersection operation
$\varnothing$	Null set

$\subseteq$	Subset
$\forall$	Arbitrary sign
α_D	Normalization constant
$M_{1\oplus 2}(A)$	final probability assignment of $M_1(B)$ and $M_2(C)$
$M_1(B)$	CNN outputs of softmax of LMCNet
$M_2(C)$	CNN outputs of softmax of MCNet
$Scene_j$	j_{th} scene
N_n^j	n_{th} normal events in j_{th} scene
AN_m^j	m_{th} abnormal events in $Scene_j$
x_i^j	Abnormal sound event
$P(\bullet)$	The level of significance of a sound event
α_P	Attention threshold
$P\left(x_i^j\right)$	Significance of sound event x_i^j in $Scene_j$
$P\left(y_i^j\right)$	Significance of previously salient sound in $Scene_j$

Chapter 1. General Introduction

1.1. Foreword

Developing a machine that possesses human-like consciousness has been the ultimate goal of artificial intelligence (AI) since the computer was invented. During the past two decades, tremendous efforts have been put out to explore the models for artificial consciousness (Churchland, 1984; Crick & Clark, 1994; Block, 1995; Chalmers, 1996; Aleksander, 2001; Edelman & Tononi, 2003; Baars & Franklin, 2009; Fekete & Edelman, 2011). In recent years, it has already been applied to numerous areas such as autonomous cars, virtual personal assistants, smart hospitals, logistics robot and so on. Where, computer vision, natural language processing, sound recognition, expert system and some other technologies are the cornerstones supported these applications. These technologies are collectively referred to as AI technology.

Generally, AI can be divided into two categories (Holland, 2009; Seth, 2009): weak artificial intelligence and strong artificial intelligence. The weak AI does not have the ability of reasoning and problem solving, can only process one specific kind of problem. The strong AI is the main goal of artificial intelligence research in recent years, it should have the ability to perform general smart behavior and can percept and aware like human beings in every aspect. At present, we are in the era of weak artificial intelligence turning to strong artificial intelligence. It is possible that adding conscious awareness, or information processing capabilities associated with the conscious mind, would open the door to a much more powerful and general AI technology (Reggia, 2013).

There are two main demands of studying consciousness of machines, the first is to improve our comprehension of the nature of consciousness (Edelman & Tononi, 2003; Reggia, 2013), the second motivation for work in artificial consciousness is the expectation of creating an intelligent machine (McDermott, 2007). For the first demands, research works on artificial consciousness generally believe that the objective

methods of science will never be able to reveal the core of consciousness due to its subjective nature (McGinn, 2004). While researchers observe that computational modeling specific parts of the human brain (consciousness) might be useful for us to understand how does the brain works. For the second, designing and manufacturing machines with consciousness are the technological goal. While this expectation is obstructed of current AI techniques. Although benefit from the tremendous advancement of computer technology, neurobiology and neuropsychological, the application of intelligent machine dramatically increased in numerous domains in the past decades. However, the level of intelligence of these applications is far from reaching the human's abilities.

Regardless of the various applied scene, cognition of the surrounding environment is the key component for all machines with artificial consciousness. Cognition is one of the mechanisms of the human brain to process acquired information and make them understandable and repeatable. Providing such a human-like mechanism to machines or robots will effectively enhance their perceptual performance in a real environment as well as the level of intelligence.

In general, cognition is the final goal of the brain for information processing, it contains three steps to realize the ultimate objective: 1) information acquisition or detection, 2) information recognition, and 3) respond to information. Human auditory and visual perception systems are the major channels of cognition to percept the environment. It is well known that these perception systems are a highly intelligent, efficient system that could perceive massive information or stimuli while sensing the surrounding environment at the same time. Yet, current research works are mainly focused on modeling human-like visual cognition and perception, the study of modeling auditory system is still in its infancy. This is because that establish a computational auditory cognition system is such a great challenge to artificial intelligence as well as the difficulties of processing complex environmental sounds in a biologically inspired way. Moreover, most of the existing research works which aim at establishing auditory models are just partial function modeling of the hearing system, like auditory attention models and sound recognition models. Consequently, the development of novel and the

comprehensive bio-inspired auditory system should be made to provide better cognition ability for artificial machines.

The human brain is a sophisticated system consists of tremendous neurons. All the information obtained by visual and acoustic channels will be uninterruptedly processed by the brain. However, human beings are surrounded and exposed to a large amount of information at all times, even when we fall asleep, neural resources are limited in our brain and not all stimuli can be processed and need to be processed to the same extent. Mechanisms exist to prompt attention toward the specific conspicuous events, thus providing a weighted representation of our environment (Desimone & Duncan, 1995). This mechanism is the selective attention mechanism and is considered as a key component of cognition as it allows the perception in the auditory channel to work efficiently for information acquirement. In this thesis, this mechanism is used as the fundamental basis for the sounds detection module of the whole system, where novel techniques are researched to obtain better performance in sound events acquisition.

Regarding the second process step of establishing an artificial auditory cognition system, deep learning-based algorithms are exploited. Deep learning-based techniques have been proved to be more efficient than conventional methods in solving complex classification problems in many domains. Multiple scientists choose deep learning models, such as CNN, in sound classification problems. CNN can solve the limitations of conventional classifiers in multiple learning and classification problems. However, there is still a long way to go when compared with CNN based image classification algorithms. For example, the longer temporal context information still cannot be captured by the original CNN. However, from the classification accuracy derived from the recently published works, it is clear that the CNN-based ESC systems still have great potentials for making further progress. Hence, novel CNN-based ESC techniques will be further researched in this thesis.

Concerning computational modeling the third step of cognition, recent research works have shown that long-term life experiences affect the ability to hear in background noise (Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013). To be specific, compared with the unconsciously detected salient sound events, the sounds

which have been heard can attract our attention more easily. This result closely parallels theories from the ‘top-down’ attention mechanism, which points out that subjective consciousness also has a great influence on attention. For example, listeners can easily attend to one speaker in a multi-speaker environment (O’sullivan et al., 2014). This phenomenon is also known as the cocktail party problem which pointed out that prior knowledge should be regarded as a crucial component of realizing artificial auditory cognition. Consequently, the impact of experience or knowledge should be taken into consideration in modeling respond function for an artificial auditory cognition system.

1.2. Biological Background

Ears are the major sensory of human cognition system, they cope with a myriad stimulus of the surrounding environment into signals of nerve impulses which generated by different kinds of nerve cells at all time, even when we are falling asleep. Compared with visual signals, sound signals will enable mankind to be aware of and avoid danger beforehand or when human vision is not available in a certain environment. From a physical point of view, sounds are the pressure wave that propagates through a medium (such as the air) and can be perceived by the human or animal auditory system. A sound has three main physical attributes: pitch, loudness and timbre. These physical characteristics are measurable properties of the sound signal while cognition is the reflection of the listener's mind on the sound.

Human is only consciously percept part of the ongoing stream of auditory information being received at each moment. The attention mechanisms select what we attend to and have the ability to focus on important aspects of sensory information. For example, listeners can easily attend to one speaker in a multi-speaker environment. Tremendous research in cognition and consciousness have proved that human attention is controlled by bottom-up attention and top-down attention (Buschman & Miller, 2007; Bayne, Cleeremans, & Wilken, 2014). These attention mechanisms process acquired information, weaken irrelevant neural activity and inhibit activity representing external objects (Kanwisher & Wojciulik, 2000; Reynolds & Chelazzi, 2004).

The bottom-up attention mechanism is also known as stimuli-driven attention, or saliency driven attention. It is the attention mechanism which transfers low-level information into higher-level information through many processing levels in the human brain. In this manner, human attention is elicited by conspicuous stimuli generated by the salient events in the environment, then, higher-level information such as response decision and next step activities could be achieved (C.-C. Liu, Doong, Hsu, Huang, & Jeng, 2009). On the contrary, top-down attention underlies our ability to concentrate on relevant stimuli and neglect irrelevant conspicuous events. The widely accepted opinion is that top-down, or goal-directed attention is undeniably important in volitionally selecting stimuli that match current task demands (Awh, Vogel, & Oh, 2006). Top-down modulation of sensory processing is not an intrinsic property of sensory cortices but rather relies on long-range inputs from and interactions with a network of 'control' regions in our brain (Gazzaley & Nobre, 2012). To be specific, life experience and memories can influence auditory cognition processing directly.

Attention and cognition are not the same processes, yet, they are generally closely connected and interrelated. Thereupon, multiple research works concentrate on modeling artificial auditory cognition have engaged in modeling acoustic attention as the first step of establishing auditory cognition models is understandable. Early studies have established some models to illustrate selective attention mechanism exists in both visual and auditory cognition system. For example, the attention model (Cherry, 1953) and response selection model (Treisman, 1960). The major assumption of these models is the ability of information-processing mechanisms in the human brain is limited. Therefore, such models could avoid the "bottleneck" of cognition in cognitive psychology research by selecting only conspicuous auditory or visual stimuli to be processed by a higher-level processing mechanism.

(Gray, Buhusi, & Schmajuk, 1997) claimed that the different processing provided to new stimuli is the key element in a stimulus entering cognition. It is pointed out in this work that a novel stimulus activates specific neural circuitry forming a separate novelty system that increases the attention system's activity and facilitates learning. This transition from low-level attention to higher-level attention is considered as the

variation from unconscious processing to conscious processing. The model presented here has been mapped onto neuroanatomical structures, and it has been related to latent inhibition occurring during classical conditioning and to the cognitive abnormalities that are characteristic of schizophrenia (Gray et al., 1997).

A model simulated the conscious processing occurs from the symbol grounding aspect of attention mechanisms is proposed by (Kuipers, 2005). It is described in this study that the cognition mechanism is surrounded by massive, continuous amounts of stimulus and events, it is a major problem of cognition processing to select the valuable information which deserves attention. In this work, the model applies symbolic memory storage and reasoning methods, this selective attention can be applied through a tracker. The tracker is a symbolic indicator of the data that preserves a relationship between low-level representation and symbolically, high-level representation in the data over time. In fact, this model partially simulates the attention mechanism through performing symbol grounding and symbolic representations to choose temporal-spatial segments of acquired data are efficient to mimic consciousness. The claim is that any system organized in this fashion, having both bottom-up and top-down attention mechanisms that create trackers along with a reasoning system of control laws that makes use of these grounded symbols, is a truly practical conscious model.

Apart from these theoretical models in cognition modeling research, multiple researchers from various relevant domains believed that selective attention mechanism enables human beings to focus on the most salient events occurred in the surrounding environment unconsciously but fast. It could also be the most efficient mechanism in searching the expectation objects when we explorer the natural world (J. Wang, 2015). (C. Kayser, Petkov, Lippert, & Logothetis, 2005) proposed an auditory saliency detection model based on the auditory saliency-driven attention mechanism. This model converts sound waveforms to a time-frequency representation, which is called an “intensity map” in this work. Then, three acoustic features: intensity, frequency contrast and temporal contrast are extracted on different scales with different sets of filters. The center-surround mechanism and normalization are applied to promote those feature maps containing prominent values. These maps are combined across different

scales to yield the saliency maps for each feature sets. Finally, linear combined the three saliency maps of each feature to generate the final auditory saliency map. Experiment results showed that this model could mimic several basic properties of the human auditory perception mechanism.

Afterward, based on Kayser's work, two more similar auditory attention models were proposed by (Kalinli & Narayanan, 2007) and (Duangudom & Anderson, 2007). In these works, saliency is considered as the key component for the auditory attention mechanism in acquiring the surrounding information. However, it is a common experience that during we focus on one salient event, our attention can be involuntarily engaged by visual or acoustic changes occurring unexpectedly in the environment (Escera, Alho, Winkler, & Näätänen, 1998; Schröger, 1996). This attention shift phenomenon of our cognitive-perceptual mechanism could also be introduced as deviancy detection. It should be noticed that deviancy can only be defined in relation to something regular (Winkler & Schröger, 2015). A novel event is determined with deviancy should satisfy that such event breaks the existing status of the current environment which it appears.

In the auditory system, deviations range from simple cases to complex ones, such as breaking a successive sound, and someone interrupting others' conversations. The deviancy should also take the environment into consideration because the environment is not the physical effects obtaining by the sensory of the observer. One's experience of environments is also a major determining element of what we acquired as deviancy. This could be regarded as the top-down attention mechanism as well. Computational modeling such a mechanism for artificial auditory cognition is more important than in vision modality. It is because that the acoustic environment is ephemeral and it lacks the ability which can be repeated at any time.

Deviance detection is an important mechanism as it represents new information that may require a response from the observer. Moreover, recent research works have shown that long-term life experiences affect the ability to hear in background noise (Anderson et al., 2013). To be specific, compared with the unconsciously detected salient sound events, the sounds which have been heard can attract our attention more

easily. This result closely parallels theories from the ‘top-down’ attention mechanism, which points out that subjective consciousness also has a great influence on attention. For example, listeners can easily attend to one speaker in a multi-speaker environment (O’sullivan et al., 2014), this phenomenon is also known as the cocktail party problem. This result pointed out that prior knowledge should be regarded as a crucial component of realizing artificial auditory cognition. Consequently, it is essential to establish novel artificial auditory cognition models that could simulate the human auditory mechanism where the deviant sound events can be identified and can respond to these events while neglecting the rest.

1.3. Motivation and Objectives

The motivation of this thesis is to achieve the ultimate goal of embedding artificial auditory cognition ability for intelligent machines, in order to precisely select the high valuable conspicuous sound events occurred in the environment and make an efficient response to them, thereby reducing the computational cost of the machines. As discussed above, the cognition consists of three major components: 1) information acquisition or detection, 2) information recognition, and 3) respond to information. Each element should be modeled to realize modeling the cognition mechanism. Most researchers exploit the saliency-driven attention principle as the basis of modeling sound information acquisition processing. In (Kaya & Elhilali, 2012), an auditory saliency map which treats the input signals as a one-dimensional temporal input was presented. In (Kim, Lin, Walther, Hasegawa-Johnson, & Huang, 2014), a saliency detection model based on the classification result was presented. (Tsuchida & Cottrell, 2012) and (Schauerte & Stiefelhagen, 2013) introduced their novel auditory saliency map based on the theory of statistics to predict the saliency in soundscapes.

In the meantime, the mechanism of deviancy detection is rarely considered in modeling auditory attention and current studies are focused on revealing how deviancy detection works and processes in the human brain. (Vachon, Labonté, & Marsh, 2017) conducted a systematic investigation whereby the impact of verbal deviants and spatial

deviants on verbal and spatial short-term memory was assessed. This study established that both verbal and spatial deviants can hinder both verbal and spatial order-reconstruction. This work suggested that this would seem intuitive because that, the warning capacity of the auditory cognition system should ensure the brain attends to the deviant events while ignoring the currently attended goal, the informational value of the task-irrelevant sound and any coupling between relevant and irrelevant information. The author finally concluded that the deviancy reflects a general form of auditory distraction as interference took place both within and across domains and regardless of the processes engaged in the focal task. Therefore, the deviancy detection mechanism could be regarded as a supplement to saliency detection, computational modeling of the bottom-up attention mechanism which constitutes both detection manner can help machines to perceive the environment in a more efficient way.

The information recognition processing can be also regarded as the processing of low-level information acquired from sensory convert into higher-level information in the human brain. With the popularity of using deep learning-based models in various categorization problems and they have been proved to be more robust than conventional methods, a growing number of researchers exploit such methods in sound recognition tasks in recent years. However, the most widely used acoustic features, such as MFCC, used for training deep learning-based models may lose some important information about environmental audio events. Meanwhile, most of the deep architectures cannot achieve satisfactory performance in categorizing the environmental sounds.

In the past decades, many studies have presented a large number of models to simulate the human auditory cognition. It can be seen in these works that most of the proposed models could only achieve a partial function of the human auditory system. The systematic artificial auditory cognition model is still relatively rare. (J. Wang, 2015) proposed a bio-inspired perception system based on visual and auditory attention mechanism, in which the functions refer to find the abnormal events in complex environment through both audio and visual information. (Fuertes & Russ, 2002) design a perceptive awareness model for reaching perceptive awareness for automatic systems. The model can recognize the environment and select the appropriate response to the

current situation. Perception Data from both microphones and cameras are prior considered in this model.

Motivated by the above-discussed research works and current obstructions in simulate human auditory cognition mechanism, the major purpose of my work is to provide machines with artificial auditory cognition that can perceive the surrounding environment in a human-like manner. Thereupon, such intelligent machines can continuously recognize the environment through the auditory channel if the visual channel is hindered. Consequently, the salient and deviant sound should be acquired efficiently and accurately. Then, the detected sound information will be further processed to obtain the higher-level information in an efficient way for realizing the artificial auditory cognition. The objectives of this thesis can be introduced in three major aspects:

- Study the state-of-art auditory cognition, perception and attention models correspond to the environmental sounds analysis tasks. Develop novel biologically inspired auditory deviant detection model for complex environmental sound deviancy detection.
- Explore the efficient acoustic feature sets and feature combination strategies, investigate the state-of-art environmental sounds classification (ESC) methods. Propose novel auditory features and deep learning-based sound classification model for ESC problems.
- Establish a novel knowledge-based system for auditory event response decisions by taking both prior knowledge of environmental context and detected sound events into account. Integrating the proposed models to form an artificial auditory cognition system that can provide a human-like auditory mechanism in a complex environment.

1.4. Contribution

The overview framework of the proposed biologically inspired artificial auditory cognition system is shown in Figure 1. Three major information processing modules

are presented to illustrate how low-level sound information transforms into high-level acoustic information.

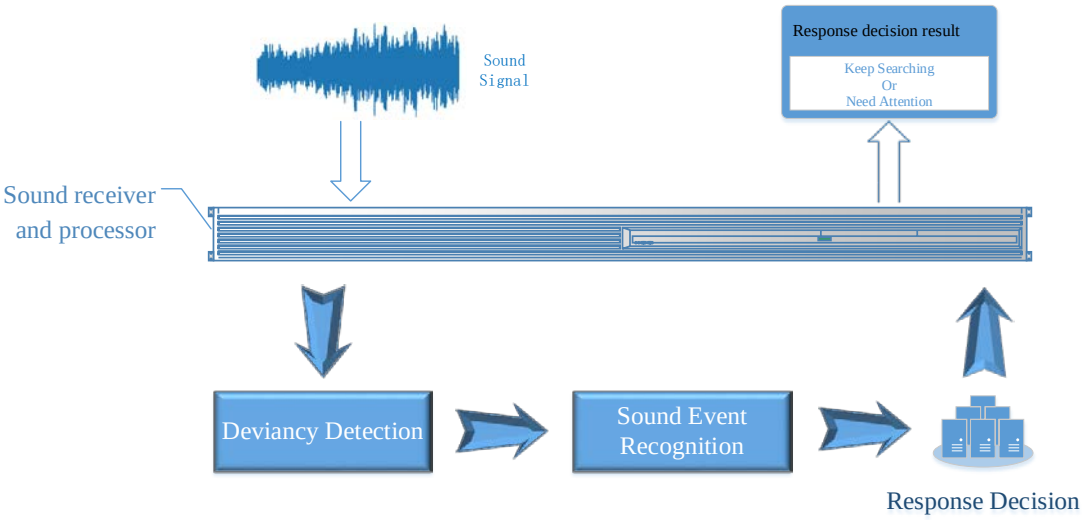


Figure 1. The overview framework of the proposed system

In Figure 1 it can be seen that when sound events occur in the environment, the sound receiver will be triggered at first to perceive and preserve the sounds. Then, the sound will be processed by a deviant detection module to identify whether the novel sound event is salient or not. It should be noticed that the deviant is relative conception, sounds are determined with deviancy or not depending on the focal auditory tasks. This means if the current environment is silent, the novel sound events could be seen as salient or deviant sounds. Thereafter, the detected deviant sound will be identified through the environmental sound recognition module. Finally, the auditory event response decision module is deployed to determine whether the deviant sound needs attention or not with prior knowledge derived from the current environment.

Several contributions relate to establishing the artificial auditory cognition for intelligent machines have been accomplished in this thesis:

- The first contribution is the studying of state-of-art research works with respect to the auditory attention models, deep-learning-based environmental sound classification techniques and human auditory response mechanisms, which shed light on the current research status and complexities of achieving the

ultimate goal in this thesis. Such studies demonstrate the obstacles and disadvantages of existing research results, resulting in the motivation of modeling auditory deviant detection mechanisms to acquire novel sound events, applying convolutional neural networks to deal with categorizing the detected sounds and exploit the knowledge-based system to simulate human auditory response mechanism.

- The second contribution is the proposition and realization of an auditory deviancy detection model, where features derived from temporal and spatial domains are extracted for sound deviancy detection. It should be emphasized that a sample entropy-based deviancy detection method is proposed to accurately extract the real deviant sounds in the temporal domain. In this method, the Shannon entropy is exploited to identify the most deviant sound peak point, and sample entropy is applied as a measurement to point out all the peak points belong to the deviant sound. Thus, the deviancy and saliency features derived from each domain are combined to yield the final result, which can be deployed in the real environment sound detection tasks.
- The third contribution is the analysis of the performance of various kinds of acoustic features in deep learning-based environmental sound classification models. Six widely used features are evaluated with a 6-layer CNN on a real environmental sound dataset. These features including cepstral features and image representation features are all derived from mel and gammatone filters. Then, eight feature combination strategies are presented based on basic features. These aggregated features are evaluated with CNN on the same dataset as well. Among these feature sets, three combined features present to be suitable in real environmental sound categorization tasks and can achieve competitive classification accuracy when compared with existing methods.
- The fourth contribution is the proposition of a two CNN fused environmental sound classification model, where DS evidence theory is applied as the fusion method. The CNN model is a novel designed 4-layer convolutional neural network while the two CNNs have the same parameters in each layer. Two

aggregated acoustic features evaluated in chapter 4 are applied to train these CNNs, separately. Then, the DS evidence theory is performed to fuse the softmax value derived from two CNN models. This deep learning-based sound classification architecture achieves an outstanding performance in real sound event taxonomic tasks, which demonstrated that this model is suitable for the auditory cognition requirement of intelligent machines in precepting the real environment.

- The fifth contribution is the conception and realization of a knowledge-based system and human auditory response decision manner inspired artificial auditory event response decision model. Motivated by the top-down attention mechanism in the human attention system, the prior knowledge of sound scene and environment is considered as the database to judge whether the detected sounds need attention or not. Each normal and abnormal sound event that might occur in a sound scene is distributed a significance value. The detected deviant or salient sound in an environment will be first compared with the possible sound events to find out its corresponding significance value. On account of the basis that the same sound may have different significance values in different environments, hence, the proposed model will judge if the new sound event deserves attention. This model can be applied to various auditory perception and cognition tasks. It can simulate the human auditory cognition mechanism to some extent and makes the artificial cognition an achievable function for intelligent machines.

1.5. Organization of Thesis

This thesis is mainly composed of five chapters. From the first chapter to the fifth chapter, the readers will be presented the current studies and results that relates to the thesis, each technique that I proposed for different module of artificial auditory cognition system, and the realization of whole system which provide artificial auditory cognition to machines for solving multiple cognition problems in complex environment.

The specific details of each chapter can be described as follows:

To help readers to fully understand the relevant biological inspirations of my work, interrelated background and research works are presented in Chapter 1 from the perspectives of human perception ability and characteristics. For the reason that the artificial cognition for an intelligent machine can be seen as the simulation of human consciousness, biological inspirations obtained from the human auditory cognition system are illustrated to give a comprehensive description of how I process the auditory information and establish the artificial auditory cognition system.

Chapter 2 illustrates the overview of my research field along with the state-of-art techniques that inspire this thesis. It has illustrated the relevant research works and models with respect to this thesis in three aspects: 1) the review of auditory saliency and deviancy detection techniques which established for auditory cognition, 2) the review of the application of deep neural networks in sound signal recognition, where the neural network-based environmental sound classification techniques are the main research orientation, 3) the overview of research works focus on auditory cognition in either theoretical level or computational modeling level published in the past decades. Several distinct approaches and observations are presented, in order to provide the general consideration of the motivation of this work. Then, the discussion regarding the state-of-art publications is connected to the problems that are researched in this thesis.

Chapter 3 focuses on auditory deviancy detection where a novel approach is proposed. It is mainly consisting of three modules. The first module is a novel approach for detecting the temporal deviancy based on the GFCC time domain curve to detect the local saliency of the sound signal. To detect the deviancy sound among those salient sound, a wavelet entropy and sample entropy-based temporal deviancy detection method are proposed. Thus, to accurately detect sounds saliency and deviancy, a module focus on frequency domain significance detection method based on the sound PSD to extract the saliency of sound in frequency domains is presented. Finally, an image indicator based on opponent color space is presented to give a better presentation of the deviant salient-sounds of sound signals. Two experiments were performed to verify the accuracy of the proposed model.

In chapter 4, the performances of several aggregated features for ESC tasks are evaluated. Since conventional sound event analysis mainly addresses time-frequency features or cepstral domain features only, and grounded on the fact that sometimes aggregate features from different domain may reduce classification accuracy. Considering that the classification performance of CNN as the classifier is sensitive to the hyperparameters and minor changes in parameters can lead to a large difference in classification results. Hence, features that comprehensively represent environment sounds and an appropriate CNN model should be carefully designed for ESC. Six basic acoustic features (Log-Mel Spectrogram, Mel Spectrogram, MFCC, Gammatone Spectrogram and GFCC) are used as features to evaluate the 6-layer CNN. Then, eight feature aggregate schemes that combined Chroma, Spectral Contrast and Tonnetz (CST) with the six basic features are presented. The performances of these feature combinations are tested on two datasets and the classification accuracy of each class include in these datasets is presented.

Chapter 5 illustrates the realization of the artificial auditory cognition system. Firstly, to further improve the performance of the CNN-based ESC model, the TSCNN model is proposed to precisely identify the class of environmental sounds. It consists of two 4-layer convolutional neural networks which are trained by two combined acoustic features. Then, the outputs of the softmax layer of both networks are fused through DS evidence theory, the fusion results are the predicted categorize of an environmental sound. Thereafter, a knowledge-based system inspired auditory events response decision model is originally proposed to better describe the significant characteristic of acoustic information obtained from the environment. The proposed method is performed by comparing the prior knowledge-based significance of detected salient or deviant sounds with sound scenes information to determine whether the system needs to respond to the abnormal sound events. Thus, abnormal sounds will be further categorized into meaningful and meaningless events, which means that meaningful deviant sounds need to respond and meaningless events do not need to respond. Meanwhile, the meaningful events need to be judged whether their significance is higher than focal tasks. At last, the proposed artificial auditory cognition

system is performed on several simulated scenarios for validation, and the results show that multiple perception tasks could be accomplished by the presented system.

The last chapter is the conclusion, where a summarized conclusion of all the research work conducted in this thesis is presented to the readers. Meanwhile, the perspectives of limitations, potential future work and ultimate goal with respect to the thesis are provided.

Chapter 2. Environment Information Perception

2.1. Introduction

Auditory cognition is an essential component of the human consciousness which helps human to perceive the surrounding environment accurately. However, the processing capacity of the human brain is limited and not all the acquired environment stimulus can be processed simultaneously. After years of evolution, a surprisingly ability was generated in our brain, called selective attention mechanism. This attention mechanism makes us can focus on the conspicuous events around us while ignore the irrelevant events. Furthermore, it is a common experience that during we focus on one salient event, our attention can be involuntarily engaged by visual or acoustic changes occurring unexpectedly in the environment. This attention shift phenomenon of our cognitive perceptual mechanism could also be introduced as deviancy detection. Neurobiologist believes that these saliency-based selective attention mechanisms could be the fastest way for humans to make responses to prominent stimulus which received from surrounding environment. Therefore, the bio-inspired saliency and deviancy detection approaches could be regarded as a feasible way for computational modeling the human selective attention mechanisms for artificial intelligence.

The saliency principle is generally used as the basis of artificial cognition models and bio-inspired human perception computational models. For computational modeling the human saliency principle, the current research works are mainly focus on the auditory saliency detection (ASD) rather than deviancy detection. It is because the ASD is the step before deviancy detection of auditory consciousness, and the ASD models could be established based on well-studied visual saliency detection models. Moreover, the researches of human auditory awareness mechanism are still at early stage. For many auditory mechanisms there are no precisely scientific and theoretical explanations of the processing details in human brain, such as “cocktail party effect” and the attention

shift phenomenon. These reasons make it harder to build exact computational models to mimic human auditory conscious. Meanwhile, though the techniques established for ASD are well researched when compared with other auditory mechanisms, and have been proved that can simulate human perception to some extent. However, the architectures of these models are all similar to the VSM model, on account of the characterizes of VSM and the feature used in this model, these aspects may lead to the loss of sound saliency information. Furthermore, there are no practically applied sound deviancy detection model that can simulate human attention shift mechanism till today.

In real life experience, various salient acoustic events generated by different sound sources occur frequently when we focus on one prominent auditory event (such as human speech or music), which attract our attention from focal task to new salient events. However, these new prominent events are not always meaningful sounds. For example, when we are talking to a colleague in an office, the car horns form outdoor are the environment noise for speech, which should not pay attention. While the phone ringing and the door knocking is the newly appeared events that should be noticed and make responses. Therefore, in order to make machines can precisely percept the surrounding environment like human, the sound event classification model and context judgment model must be established as well, in addition to saliency principle computational modeling.

Inspired by the perception mechanism of human beings, a practical solution is to apply the saliency principle for auditory feature extraction in different domains in order to obtain the saliency information in an audiovisual way. The initial characterization of saliency is to describe an event that is prominent relative to surrounding environments. This problem is well studied in human visual system and computer vison application, but less in auditory system. Till today, only few research works announced they have successfully embedding machine with human-like auditory cognition ability for autonomous environment perception. After considerable number of psychological acoustics experiments conducted by neurobiologists, they believe that mimicking the saliency principle and attention shift mechanism could be the potential way of modeling artificial auditory cognition. Hence, the related research works will be introduced first

as the fundamental basis in the following sections.

After the salient and deviant sound events are detected, they should be recognized, in other words: the classification system should be applied to identify the class of the prominent sounds. Recently, a growing number of researchers have begun to apply deep neural networks for environment sound events classification and recognition (ESC). In the past decades, Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are widely used classifiers in sound classification problems. However, deep learning-based models have been proved to be more competitive than the traditional methods on solving complex learning problems in various domains. At present, the deep architectures have conquered the field of image, music and speech recognition, but the application in ESC tasks still falls behind.

Moreover, as illustrated above, not all the conspicuous sound events which cause attention shift should be noticed and responded accordingly. In view of this fact, a decision-making system is also needed to decide whether the detected prominent sound events should be responded or not. The previous works of ESC will be illustrated, along with the research works related to decision making system based artificial cognition

Since my work is inspired by the previous researches done in the fields of sound processing, deep learning-based ESC and decision making, the state-of-art works of each research filed will be introduced respectively. Although not all of the previous works are proved to be efficient to auditory cognition, they are still included in this thesis is for presenting a general review of related researches and to let readers have a better point of view of why and how I conduct my research work, which could be regarded as the motivation and methodology of my work as well. To be specific, the previous works focused on the acquisition of auditory salient information will be illustrated in section 2.2, including classical and improved auditory saliency detection models. In section 2.3, the theoretical research works of auditory attention shift and biological basis of deviancy detection will be presented. The deep learning-based environmental sound classification system will be discussed in section 2.4. Finally, the decision-making mechanism based artificially acoustic cognition will be discussed in section 2.5.

2.2. Auditory Saliency Detection

Auditory saliency detection is one of the most important research fields of realizing machine awareness which aims at detecting the abnormal or conspicuous sound events in the real 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). It should be noticed that, the ASM is basically established followed the pioneering research work of saliency-driven attention (Koch & Ullman, 1987) and the visual saliency map proposed by (Itti, Koch, & Niebur, 1998). This model is a visual attention system inspired by the behavior and the neuronal architecture of the early primate visual system. In this model, feature maps of color, intensity and orientations are extracted from image inputs at first. Then, the center-surrounding process and normalization are performed on each set of feature map. A biologically inspired normalization operator is proposed to promote maps where a small number of strong peaks of activity is present, while suppressing maps contain comparable peaks. With the normalization operator and across-scale combination, each set of feature maps are combined in to three saliency maps. These maps are then summed into one visual saliency map (VSM) followed by “winner-take-all” and inhibition of return processing to prevent the model to subsequently jump to salient locations spatially close to the currently attended location. The model is able to reproduce human performance to some extent and shows a better performance than conventional visual saliency detection models.

2.2.1. Classical Models

A considerable research works conducts various experiments on the relevance of audio-visual perception information, point out that there are correlations between image processing and sound processing in human perception system. This is why the sound saliency can be transformed into visual representation, to be specific, visual and auditory perception channels have perceptual correlations in high-level perceptual processing of human brain. 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 the success of VSM and the theoretical basis, almost all research works in auditory saliency field translate sound signals into two-dimension images (spectrograms), and use similar method to detect sound saliency.

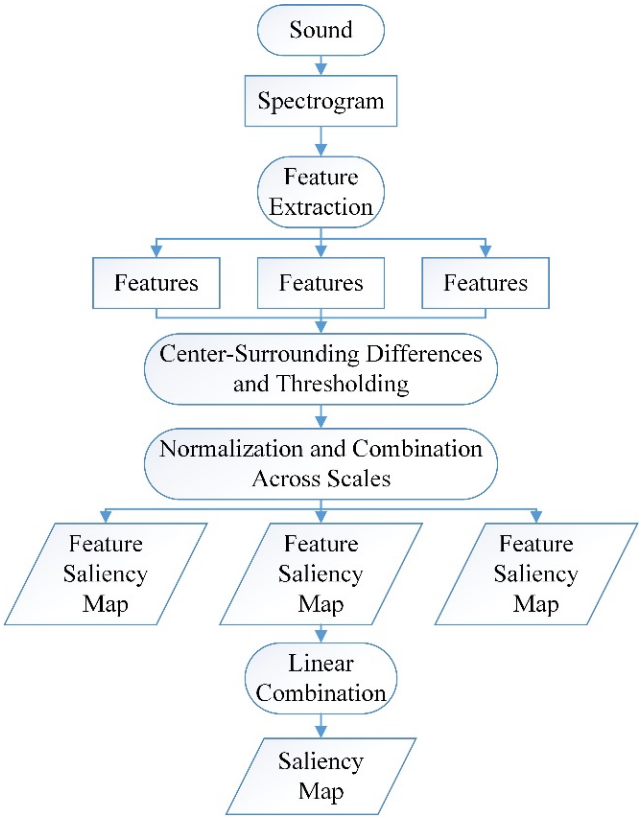


Figure 2. The auditory saliency map proposed by Kayser.

Several ASD models have been proposed for salient sound detection based on Itti's visual saliency map. These models can be regarded as the classical ASD models since they use original concept of VSM in their works. (C. Kayser et al., 2005) first proposed an auditory saliency map (ASM). Afterward, based on Kayser's work, two improved ASM approaches were proposed by (Kalinli & Narayanan, 2007) and (Duangudom & Anderson, 2007).

The auditory saliency model proposed by Kayser converts sound waveforms to a time-frequency representation, which is called "intensity map" in this work. Then, three acoustic features: intensity, frequency contrast and temporal contrast are extracted on different scales with different sets of filters. The center-surround mechanism and normalization are applied to promote those feature maps containing prominent values. These maps are combined across different scales to yield the saliency maps for each feature sets. Finally, linear combined the three saliency maps of each feature to generate the final auditory saliency map. The structure of Kayser's ASM which is identical to the visual saliency map is shown in Figure 2.

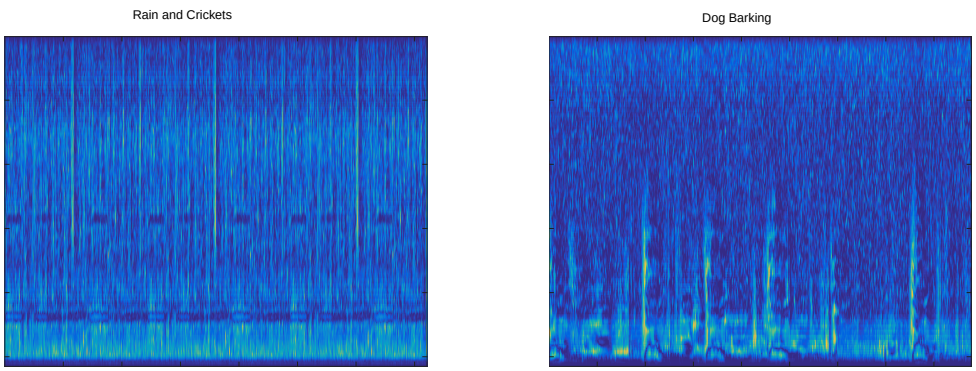


Figure 3. The saliency maps of rain and crickets and dog barking sounds obtained from Kayser's model

This model is tested through two environmental recordings collected from real environment. Sound a is the rain and crickets, the crickets are the salient sound, and the background noise (rain) has almost same intensity with the salient events. Sound b is dog barking with children talking in which the barking is the salient events. The results

are shown in Figure 3. From Figure 3 (a) we can find out that, when the intensity of background noise is high, the saliency map could not give a clearly representation of the salient sound events of cricket. The Figure 3 (b) pointed out that, when the intensity of background noise is low, this model could detect the sound saliency effectively.

In order to improve the detection accuracy of ASM, (Kalinli & Narayanan, 2007) presented the second ASD model, which extract two more features: the orientation and the pitch. 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. There are two hypotheses proposed by Shamma for the coding of pitch in the human auditory system: temporal and spectral (S. Shamma, 2001). In this model, the temporal hypothesis has been chosen to extract the pitch features and then project to the spectrogram frequency axis to obtain the feature map.

(Duangudom & 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 there are two main differences. The first is the acoustic features, where global energy, time modulation, spectral modulation and high temporal-spectral modulation are extracted in this model. The second is the processing schemes of the feature maps. First, the inhibition is performed to each individual map, resulting in the demotion of maps with no salient features. Then, the individual feature maps in each of the 4 categories are then combined into a “global” feature map for each class. At last, combined the 4 global maps through inhibition and summation to generate the final saliency map.

2.2.2. Improved Models

In order to improve the performance or auditory saliency detection, 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 & 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. Features have been selected in this ASM were: waveform envelope, spectrogram, rate, bandwidth, and pitch. All these features were obtained in eight scales. It should be noted that the waveform envelope and the pitch were kept in one dimension throughout whole processing, the rest of features were first computed in two dimensions. Followed processing steps were the same as Kayser's ASM to achieve the final temporal auditory saliency map. The peak in the saliency map represents a prominent event of the sound signals.

The experiment results derived from the three classical ASD models show that these methods can only achieve acceptable detection results when the salient sounds are short-term sound signals. For overcoming this drawback, (Botteldooren & 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, & 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.

Except for Itti's visual saliency map, there is another representative saliency map was presented in (L. Zhang, Tong, Marks, Shan, & Cottrell, 2008), called the Saliency Using Natural Statistic (SUN). This model measured saliency from natural image statistics, obtained from a collection of natural images. Based on this, (Tsuchida & Cottrell, 2012) proposed a novel auditory saliency map called the Auditory Saliency using Natural statistic (ASUN). ASUN uses the same method which has been applied in SUN to estimate the local statistics and compared it with learned statistics, in order to find if there are some differences between them. The differences could be treated as the sound saliency. Results showed that when the sounds were short time signals, it could reproduce psychophysical phenomena.

In order to understand how does human divert our attention in different voices over time, (De Coensel & 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 Kayser's ASM to obtain the saliency map and Time-Frequency masks for the spectrograms of each sound resources was calculated at the same time. Afterward, T-F masked spectrograms and auditory saliency map are combined to yield the saliency score of each acoustic signal. Based on these, the author proposes an attention model which based on the saliency scores calculation, and the winner-takes-all competition is implemented to identify the most salient sound events. The model was tested with the traffic recordings and the experimental results indicate that, this model can mask undesired sounds in the real environment.

Energy linear superposition theory was used in (Pan, Long, Cheng, & Chen, 2013) to detect the saliency of auditory. According to this theory, a mixed sound is the result

of multiple linear superpositions of individual sounds. Therefore, the energy of a salient sound could be obtained by subtracting the energy of background noises from the energy of the mixed sound signal in the energy domain. This model is similar to Kayser's saliency model but only consider the features of intensity and orientation to simplify the model. The linear combination was applied on the two feature maps to yield the ASM while a prominent area is pointed out on it. The author assumed that the background noise will not change in a short time, the energy of background noise could be estimated by taking a short period before and after the salient area of the sound signal on the auditory saliency map. Based on this, the theory described above was used to acquire the salient area on the final ASM which is the auditory saliency detection result of the proposed model. Experimental results proved that the proposed method could achieve high performance on detecting salient sound in a smooth and steady background.

(Schauerte & 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.

(Kim et al., 2014) considered the Bark-frequency loudness based optimal filtering for auditory salience detection and researched on the collecting annotations of salience in auditory data, in which linear discrimination was used. Though the experiment results shown 68.0% accuracy, the sound signals for validation are collected from meeting room recordings. This means that only indoor environment is considered.

Inspired by the research results of bird auditory system, a task-related sound

locating method through interaural time difference and interaural level difference was presented in (Mosadeghzad, Rea, Tata, Brayda, & Sandini, 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 environment. (J. Wang, Zhang, Madani, & Sabourin, 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 Kayser's 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 (J. 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.

2.3. Acoustic Deviancy Detection

One of the important aspects of our acoustic perceptual skills is auditory deviancy detection. This acoustic mechanism allowed human beings to percept the novel stimulate while regardless of the processes engaged in the focal task. It seems like the definition of deviancy detection is similar to saliency detection, in fact, they are different in nature. The main purpose of saliency detection is to identify those features in a scene are conspicuous based on their context and are salient, and could attract attention. While the main purpose of deviancy detection is to identify the unusual or deviant events when we focused on the objects or events which attract attention at first.

A similar ability, imitating this auditory awareness mechanism will greatly improve the efficiency of artificial perception in complex environment.

The current study mainly aims to reveal the theoretical basis and use the electroencephalograph (EEG) and the mismatch negativity (MMN) to find out how does auditory cortex process and responses the deviant signals. With carefully designed experiments, (Escera et al., 1998) point out that there are two different neural mechanisms in triggering involuntary attention to acoustic deviancy: a transient-detector mechanism activated by auditory deviancy, and a stimulus-change detector mechanism activated by deviant tones and novel sound events. The attention shift signals derived from the activation of the two mechanisms trigged an effective engagement of attention. These results indicate that, small changes in the acoustic environment capture attention involuntarily by activating the stimulus-change detector mechanism reflected in the mismatch negativity (MMN).

Through the study of anterior insula (AI) and considered it as a hub of a “salience network”, a possible framework of how does our brain response to stimulus is presented in (Menon & Uddin, 2010) for better understand brain mechanisms in important environmental stimuli detection tasks. This model helps to aggregate different findings into a common framework and suggests that AI could be a core component in cognition control. The author also proposes that a transient signal from the AI engages the brain’s attention, working memory and higher-order control processes while neglecting other systems that are not immediately task relevant could be a basic function within cognition control.

The mechanisms underlying human auditory perception of environmental sound is a fundamental principle in soundscape design. A computational model for soundscape analysis was presented by (Oldoni et al., 2013), with the goal of simulating how listeners would switch their attention over time between different sounds. In this model, there are three mainly processing stage: a) peripheral auditory processing, b) co-occurrence mapping of features, and c) modeling auditory attention. In the first stage, the sound wave is first transform to 1/3-octave band spectrogram followed by the same feature extraction strategy presented in (C. Kayser et al., 2005). Then, a measure for the

saliency of the sound at each timestep is calculated based on the scheme presented by(Kalinli & Narayanan, 2007), where the effects of spectral-temporal orientation and pitch are not considered. At last, a single saliency score at each timestep is calculated by summing all values of the saliency vector. In the second stage, an unsupervised learning strategy based on feature cooccurrence is used, which is implemented as a self-organizing map (an abstract model of topographic mapping in the sensory cortex). In the last stage, an excitatory-inhibitory artificial neural network (ANN), simulating the auditory cortex, is applied to identify sounds that acquired of the trained self-organizing map. Although this model does not provide abundant detail, it still complements already existing models of attention-based auditory scene analysis, promoted the understanding of the attention shift mechanism as well.

In (Kaya & Elhilali, 2013), a biologically motivated model which based on MMN and Kalman filters is proposed as a supplement to other sound relevant models that might need deviancy detection. Based on the “predictive coding” theory, in this model the MMN will be regarded as the representation of deviancy. To be specific, when a sound occurs and is different from the focal sound events, will elicit the MMN. The standard of the incoming sound feature is detected and for each detected stream, two Kalman trackers are triggered. One tracks the value of feature and one tracks the timing of the values. If no tracking Kalman filter has predicted this value will trigger the MMN. If the value has been predicted by a filter, then it will be compared with the time tracking Kalman filter, and the MMN will be elicited if the time occurrence is far from the prediction of the time filter. This model is tested by finding the deviant onset times of simple oddball paradigms and simple sound patterns.

Two parallel but separate lines of research on auditory novelty detection is presented in (Escera & Malmierca, 2014), in order to give a better understanding of the functional organization of the auditory system. The first line is human studies of the MMN, and the second line is animal studies of single neuron recordings of stimulus-specific adaptation (SSA). These two studies reveal that novelty detection should be a key principle consisting the auditory awareness, and the generation of MMN recorded from human studies show that when deviancy occurs with regard to a single physical

attribute of the acoustic input, a concatenation of processes taking place at different levels of the auditory system's hierarchy. Based on the experiment results and compared with several recent results in other works, the author finally proposal that the auditory novelty system should be organized in a hierarchical manner.

In (Escera, Leung, & Grimm, 2014), after review the evidence of three kinds of human brain response to deviant sounds along with animal studies on SSA, the author concluded that deviance detection is a basic principle of the functional organization of the auditory system. Furthermore, the phenomenon that conspicuous in complex environment cannot trigger the deviancy detection mechanism but MMN is elicited indicate that, regularity encoding based deviancy detection is organized in ascending levels of complexity along the auditory system.

By assessing the sensitivity of Middle-Latency Responses (MLR) components to deviant probability manipulations, the study of (López-Caballero, Zarnowiec, & Escera, 2016) further characterize the auditory hierarchy of novelty responses. MMNs and MLRs were recorded in 24 healthy participants, using an oddball location paradigm with three different deviant probabilities (5%, 10% and 20%), and a reversed-standard (91.5%). The differences in the MLRs elicited to each of the deviant stimuli and the reversed-standard are analyzed. The results verified that the deviancy detection occurred at the level of both MLRs and MMN. However, conspicuous differences for deviant probabilities only found in MMN. Which further pointed out that this process only occurs at higher stages of the auditory hierarchy.

(Liao, Yoneya, Kidani, Kashino, & Furukawa, 2016) present a study shows that the human pupillary dilation response (PDR) is sensitive to the stimulus properties and irrespective whether attention is directed to the sounds or not. Three experiments were conducted in this work, the PDR of subjects were recorded while they listened to the auditory oddball sequence. When the participants only listen to the noise oddball, their pupils expand for approximately 4 seconds, but no PDR for 2000 Hz oddball tones. When the participants were expose to visual oddballs along with auditory strange recordings, they separated the auditory or visual oddballs when trying to ignore stimuli from another modality. When visual and auditory stimuli were presented to the subjects

asynchronous, the pupils dilated to both kind of tones. These results point out that the PDR can be regarded as a measurement for detection of deviant auditory stimuli.

(Vachon et al., 2017) conducted a systematic investigation whereby the impact of verbal deviants and spatial deviants on verbal and spatial short-term memory was assessed. This study established that both verbal and spatial deviants can hinder both verbal and spatial order-reconstruction. This work suggested that this would seem intuitive because that, the warning capacity of the auditory cognition system should ensure the brain attends to the deviant events while ignoring the currently attended goal, the informational value of the task-irrelevant sound and any coupling between relevant and irrelevant information. The author finally concluded that the deviancy reflects a general form of auditory distraction as interference took place both within and across domains and regardless of the processes engaged in the focal task.

(Marchi, Vesperini, Squartini, & Schuller, 2017) presented a broad and extensive evaluation of state-of-the-art methods with a particular focus on novelty detection and recent unsupervised approaches based on RNN-based autoencoders. A broad evaluation on three different datasets is illustrated to present complete evaluation in the field of acoustic novelty detection. It is pointed out that RNN-based autoencoders outperform conventional methods in auditory novelty detection. Furthermore, combining the binary-LSTM autoencoder architecture with the nonlinear prediction scheme could achieve significant improvement in detecting accuracy.

In general, deviancy detection is a key characteristic of the auditory system that allows pre-attentive discrimination of incoming stimuli irrespective the ongoing constant stimulation. Hence, providing artificial intelligence with such auditory mechanism will effectively enhance its perceptual performance in real environment.

2.4. Modelling Auditory Cognition

Auditory cognition is becoming a hot issue in recent years, which can be applied in many arears such as remote surveillance and mobile devices. This problem is mainly

consisting of three components: sound events detection model, environmental sound classification model and decision-making model. Although appropriate frameworks for automatic speech recognition (ASR) and music information retrieval (MIR) have been well established by a growing number of researchers (Juang & Rabiner, 2005; Klapuri & Davy, 2007; H. Xu et al., 2018; Yakar, Litman, Sprechmann, Bronstein, & Sapiro, 2013), etc., the ESC research is still at the early stage. (Piczak, 2015b) has pointed out that environmental sounds are very diverse group of everyday audio events on account of considerably non-stationary characteristics that cannot be described as only speech or music. Hence, there is a strong need to establish suitable acoustic features and sound events categorization models for ESC tasks. Finally, after recognizing the sound events, they should be further categorized into two classes: valuable events and non-valuable events. It is because that, not all the environmental events present to be salient need acoustic attention, some salient sounds may also be the high-intensity noises relative to focal auditory tasks. That is reason why the decision-making system is required for establishing the artificial auditory cognition. Therefore, in this section we will introduce the state-of-art research works related to these three aspects.

2.4.1. Acoustic Features

According to the conclusion of (Chachada & Kuo, 2014), the feature extraction methods are established mainly based on two aspects: sound signal processing scheme and characteristics of features. For sound signal processing strategies, there are three commonly used schemes, which is framing-based processing, sub-framing-based processing and sequential processing. 1) In the framing-based processing scheme, sound signals are first divided into frames based on Hamming or Hanning window. Features are extracted from each frame and their combination is used as one feature set for training or testing. However, each frame gets a classification label, lead to successive frames may belong to different classes. Meanwhile, since some sound events are short-time signals and some are long, hence, it is hard to select a satisfied window length for all classes. These two aspects are the main drawbacks of this processing

scheme. 2) For sub-framing-based processing strategy, each frame obtained by framing-based processing are further segmented into shorter sub-frames, features are extracted from these sub-frames. The extracted features are concatenated together as a feature vector or averaged to represent a single frame, which used to train classifiers. This signal processing scheme present to be more flexibility in segmenting consecutive sounds based on labels of sub-frames. 3) The sequential processing strategy still divides sound into smaller segments, which is generally of 20-30 ms long with 50% overlap. The classified decisions are made based on features extracted from these segments. This scheme is unique in its aims to acquire the correlation of intersegment and the long-term variations of the sound signal, when compared with the other two strategies.

Sounds can be analyzed in both temporal and frequency domain. From a physical point of view, both representations from these domains provide different perspectives of the signal. Temporal domain information provides exact measurable feature of sound signal, such as the vibrations. Frequency domain features describe the nature of the physical phenomenon constituting the signal. Furthermore, on account of the assumptions that whether the sound signals vary with time or not, the features could be divided in to non-stationary features and stationary features (Cowling & Renate, 2003).

2.4.1.1.Stationary features

Stationary features including both temporal and spectral features, such as the Zero-Crossing Rate (ZCR), Short-Time Energy (STE), Sub-band Energy Ratio and Spectral Contrast, which are easy to compute and widely used in many arears (Gouyon, Pachet, & Delerue, 2000; Higashi, Kim, Jeon, & Ichikawa, 2010; Swee, Salleh, & Jamaludin, 2010). Cepstral features are also generally used in ESC, the most famous one is Mel Frequency Cepstral Coefficient (MFCC) with its first and second order derivations (ΔMFCC and $\Delta\Delta$MFCC), which is often used in human voice or music related audio signal processing scenarios, such as speech recognition and music genre recognition tasks. Other widely used cepstral features including Gammatone Filterbank Cepstral Coefficient (GFCC), Linear Predictive Cepstral Coefficient (LPCC), Homomorphic

Cepstral Coefficients (HCC) and Bark-Frequency Cepstral Coefficients (BFCC) (Burgos, 2014; Hu, Mitchell, & Pang, 2012; Ittichaichareon, Suksri, & Yingthawornsuk, 2012; Schafer, 2008; Zheng, Zhang, & Song, 2001).

2.4.1.2.Non-stationary features

However, real-life or environmental acoustic events have time vary characteristics, they are always non-stationary. Non-stationary features are referring to two categorizes, first is the time-frequency features derived from Short-time Fourier Transform (STFT) based spectrograms, or the features generated by Discrete Wavelet Transform (DWT) or Continuous Wavelet Transform (CWT) based scalogram. The second is Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP) based sparse domain features(Chu, Narayanan, & Kuo, 2009; Uzkent, Barkana, & Cevikalp, 2012). Moreover, despite the species of features, (Chachada & Kuo, 2014) pointed out that, combined acoustic features always perform better than single features in ESC tasks.

From the research works published in the past decades, we can notice that the MFCC is the most widely used acoustic feature in both speech, music and environment sound recognition problems. It is derived from STFT based spectrograms with framing-based signal processing scheme. This feature is originally developed for speech and music recognition and achieve outstanding performance in these applications. (Chia Ai, Hariharan, Yaacob, & Sin Chee, 2012) conduct a series experiments in order to find out the optimal configuration of MFCC and LPCC in speech recognition problems. The experiment results showed that 25 MFCC features present the best accuracy of 92.55%. (Ali, Tran, Benetos, & d'Avila Garcez, 2018) propose a method to combine the learned features derived form neural networks and the MFCC features for speaker recognition task, which can be applied to audio scripts of different length. (Ghosal & Kolekar, 2018) combined MFCC with several conventional acoustic features to train a convolutional long short term memory neural network for music genre recognition. The results indicate that this approach can achieve the state-of-art performance.

In addition, a considerable number of studies also reported that the robustness of

MFCC is not sufficient in noise-background while GFCC shows better performance and robustness (Zhao, Shao, & Wang, 2012; Zhao & Wang, 2013). GFCC is similar to MFCC, it is a sound feature for simulating human auditory characteristics as well. It mimics human auditory system which has different modalities of non-linear response to the different frequencies of signal components through a set of Gammatone Filterbank (Shao, Jin, Wang, & Srinivasan, 2009). It is also reported in (Chachada & Kuo, 2014) that GFCC has a strong ability in representing impulsive signals.

(Zhao et al., 2012) employ the GFCC in speaker identification system (SID), where computational auditory scene analysis (CASA) is applied to separate the background noise and speech. With systematic investigation, the author pointed out that, nonlinear log rectification is the reason why GFCC shows superior noise robustness compared with conventional features. Inspired by the characteristics of human peripheral auditory systems, (Adiga, Magimai, & Seelamantula, 2013) proposed a GFCC and wavelet based features, called GWCC. The extraction method is similar to that of the MFCC, with the difference of replacing the mel filterbank in MFCC with a Gammatone wavelet filterbank. The experiment results showed that the GWCC performed better than MFCC at low signal-to-noise ratios (SNR). (J.-M. Liu et al., 2013) use GFCC in cough recognition problems. The accuracy of GFCC comparing with MFCC is evaluated on a designed cough dataset following a 10-fold cross-validation, where weighted SVM is applied as the base classifier. After aggregating GFCC and MFCC, this model presents a better performance in cough recognition tasks.

The analysis of sound scenes or events is a relatively field of research in the context of sound signal analysis, meanwhile, the features used in speech or music processing often brings interesting insight on the content of environmental sound events. Hence, multiple researchers prefer to use these features and their combinations rather than develop new acoustic features in ESC tasks.

(Rakotomamonjy & Gasso, 2015) propose a novel feature for classifying audio scene, which show a good performance in capturing relevant discriminative informations. The novel feature has been obtained by computing histogram of gradients of a constant Q-transform followed by an appropriate pooling. The experiment results

on several datasets proved that this feature can achieve outstanding classification accuracy. (Adavanne, Parascandolo, Pertilä, Heittola, & Virtanen, 2017) present a long short-term memory (LSTM) recurrent neural network (RNN) based automatic sound event detection (SED) model. Where log mel-band energies, pitch frequency and its periodicity, and time difference of arrival (TDOA) in sub-bands are extracted to form the feature vectors for training the proposed SED model.

A considerable number of conventional sound event classification methods that mainly address local temporal-spectral patterns, (J. Ye, Kobayashi, & Murakawa, 2017) propose an aggregation scheme to combine both local and global acoustic features. In order to characterizing local patterns, the unsupervised feature learning method is performed. This model use dictionary to code representative patterns of sound events, followed by mapping to generate new features regard to the dictionary. Variability and recurrence are extracted as global features through long-term descriptive statistics. Finally, the mixture of experts model is exploited to aggregate the local and global features for classification. The experiment results indicate that this model can achieve superior performance compared with 3 other models.

(Lian, Xu, Wan, & Li, 2017) exploit modified GFCC in underwater acoustic target classification. The author found the conventional GFCC is not suitable for underwater acoustic events since the background sounds are quite different from environment. Therefore, a sum-of-squares approach is used to replace the rectangular window in primary feature extraction stage. The experiment results proved that the modified GFCC features are more robust than conventional features for underwater sounds.

A companion robot used in fire environments always work under low visibility conditions, where visual information is hard to be acquired. For solving this problem, the ESC techniques are applied in fire-fighting mobile robots by (Baum, Harper, Alicea, & Ordonez, 2018). In this system, the Mel-spectrogram, MFCC, chromagram of the power spectrogram, octave-based spectral contrasts and the tonal centroids are extracted as features to train the classifier. This model obtains classification results with an overall accuracy of 85.7%.

(Serizel, Bisot, Essid, & Richard, 2018) presented an overview of the different

blocks of a standard feature extraction method. The first step in most feature extraction techniques is the choice of a suited time-frequency representation. The performance of using such representations will be limited to the quality of the representation used for training. Therefore, it is needed to studies of the advantages and drawbacks of certain representation to accurately describe and discriminate the useful information in sound scenes. Moreover, the most frequently used hand-crafted features are also described. The features used for sound scene and event analysis are mainly inspired from speech, music or image processing. However, they are often limited to describing only specific aspects of the time-frequency information. It is pointed out that combining a large variety of different features is often required to improve performance over features taken in isolation.

2.4.2. Deep learning-based Environment Sound Classification

Support-vector machines (SVM) and Gaussian mixture model (GMM) are two widely used classifiers in both ASR, MIR and ESC tasks in the past decades(Shao & Wang, 2008; J.-C. Wang, Wang, He, & Hsu, 2006). However, these conventional classifiers are designed to model small variations which result in the lack of time and frequency invariance. In recent years, deep neural network-based models have been proved to be more efficient than traditional classifiers on solving complex categorize problems. 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.

(Mohamed, Dahl, & Hinton, 2012) applied the generative pre-trained input based deep belief networks (DBN) for acoustic modeling in phone recognition. (Gencoglu,

Virtanen, & Huttunen, 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. (Espi, Fujimoto, Kinoshita, & Nakatani, 2015) proposed a deep learning (DL) based acoustic event detection 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 order to detect the temporally overlapped environmental sound, (Cakir, Heittola, Huttunen, & Virtanen, 2015) propose a DNN based multi label neural networks use log-mel band energy as features for this problem. The DNN consists of two hidden layers, where maxout function and sigmoid function are applied as activation function for hidden layers and output layer, respectively. This system is compared with another model and improves the accuracy by 19% overall.

(Krizhevsky, Sutskever, & Hinton, 2017) first use the CNN in image recognition and outperform all the traditional methods in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). CNN has been successfully used for ASR (Palaz, 2015) and MIR(Ghosal & Kolekar, 2018). 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. Based on this, (Piczak, 2015a) first evaluated the performance of using CNN in ESC tasks. In this work, an ESC system consists of 2-layer CNN with max-pooling and 2 fully connected layers is proposed. Log-mel spectrograms are extracted as an auditory feature to train the neural network. The experiment results indicate that the classification accuracy of this model is 5.6% higher than traditional methods. Zhang et al.(H. Zhang, McLoughlin, & Song, 2015) propose to use CNN with smoothed and de-noised spectrogram image feature in sound recognition tasks. (Meyer, Cavigelli, & Thiele, 2017) present a CNN model using mel-

spectrograms as features. The performance of three neural network layers as classifiers are investigated, which is a fully connected layer, convolutional layer and convolutional layer without max-pooling. The results indicate that using convolutional layer as classifier outperform the model applying fully connected layer as the classifier. (Takahashi, Gygli, Pfister, & Van Gool, 2016) present a 6-layer CNN model for acoustic event recognition. In this work, the log-mel spectrograms with their first order derivation and second order derivation are extracted for each recording without segmentation. Then, multiple instance learning is applied and the softmax layer is replaced by an aggregation layer to aggregate the outputs of each network. The data augmentation is applied to prevent over-fitting and improve the robustness of the model. CNN has a strong ability to extract features directly from raw inputs, which has been verified in various image recognition problems. Based on this, (Pons & Serra, 2018) propose to use CNN to extract features from raw waveform and use SVM or extreme learning machines as classifiers in ESC tasks. The results denote that this architecture outperforms the CNN trained by MFCC. However, the work presented by (Dai, Dai, Qu, Li, & Das, 2017) show that the accuracy is only 70.74% when using raw waveforms to train CNNs as well. In this work, the problem of how many layers are the most suitable for CNNs has been studied. With considerable experiments, it is pointed out that deeper layers do not give better performance. Meanwhile, the results also indicate that using waveform just achieve an approximative performance of models using log-mel features.

Traditional CNN models have several drawbacks in auditory recognition. For example, pooling layers are generally applied in CNN models for feature dimensional reduction, however, these processes can lead to information loss and hinder the performance of neural networks. Therefore, a considerable number of works attempt to use improved CNNs for ESC tasks.

A sound events detection model consists of a stacked convolutional and recurrent neural network with two prediction layers is proposed by (Adavanne & Virtanen, 2017), where log-mel band energy is extracted as features. One of the prediction layers is for the strong label and one for predicting the weak label. A method is proposed to control

what the network learns from the weak and strong labels by different weighting for the loss computed in the two prediction layers. The experiment result indicate that this model can achieve acceptable detection accuracy. Dilated convolution layers are exploited for ESC (X. Zhang, Zou, & Shi, 2017; Chen, Guo, Liang, Wang, & Qian, 2019) to avoid the above-discussed obstacles. Several research works exploit CNN models which originally developed for image recognition tasks, and achieve outstanding performance in ESC as well. (Boddapati, Petef, Rasmusson, & Lundberg, 2017) the environment sound classification accuracy of AlexNet (Iandola et al., 2016) and GoogLeNet (Szegedy et al., 2015) are evaluated on UrbanSound8K, ESC-10 and ESC-50 (Piczak, 2015b) datasets. Spectrograms (Spec), MFCC and Cross Recurrence Plot (CRP) feature sets are extracted and concatenated as three-channel image feature to train both models. The experiment results indicate that the image recognition models could also obtain good taxonomic accuracy for sound recognition problems. (Tokozume & Harada, 2017) end-to-end ESC system using a convolutional neural network. In this model, raw waveforms are used as inputs and two convolution layers are applied to extract features. Then, three max-pooling layers are performed for feature dimensional reduction followed by two fully connected layers as the classifier. A VGGNet (Simonyan & Zisserman, 2014) based ESC system is presented (Z. Zhang, Xu, Cao, & Zhang, 2018), where the convolution filters are set to 1-D for learning frequency patterns and temporal patterns respectively. (Zhu et al., 2018) propose a CNN based model called WaveNet, which use multi-scale features to make CNN learns comprehensively information of environment sounds. First, features are extracted from one recording through the first convolution layer using three types of filter size. The second convolution layer uses corresponding pooling stride to equal the dimension of these features and then, the three features are concatenated to form the multi-scale features. This feature is further combined with log-mel spectrogram and perform better than other systems on ESC-50 dataset. The DS-CNN model presented by (S. Li et al., 2018) also uses raw waveform and log-mel spectrogram as inputs to train CNN based ESC system. The difference between WaveNet and DS-CNN is: the WaveNet combined two kinds of features together while in DS-CNN, two different CNN use raw waveform

and log-mel spectrogram as inputs respectively, and the outputs are fused by DS theory.

2.4.3. Artificial Auditory Perception

Life experiences proved that deviancy from sound events generally break into our conscious even they are not attended previously. However, it is also illustrated by every-day experience that, not all the deviant sound events are meaningful. Therefore, the detected and recognized deviant environmental sound events should be further identified whether they need attend or not. Decision making is a key component of cognition system of selecting an action or an event within a series of more alternatives (X.-J. Wang, 2008). Unlike visual cognition, physiological studies of decision making based auditory cognition are still at the theoretical research stage. Although multiple published works have claimed that their established models could mimic human auditory cognition processing, however, these models are just kind of primary simulating the basic functions of auditory cognition, such as sound event detection and sound scene analysis. These models are all lack of judgements about the content of detected events or stimulus and the ability to decide the action of next step. The auditory system can not only possess the ability to detect and classify the sound events, it also needs to make decision of following activation or reaction to the sound events. Therefore, they cannot be regarded as real auditory cognitive models.

2.4.3.1. Theoretical Research Works

The majority research works focus on auditory cognition are theoretical research, which try to answers how does our brain process the obtained stimulus and what is the neuronal underpinnings of auditory cognition. (Romo & Salinas, 2001) conduct a study on what components of the neural activity evoked by a stimulus are directly related to decision making, and how are they related. The experiment results suggest that the ability to make decisions occurs at the sensory-motor interface. (Roitman & Shadlen, 2002) study the neural correlate of gradual decision formation by recording activity

from the lateral intraparietal cortex (area LIP) of rhesus monkeys during a combined motion-discrimination reaction-time task. (Binder, Liebenthal, Possing, Medler, & Ward, 2004) conduct an experiment where the blood oxygenation signals in the brain of human participants were recorded when they were asked to identify speech sounds masked by varying levels of noise. The results provide evidence for a functional distinction between sensory and decision mechanisms underlying auditory events identification. Meanwhile, it is also pointed out that there is a link between inferior frontal lobe activation and response-selection processes during auditory perception tasks.

A review of human neuroimaging studies in conjunction with data analysis methods that can directly link decisions and signals in the human brain on a trial-by-trial basis is presented by (Heekeren, Marrett, & Ungerleider, 2008). (X.-J. Wang, 2008) present a review of decision making in recurrent neuronal circuits from four aspects which are the computations at the core of decision processes as well: 1) the cellular basis of temporal accumulation of information, 2) the termination conditions for a deliberation process in neuron, 3) reward-based adaptation, 4) stochasticity inherent in choice behavior (this is mainly about to study what is the representation of uncertainty in our brain and what are the intrinsic neuronal sources of randomness in choice behavior). These computations are the key component of decision-making. Hence, it is essential to know their neuronal underpinnings or a biological foundation of decision making. An overview of research works concerning the neural basis of auditory scene analysis is presented by (S. A. Shamma & Micheyl, 2010). Three most significant questions are summarized in this review: do auditory streams emerge below, in, or beyond the auditory cortex, the role of temporally coherent, and how does attention influence auditory stream formation with neural. After a comprehensively investigation, the author briefly answered these questions. For the first question, the perception of sound sequences such as those used in studies of auditory streaming emerges from interactions between the auditory cortex. For the second one, the grouping of temporally coherent responses across neurons tuned to different frequencies or different stimulus attributes. Finally, the abundance of descending (efferent) connections in the

auditory system provides ample opportunity for “top-down” influences, and makes it quite possible that effects of selective attention affect early stages of the neural analysis of auditory scenes.

In order to find out the specific and causal contributions of different brain regions in the ventral auditory pathway to auditory decisions, (Tsunada, Liu, Gold, & Cohen, 2016) let monkeys to decide whether an auditory stimulus contained more low-frequency or high-frequency tone bursts, and record from and microstimulated middle-lateral and anterolateral sites. The results indicate that anterolateral directly and causally contributes sensory evidence used to form the auditory decision. (S. J. Kayser, McNair, & Kayser, 2016) believe that the qualities of perception depend not only on the sensory inputs but also on the brain state before stimulus presentation. For proving such assumption, behavioral and EEG data in human participants performing two auditory discrimination tasks relying on distinct acoustic features are collected. They find that, power in task-specific frequency bands affected the encoding of sensory evidence while phase has no influence on decision.

2.4.3.2. Computational Modeling the Auditory Cognition

Considering the application of auditory cognition on artificial machines will greatly improve its ability of recognizing the surrounding environment. The theoretical research results have already answered the question about where and how the auditory cognition processing in our brain to some extent. Based on these works, multiple researches have engaged in establishing computational auditory cognition models in recent years.

(J. Wang, 2015) proposes a heterogeneous audio-visual information-based model for realizing artificial cognition, which consists of three main modules. The first module uses various saliency features obtained from both spectral and temporal domain to realize auditory saliency detection. A biologically inspired computational inhibition of return model is proposed to extract the salient temporal information, power spectral density is applied to extract spectral information. Then, a fuzzy vector

based acoustic feature is presented for real environment sound classification. The second module is about realizing a salient foreground object detection approach from visual channel. The third module is an information probability model based heterogeneous information fusion model, which fuse the salient auditory and visual information.

A novelty detection algorithm detects abnormal acoustic events to alert the user of a possible emergency is presented by (Principi, Squartini, Bonfigli, Ferroni, & Piazza, 2015). In this system, an acoustic novelty detector is employed in order to be able to deal with unknown sounds, thus not requiring an explicit modelling of emergency sounds. This detector is a machine earning model use Power Normalized Cepstral Coefficients, Critical Band-based Teager Energy Operator Autocorrelation Envelope and MPEG-7 as features with GMM as classifier. After an alert event is detected, the system integrates a VoIP infrastructure so that emergencies can be communicated to relatives or care centers. Two datasets are exploited to evaluate the efficiency, and the obtained results show that the adopted solutions are suitable for speech and audio event monitoring in a realistic scenario.

2.5. Conclusion

This chapter illustrates the overview of my research field along with the state-of-art techniques that inspire this thesis. It has illustrated the relevant research works and models with respect to this thesis in three aspects: 1) the review of auditory saliency and deviancy detection techniques which established for auditory cognition, 2) the review of the application of deep neural networks in sound signal recognition, where the neural network-based environmental sound classification techniques are the main research orientation, 3) the overview of research works focus on auditory cognition in either theoretical level or computational modeling level published in the past decades. Several distinct approaches and observations are presented, in order to provide the general consideration of the motivation of this work. Then, the discussion regarding the

state-of-art publications are connected to the problems that are researched in this thesis.

Based on the three above mentioned review works, there are three observations need to be emphasized. First of all, auditory cognition is one of the most important components of the human awareness system. The computational model of this system should consist of the sound events detection module, the recognition module and the following activation decision-making module. Saliency detection is the basic principle for auditory perception, which is the detector of auditory cognition. Aware of the most conspicuous objects or events will lead to a faster and simple procedure in the perception of the surrounding environment. Meanwhile, the deviancy detection mechanism could be regarded as a supplement to saliency detection, a bottom-up selection mechanism made up of both helping us to perceive the environment more precisely. However, the research issue of auditory deviancy detection is more complex than auditory saliency detection, because a sound event should be salient at first, then, it could be deviant. In other words, detected salient sound events should be judged whether it is deviant or not. This will increase the difficulty of computational modeling the detector of artificial cognition system.

Secondly, although appropriate frameworks for automatic speech recognition (ASR) and music information retrieval (MIR) have been well established by a growing number of researchers, the ESC research is still at the early stage. This is because the environmental sounds are a very diverse group of everyday audio events on account of the considerably non-stationary characteristics that cannot be described as only speech or music. Furthermore, the environment sounds do not have meaningful patterns or sub-structures, such as rhythm for music and phonemes for speech. It is reported that the use of these features generally fails to precisely describe the content of environment sounds, since they cannot comprehensively represent the information in environment mixtures individually, leading to the classification accuracy of ESC failing to reach the same level as visual events categorization. On the other hand, deep neural network-based models have been proved to be more efficient than traditional classifiers on solving complex categorize problems. Despite various research works attempt to use deeper neural networks to improve the taxonomic accuracy like (Boddapati et al., 2017;

Dai et al., 2017; S. Li et al., 2018; Z. Zhang et al., 2018), however, the classification accuracy presented in these works is still unsatisfactory. Hence, there is a strong need to develop appropriate auditory features and novel neural network models to achieve high categorization accuracy for ESC tasks.

Thirdly, a growing number of investigations and analyses have been made on decision-making based on auditory perception. However, the main achievements are almost all on the theoretical level, which tries to find out how does our brain process the obtained information and what is the neuronal underpinnings of auditory cognition like (Binder et al., 2004; Lotto & Holt, 2011). Only a few published works present applicable computational models. Meanwhile, it can be noticed that these models are either elementary simulation of human auditory perception (J. Wang, 2015) or developed for indoor conditions (Principi et al., 2015), which may be insufficient for environment perception. Consequently, in order to realize artificial auditory cognition for complex environment awareness, novel approaches need to be researched and established.

Chapter 3. Computational Modeling of Environment Deviant Sound Detection

3.1. Introduction

Eyes and ears are the two major sensory organs of the human perception system, and they cope with myriad stimuli of the surrounding environment almost all day. Receiving these tremendous amounts of stimulus, our brains is capable to extract the pertinent information constructing our cognitive awareness about the environment in which we evolve. Research works relating cognitive psychology (Frintrop, Rome, & Christensen, 2010) have shown that the human's saliency-based selective attention mechanism greatly contributes to human's perception of surroundings and in his actions' efficiency regarding his interactions with the environment. In other words, this cognitive perceptual mechanism acts as a foremost process in construction of our effective awareness about the surrounding environment, helping us to focus on the objects, sounds or events which is conspicuous to us and to reject those (objects, sounds or events) which appear as background noise regarding the target we deal with at a given time. Furthermore, it is a common experience that during we focus on one salient event, our attention can be involuntarily engaged by visual or acoustic changes occurring unexpectedly in the environment (Escera et al., 1998; Schröger, 1996). This attention shift phenomenon of our cognitive perceptual mechanism could also be introduced as deviancy detection.

Compared with visual signals, sound signals will enable mankind to be aware of and avoid danger beforehand or when human vision is not available in certain environment. However, most of the auditory selective attention mechanism-based technologies mainly focus on sound saliency detection (Duangudom & Anderson, 2007; Kalinli & Narayanan, 2007; Kaya & Elhilali, 2012; C. Kayser et al., 2005; J. Wang et al., 2015). The research on acoustic deviancy detection is still in the theoretical research stage (Menon & Uddin, 2010; Vachon et al., 2017).

Deviancy detection aims at recognizing situations in which unusual events occur. It seems like the definition of deviancy detection is similar to saliency detection. In fact, they are different in nature. The main purpose of saliency detection is to identify those features in a scene are conspicuous based on their context and are salient, and could attract attention. While the main purpose of deviancy detection is to identify the unusual or deviant events when we focused on the objects or events which attract attention at first. For example, when we listen to music at home, someone knocked on the door, the attention will shift from the music to the knocking. The deviancy detection mechanism could be regarded as a supplement to saliency detection and a bottom-up selection mechanism made up of both helping us to perceive the environment more precisely.

Anomalous sounds which could cause auditory attention shift possess two characteristics：1) It is salient in the entire sound clip; 2) It is deviant relative to the salient sounds that have been detected or focused before. Therefore, the purpose of our goal is to detect the sounds have both the above two properties and irrespective the prominent sound that has been already detected.

For the auditory saliency detection part, since the research results of neuropsychology (Itti et al., 1998) proved that visual and auditory perception channels have perceptual correlations in high-level perceptual processing. Thus, it's reasonable to expect that the perception of auditory saliency could be convert into the perception of saliency of the visual channel. Based on this, (C. Kayser et al., 2005) initially proposed an auditory saliency map for salient sound detection. Experiment results showed that this model could mimics several basic properties of human auditory perception mechanism. (Kalinli & Narayanan, 2007) proposed an innovative ASM in for improving the performance of Kayser's model, the new model added the orientation and pitch as new sound features. (Duangudom & Anderson, 2007) proposed another ASD model in which the time-frequency receiver domain model and adaptive suppression were used to provide the final auditory saliency map. In (Kaya & Elhilali, 2012), an auditory saliency map which treat the input signals as a one-dimensional temporal input was presented. In (Kim et al., 2014), a saliency detection model based on the classification result was presented. (Tsuchida & Cottrell, 2012) and (Schauerte

& Stiefelhagen, 2013) introduced their novel auditory saliency map to predict the saliency in soundscapes, based on the theory of statistics. In (J. Wang et al., 2015) proposed a composite system that combined parallel paths including: temporal analysis, spectral analysis and the image salience model. It is reported that this model provided better robustness to saliency detection especially in real noisy soundscapes then conventional methods.

For the deviant sound detection part, the current study mainly aims to reveal the response and mechanism of auditory cortex to deviant sound through electro-encephalograph (EEG) and mismatch negativity (MMN) auditory evoked potential. It is reported in (Escera et al., 1998) that small changes in the acoustic environment capture attention involuntarily by activating the stimulus-change detector mechanism reflected in the mismatch negativity (MMN). Through the study of the anterior insula (AI) which is considered as a hub of a “salience network”, a network model is presented in (Menon & Uddin, 2010) for better understand brain mechanisms in important environmental stimuli detection tasks. Two parallel but separate lines of research on auditory novelty detection is presented in (Escera & Malmierca, 2014) and indicated that auditory novelty system should be organized in a hierarchical manner. In (Escera et al., 2014), after reviewing the evidence of three kinds of human brain response to deviant sounds, the author concluded that deviance detection is a basic principle of the functional organization of the auditory system. By assessing the sensitivity of Middle-Latency Responses components to deviant probability manipulations, the study of (López-Caballero et al., 2016) further characterized the auditory hierarchy of novelty responses. In (Kaya & Elhilali, 2013), a biologically motivated model is proposed to building a computational model of MMN based on Kalman filters. This model is tested by finding the deviant onset times of simple oddball paradigms and simple sound patterns. The study of the relationship between human pupillary dilation response (PDR) and deviant auditory stimuli (Liao et al., 2016) showed that a salient event which is deviant from the background attracts attention and reflected in the PDR. The experiment results presented in (Vachon et al., 2017) demonstrate that the deviation effect reflects a general form of auditory distraction as interference took place both

within and across domains and regardless of the processes engaged in the focal task.

Motivated by the shortcomings and limitations of previous research works from both auditory saliency detection and deviancy detection, a computational model to capture the deviant salient-sound in the real environment which mimics human auditory attention shifting mechanism. This approach is based on the detection of deviant salient-sounds in the temporal domain combined with the frequency domain saliency detection. Then it presents the detected deviant sound in the image domain at last. The model first obtains the local salient sounds in the time domain through a combined feature of Gammatone Filterbank Cepstral Coefficient (GFCC). Then, an entropy-based analysis method is applied to find the sound with deviancy which elicit the acoustic attention shift. Moreover, the sound saliency in the frequency domain is derived from the Power Spectral Density (PSD) based frequency saliency detection method and been considered as frequency deviancy of sounds. Finally, in the opponent color space, the gammatone spectrogram blue-yellow channel information is calculated as the indicator to present the deviant salient-sound which lead to the auditory attention shift.

3.2. Overview of the Approach

The environmental sound signals are varying in both temporal domain and frequency domain while the auditory deviancy detection has some similarity with auditory saliency detection to some extent. Hence, we will analysis the saliency and deviancy of sound signals in both of these domains. Since MFCC has been well studied in speech recognition and made great achievements, many researchers choose MFCC as features for sound saliency detection and sound event detection (Adavanne, Parascandolo, et al., 2017; Parascandolo, Heittola, Huttunen, & Virtanen, 2017; McLoughlin, Zhang, Xie, Song, & Xiao, 2015; Takahashi, Gygli, & Van Gool, 2017). However, many studies also reported the robustness of MFCC is not sufficient in noise-background while GFCC (Gammatone Frequency Cepstral Coefficient) shows better performance and robustness (Zhao et al., 2012; Zhao & Wang, 2013). GFCC is similar to MFCC and it is a sound feature extraction method for simulating human auditory

characteristics as well. It mimics human auditory system which has different modalities of non-linear response to the different frequencies of signal components through a set of Gammatone Filterbank (Shao et al., 2009). It is also reported in (Chachada & Kuo, 2014) that GFCC has a strong ability in representing impulsive signals. Therefore, the GFCC feature of sound signals is chosen to represent human auditory perception model for temporal saliency detection.

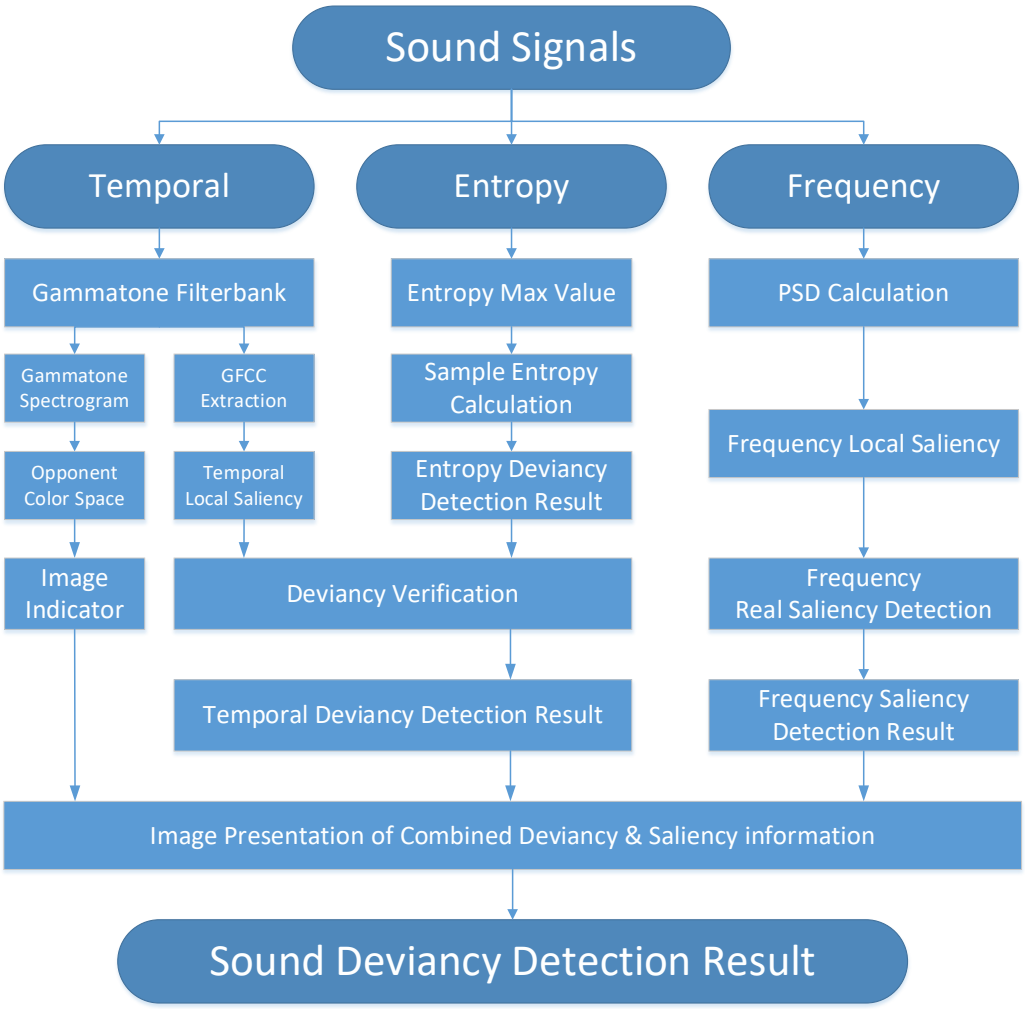


Figure 4. The proposed auditory deviancy detection model

Afterwards, since entropy could measure the uncertainty of a signal while deviancy could be considered as the most surprising component of a signal, hence, an entropy-based deviant salient-sound detection method is proposed to highlight the sound with deviancy in the temporal domain. Then, we calculate the Power Spectral Density (PSD)

of the sound signals to obtain the salient information derived from the frequency domain. Furthermore, the Gammatone Filterbank spectrogram is acquired in the opponent color space to eliminate background noise while prominent the salient part in the image domain. Finally, integrating the salient information both in temporal domain and frequency domain to point out the deviant salient-sounds, and present them in the image domain. The overview structure of the proposed model is shown in Figure 4.

3.3. Heterogeneous Deviancy Features Extraction and Fusion

3.3.1. GFCC

The input sound is first decomposed into the time-frequency spectrogram through a bank of Gammatone filters in our model. Gammatone filterbank (GF) is derived from psychophysical and physiological observations of the auditory periphery and this filterbank is a standard model of cochlear filtering (Zhao et al., 2012). GFCC is similar to the MFCC, the main differences are the non-linear rectification step before DCT where MFCC uses log operation and GFCC uses cubic root, and MFCC use log scale where GFCC is based on the ERB scale (Glasberg & Moore, 1990):

$$ERB(f) = 24.7 \times (4.37 \frac{f}{1000} + 1) \tag{0.1}$$

The GF impulse response in the time domain is shown as follows:

$$Gamma(n) = \alpha n^{\gamma-1} e^{-2\pi bn} \cos(2\pi f_c n + \phi) \tag{0.2}$$

where f_c is the center frequency, ϕ is the phase of the carrier, α is the amplitude, n is the order of the filter, b is the bandwidth of the filter and t is time. The relationship between b and f_c is:

$$b = 24.7 \times (4.37 f_c / 1000 + 1) \tag{0.3}$$

The input signal is processed by a n-channel Gammatone filterbank (GF). Its center frequency is usually setup between 50 Hz and 8 000 Hz. This center frequency is equally distributed on the ERB scale and the filters will have wider bandwidths with higher center frequencies. After processing by the filters, the signal still retains its

original sampling frequency. Therefore, alone the time dimension we refined the n-channel filter response to 100 Hz. This yields a corresponding frame rate of 10 ms and the magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation (Zhao et al., 2012):

$$g_m(i) = \left| g_{d-s}(i,m) \right|^{1/3} \tag{0.4}$$

where n is the number of filters and $i = 0 \ldots n-1$ and M is the number of frames and $m = 0 \ldots M-1$. The $g_m(i)$ form a matrix which represent a variant of cochleagram. Each frame of the cochleagram is a GF feature. When compared with spectrogram has the linear frequency resolution, cochleagram has the advantage of its resolution is better in the low frequency than the high frequency.

The GFCC extraction method is similar to the MFCC extraction which applies the discrete cosine transform (DCT) to $g_m(i)$ for dimension and irrelevant components reduction. The dimension of GFCC is set to 22 in this chapter.

$$GFCC = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} g_m(i) \cos\left(\frac{j\pi}{2N}(2i+1)\right) \tag{0.5}$$

3.3.2. Temporal Deviancy Detection

Human beings always intend to be attracted by the sounds with higher frequency components or higher loudness level. This phenomenon could be explained by the theory of the inhibition of return (Posner, Rafal, Choate, & Vaughan, 1985) and the conclusion presented in (Southwell et al., 2017) showed that attention mechanism prefers to perceive novel stimuli in the environment by an involuntary attention mechanism for efficiently percept the environment. This is the reason why we are sensitive to the emerging sound signals which are deviant to a current salient sound has been attended.

To mimicking the acoustic saliency detection of human beings, the GFCC of sound is considered to be the main representation of human hearing system and the salient sound is indicated by the peak value of GFCC curve. The GFCC curve is computed by

sum each row of GFCC matrix. Since the sound signal is first processed by a n-channel GF, then each row is the GFCC feature of different GF channels. The GFCC curve is the sum of the GFCC of each channel with normalized to 0 and 1, which is defined as:

$$GFCC_{curve} = \sum_{i=1}^{n} GFCC(i) \tag{0.6}$$

From the variation of sound signal in the temporal domain we can see that the GFCC curve could reflects the bionic characteristic of the human auditory perception. Meanwhile, the peak points of the GFCC curve could be regarded as the local salient components of a sound signal in the temporal domain.

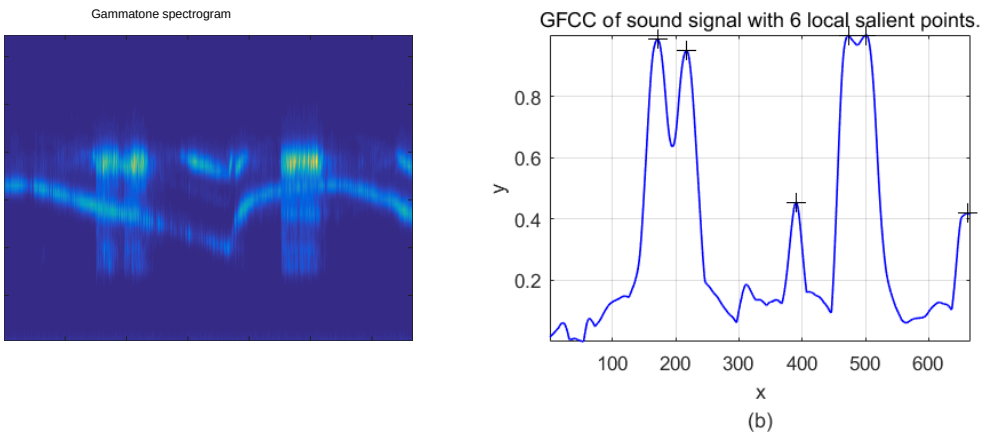


Figure 5. a) The GF spectrogram of sound example, b) The local saliency detection results of temporal domain

Figure 5 a) shows the GF spectrogram of a sound sample which is recorded in the real environment. The sound sample consists of a siren exist from the beginning to the end and two consecutive track honks. The siren and track honks are both salient compared to the background noise. It is obvious that in this example, the siren is the first salient component which attracts auditory attention and the track honks are the deviant salient-sounds which elicit the auditory attention shift. However, through Figure 5 b) we can notice that the GFCC curve could not identify the horn sounds while ignoring other peak points. Therefore, we propose an entropy-based sound deviancy detection method to locate the true deviant salient-sound in the temporal domain.

In Shannon's information theory, the concept of entropy is a measurement of uncertainty associated with a random variable (Shannon, 2001). The saliency components could be considered as a surprising component compared with its temporal neighborhoods within a time period. Since the deviant salient-sound should be surprised to the precepted salient sounds, we can image that it would presents a more uncertain value than the previous salient components in entropy domain. Hence, the highest Shannon entropy value could be considered as the deviant salient-sound of a sound signal. Here, the short-term wavelet packet Shannon entropy approach is applied to represent and estimate the saliency characteristic of real sound signals (J. Wang et al., 2015). The sound signal is divided into short-term frames with overlap of 50% and the Shannon entropy of each frame is calculated to represent the average change of the sound signal. The j_{th} frame of signal S is defined as:

$$E(S_j) = -\sum_i s_{i,j}^2 \log(s_{i,j}^2) \tag{0.7}$$

The short-term Shannon entropy of the entire signal is:

$$E(S) = \sum_{j=1}^{n} E(S_j) \tag{0.8}$$

We normalized the entropy to 0 and 1. Figure 6. shows normalized the entropy of the sound signal while the max value of entropy is pointed out. However, as we illustrated above, the consecutive track honks which appeared twice would elicit auditory attention shift. Obviously, employing the max value of entropy as representation of the deviant sounds is still inefficient for deviancy detection. Since the deviant sound may appear continuously and should have the same attributes while the surprise value of rest deviant sounds is similar to the first one. Therefore, only if all the deviant components are identified, we can say that the deviancy of a sound signal have been detected. Here, a sample entropy-based method in order to find out the real deviant sounds is proposed.

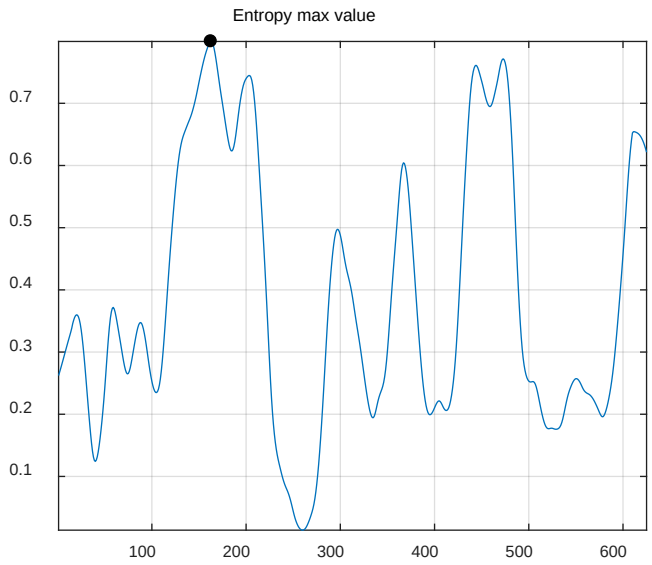


Figure 6. The max value of normalized Shannon entropy of sound sample

Sample Entropy (SampEn) (Richman & Moorman, 2000) is a traditional methods of measuring complexity, which determines the probability of finding specific patterns or resemblance between time series to examine the irregularity or the predictability of one particular time series. SampEn measures the negative logarithm of the conditional probability that two sequences that are similar for m points remain similar at the next point, within a tolerance r, the sample entropy is given as:

$$SampEn(m,r,N) = -\ln\left[A^m(r)/B^m(r)\right] \tag{0.9}$$

Where, $A^m(r)$ is the probability that two sequences will match for $m+1$ points while $B^m(r)$ is the probability that two sequences will match for m points, N is the time series of a signal.

Since SampEn measures the complexity of time series, as the background noise of a signal is strong, the value of SampEn will be large. Conversely, when the background noise is weak, the value of SampEn will be small. To be specific, when the deviant sounds could be precepted in high-intensity background noise with previous salient sounds, their values of entropy should be similar to the entropy value (max value) of the first detected deviant sound. Otherwise they could be masked by the background noise or by the current salient sounds which already attracted auditory attention. On the

other hand, when the deviancy is first precepted in low-intensity background noise, the entropy value of rest deviant sounds could change slightly wider than in high-intensity background noise situation. Moreover, in low-intensity background noise environment, the deviant salient-sound might be the only exist salient sounds. Therefore, the expectation value of deviancy in entropy domain can be calculated as:

$$D_i = \begin{cases} 1, & if\ E_i \geq \max(E(S)) \times e^{-SampEn} \\ 0, & else \end{cases} \tag{0.10}$$

Where E_i denotes the value of each point in entropy domain, $max(E(S))$ is the max value of entropy which also represents the first deviant salient-sound, D_i is the point which represents the rest deviancy point in entropy domain. Eq. (3.10) gives the deviancy detection principle in entropy domain. Since the range of SampEn is 0 to 1, it is obvious that when the complexity of sound signal is high, only those points which have similar value with $max(E(S))$ could represent the deviancy.

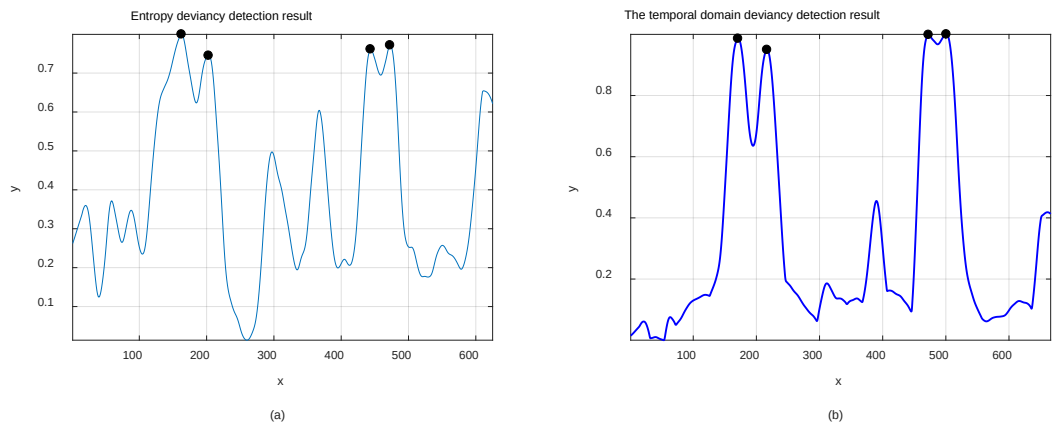


Figure 7. a) The entropy deviancy detection result; b) The deviant salient-sounds detection results in temporal domain

If the complexity of sound signal is low, the values of the remaining abnormal sounds may fluctuate over an acceptable range. The corresponding time of each detected entropy deviancy point could be acquired as the entropy is calculated frame by frame. Therefore, if the local saliency points in temporal domain matches the deviancy points in entropy domain which means the sounds appeared in this period

possess both saliency and deviancy, these points are the representation of deviant salient-sounds in the temporal domain.

3.3.3. Frequency Saliency Detection

As we mentioned above, one of the characteristics of deviant sounds is saliency. Since the environment sounds are non-stationary, so merely via the saliency and deviancy detection in the temporal domain to acquire the deviant salient-sounds is inadequate. Thus, obtaining the salient frequency component as a complementary part to saliency detection is also necessary. Hence, we propose a Power Spectral Density (PSD) domain saliency detection method. The PSD estimation results give the energy distribution of sound signals at different frequencies, so the average of PSD estimation represents the average level of sound frequency change in the spectral domain. Therefore, the points of the PSD curve which are greater than the mean value of PSD estimation can be regarded as the potential salient components of sounds in the frequency domain. It can be illustrated by:

$$P_i = \begin{cases} 1, & P_{peak}(i) \geq P_{avg} \\ 0, & P_{peak}(i) < P_{avg} \end{cases} \tag{0.11}$$

Where $P_{peak}(i)$ is the maximum point, P_{avg} is the mean of PSD curve and P_i is the maximum point greater than P_{avg}.

Related research work pointed out the frequency range that the human auditory system can perceive is usually between 20 Hz and 20 kHz, but only a handful of people can hear the upper and lower frequency limits. For most adults, the frequency range that can be heard in real life is between 40Hz and 16 kHz. Thus, a salient frequency distribution band is defined from 40 Hz to 16 kHz, a conspicuous point below or above the frequency band will also be classified as non-saliency. According to the frequency masking we can know that a strong pure tone will mask weak tones that sound simultaneously in its vicinity. Meanwhile, a pair of sounds with different loudness can be distinguished if the physical level increases by 10 dB (Petit, El-Amraoui, & Avan, 2013). Moreover, considering the concept of the critical band and computational cost,

each P_i will be compared with P_{i+1} and P_{i-1} if their frequency gap is less then 1kHz to identify the P_i with real saliency. Hence, the final frequency salient point $S_{S,i}$ could be obtained by:

$$S_{s,i} = \begin{cases} 1, & Px_i - \forall(Px_{i-1}, Px_{i+1}) > 1kHz \\ 1, & Px_i - \forall(Px_{i-1}, Px_{i+1}) < 1kHz \ and \ P_i - \forall(P_{i-1}, P_{i+1}) > 10dB \\ 0, & else \end{cases} \tag{0.12}$$

Where Px_i, Px_{i-1} and Px_{i+1} is the location in frequency axis of P_i, P_{i-1} and P_{i+1}. The first condition of Eq. (3.12) means when P_i is the only point within the range of $Px_i \pm$ 1kHz, it would not be masked. Therefore, those salient points with $S_{S,i} = 1$ are the real salient frequency components. If all of these points are equal to 0, it means that no salient frequency component exists.

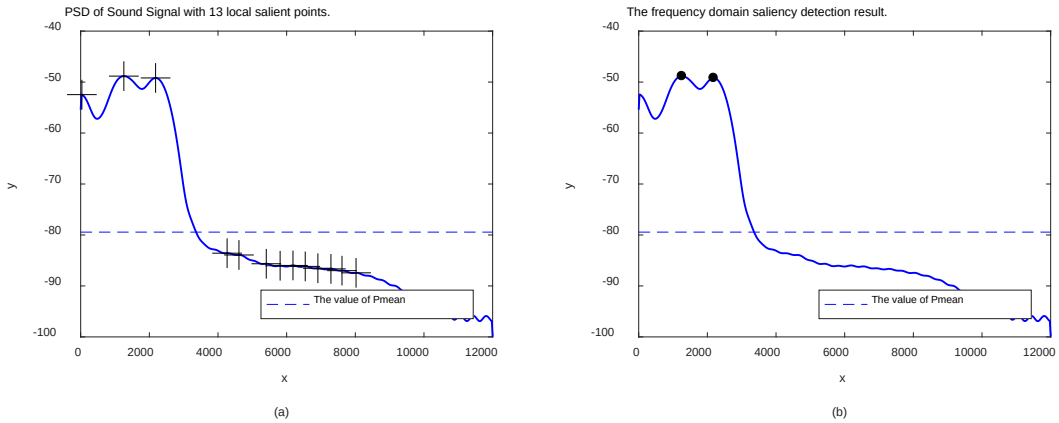


Figure 8. a) The frequency domain local saliency detection result; b) The frequency domain true saliency detection result

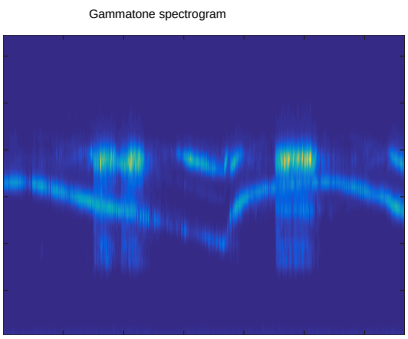
3.3.4. Image Indicator

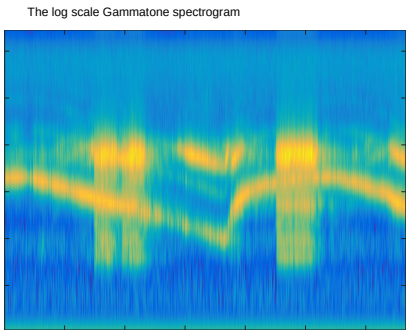
The original spectrogram is transformed into the log scale to depress the effect of background noise and emphasize the salient sound signal components. From Figure. 6(b) we can see that the log scale gammatone spectrogram of a sound signal which mainly consists of blue, green and yellow where yellow denotes the salient time-frequency component. The yellow components are much easier to be perceived by the human visual perception system than the components presented in blue which represent

the background noise. However, the representation of color in traditional RGB color space is not the best choice for human visual perception system (Evangelopoulos, Rapantzikos, Maragos, Avrithis, & Potamianos, 2008). Meanwhile, the computational efficiency will be affected while the indication may not be obvious for further processing if the colorful log scale gammatone spectrogram is used to indicate the deviant sounds. Therefore, we prefer to simplify the log scale gammatone spectrogram in the opponent color space (van de Sande, Gevers, & Snoek, 2008). There were three types of photo receptors: white-black, yellow-blue and red-green (Anwer, Vázquez, & López, 2011). As the component of log scale gammatone spectrogram with yellow color are more salient to human among background noises with blue and green. Hence, we can apply the concept of opponent color space to reduce the three-layer log scale image into a single-layer image for a better indication of sound deviancy:

$$S_I = (I_c(R) + I_c(G) - 2I_c(B)) / \sqrt{6} \tag{0.13}$$

Where I_c is the log scale gammatone spectrogram of a sound signal. $I_c(R)$, $I_c(G)$ and $I_c(B)$ are the red, green and blue color values of the pixels in the original RGB color space of I_c. S_I is the image indicator. Finally, the combined deviancy information and saliency information are presented on the image indicator to highlight the deviant salient-sounds.





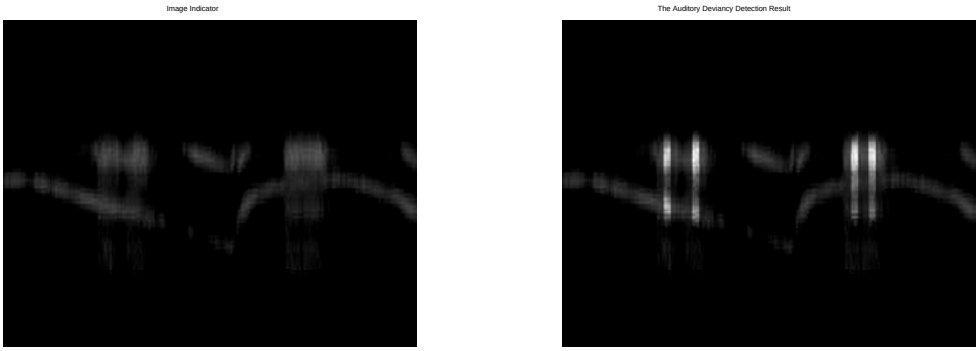


Figure 9. a) The gammatone spectrogram of sound signal; b) the log scale gammatone spectrogram of sound signal; c) the image indicator of sound signal; d) the deviant salient-sounds presented in the image indicator

3.3.5. Verification of the Proposed Model

To verify the performance of the auditory deviancy detection framework proposed in this paper in dealing with the actual environment sound signals, an experiment was conducted using three samples recorded in different soundscapes. The background noise of these sound samples consists of environment sounds or urban sounds. All the samples have a salient sound could attract auditory attention that always exist. While a deviant salient-sound also exists, which lead to the auditory attention shift in each sound sample.

Specifically, the sound sample A is a 11s sound recorded in a restaurant, the human talking voices is the salient sound attract auditory attention at first. The auditory attention shift evoked around 5s caused by breakage of window. Sound sample B is the sound of two owl's hooting recorded in the natural environment. This sample includes the sounds of owls as salient sounds and several other birds chirping as background noise. The difficulty of deviancy detection is to distinguish the second owl hooting from the first one. Sample C is recorded on a rainy day that consist of raining as background noise while the buzzing is the only salient sound in this sample. The sample C used here is to verify if there is no deviancy, whether the proposed model could detect the auditory saliency or not.

These sound samples are representatives which include recordings in different soundscapes. Hence, we could verify the performance of the proposed framework by the samples deviancy and saliency detection results. The frame length of $E(S)$ is 512 with an overlap of 256 and the scale of gammatone filterbank is 23. The GFCC curve and entropy curve are all normalized and smoothed to stabilize the detection of peak points in each domain.

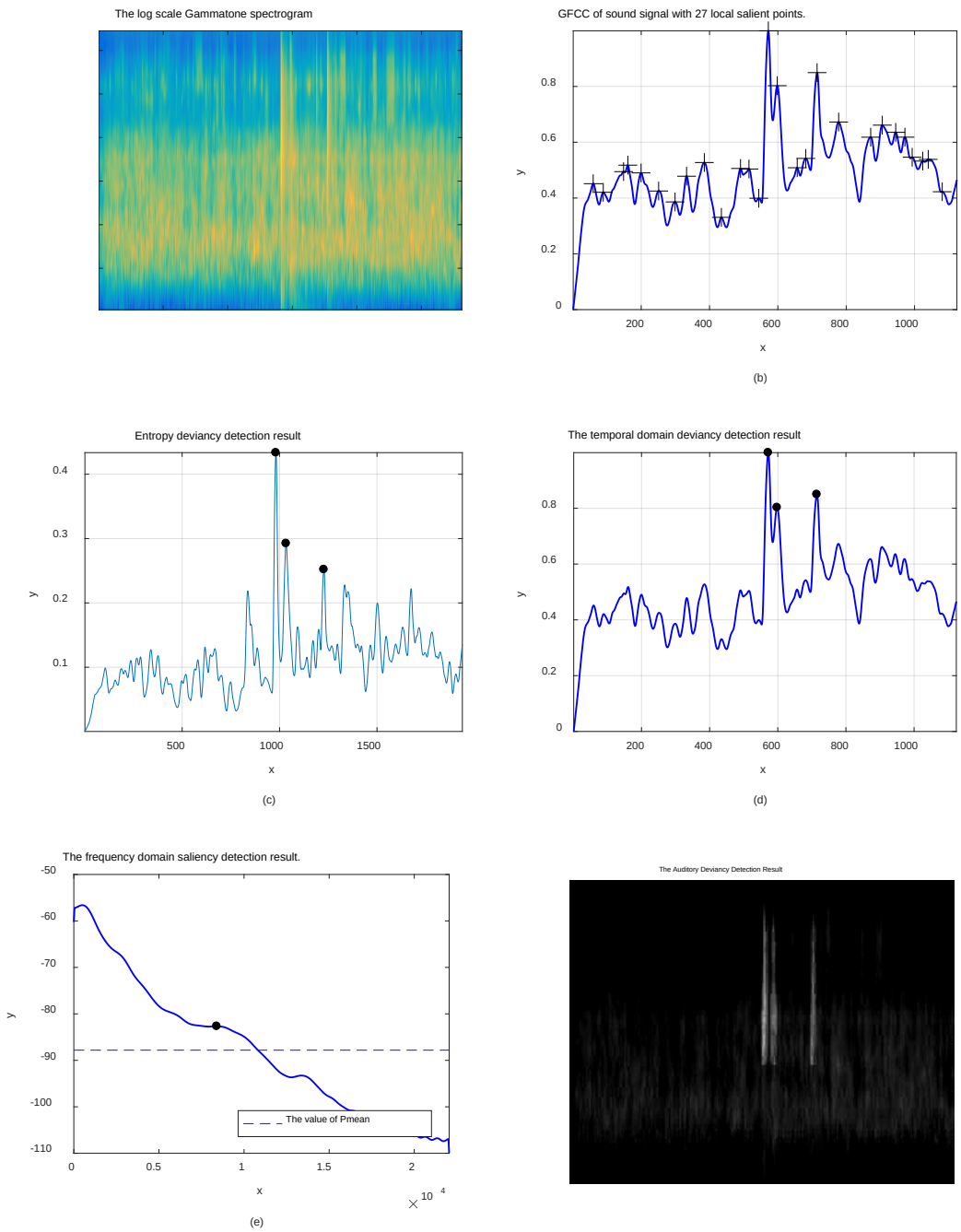


Figure 10. The result of each domain and the deviancy detection result of sample A.

The results are shown in Figure. 10-12 The subgraphs (a) - (f) are the log scale Gammatone spectrogram, temporal local saliency detection result, entropy deviancy detection result, temporal domain deviancy detection result, frequency domain saliency detection result and the sound deviancy detection result respectively.

Figure 10 shows the process and detection result of sound sample A. Obviously, from Fig. 10 (a) we can see that the human talking voice is salient to the background noise while the deviancy appeared around 6s is almost masked by the salient sound. The temporal local saliency detection could not identify the deviancy. However, after comparative analysis of entropy domain deviancy detection result, the temporal deviancy is successfully detected. Then, as illustrated in Figure 10 (e), the frequency domain true salient point is identified through Eq. (11) while the mismatch points are eliminated. Finally, the auditory attention shift caused by breakage of window is accurately presented in Figure 10 (f).

The auditory deviancy is also detected in sound sample B and the process and result is showed in Figure 11. The deviant salient-sound which cause the human auditory attention shift is the second owl hooting appeared for about 6 seconds. The difficulty of deviancy detection in sample B is that the deviant salient-sound is overlapped with the first owl hooting. Meanwhile, they all sound from the same species which means the two sounds have the same features and properties. From Figure 11 (a) and (b) we can see that the second owl's hooting is not presented clearly. The reason is that this sound is overlapped by the first owl's hooting and is hard to identify in temporal domain and frequency domain. However, from Figure 11 (c) and 11 (d) it can be found that, through the proposed entropy-based deviancy detection method, it could successfully find the deviant sounds since it is a novel sound in this sample and has a higher uncertain value in entropy domain. Then, Figure 11 (e) illustrates that the frequency components of background noise around 2.4 kHz always exist have been accurately eliminated. The true deviant sound is correctly presented in the image indicator.

The deviancy detection results of sound sample C are shown in Figure 12. This sound clip has no deviancy while the buzzing is the salient sound compared to the sound of rain. Despite the background noise do not show a high level in log scale gammatone

spectrogram, however, the noise almost masked the buzzing when listening to this snippet. Nevertheless, the most salient sound has been accurately detected, it is clearly demonstrated in Figure 12 (f) that the buzzing around 3.7 kHz is correctly highlighted. In other words, the proposed model is also applicable in salient sound detection tasks.

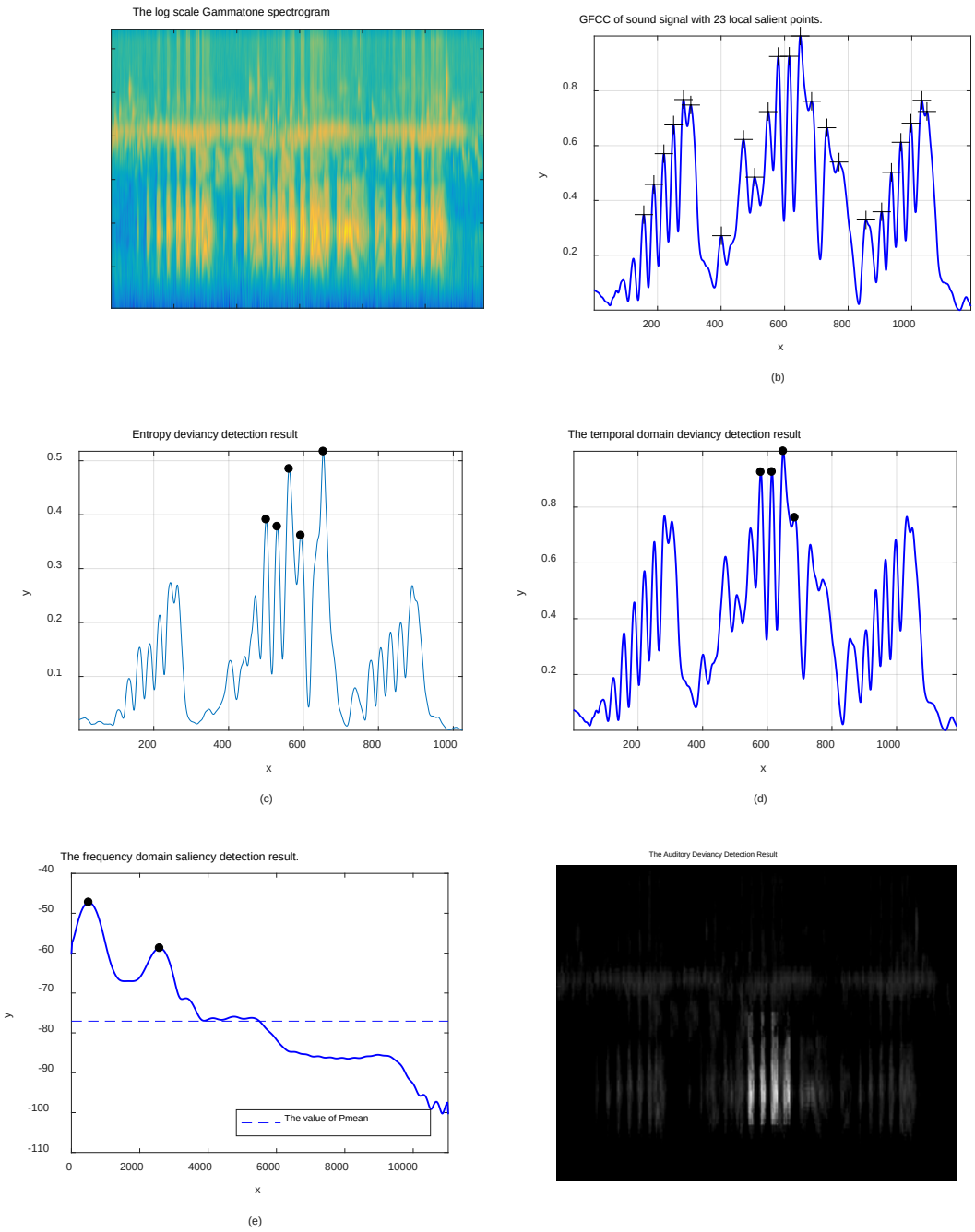


Figure 11. The result of each domain and the deviancy detection result of sample B.

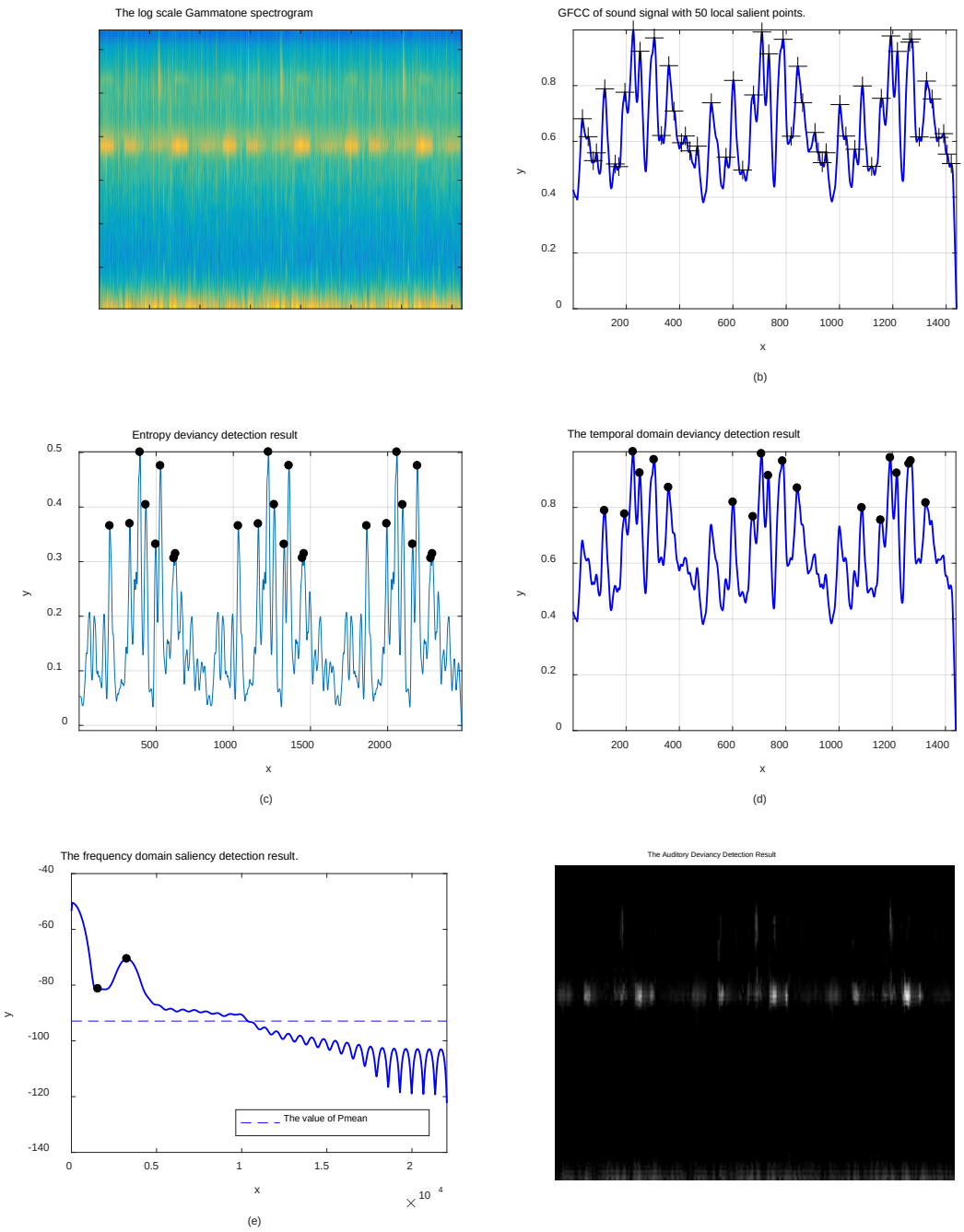


Figure 12. The result of each domain and the deviancy detection result of sample C.

Since the GFCC curves presented in Figure 10. (d), Figure 11. (d) and Figure 12. (d) show a considerable performance in salient sound representation while inhibit the background noise, it has been proved that the GFCC is a robust feature for representing the unstable environment sounds while it is an appropriate choice in our model. The detection results derived from temporal domain deviancy verified the efficiency and accuracy of the proposed entropy-based deviant salient-sound detection method.

Meanwhile, the frequency saliency detection results of these sound clips illustrate that no authentic salient points exist after using the frequency saliency verification in Eq. (3.12). The detection results of sound sample C demonstrate that the model could also be exploited for saliency detection tasks. Therefore, it can be concluded that the proposed model could effectively simulate the human auditory attention shift mechanism.

3.4. Experiments

3.4.1. Experiment Set Up

To test the validity of the model in a more quantitative manner, a set of 180 recordings of sound snippets which are synthetic from various sources including the Freesound database and the SoundBible database. Each recordings of our own database is a synthetic mixture using isolated sound events from the two above mentioned databases. The synthesized sounds are consisting of three components: background noise, salient sound which always exist and the deviant sounds which could cause auditory attention shift. Scenes were normalized based on the root mean square (RMS) energy of the loudest 20% and 60% of each wave file for creating 90 weak background noise instances and 90 strong background noise instances. All the synthetic mixtures have the same sample rate: 44100 Hz and sample bit: 16 bits. An overview of all components included in this dataset is given in Table 1.

We exploit Event-based metrics to compare system output and corresponding reference event by event (Mesaros, Heittola, & Virtanen, 2016). It is a widely used measurement to identify the efficiency of classification and recognition systems (Adavanne, Pertilä, & Virtanen, 2017; Cakir et al., 2015). Event-based metrics have no meaningful true negatives, except in the case when measuring actual temporal errors in terms of length, in which case the total length of time segments where both model output and reference contain no active events is measured. As the evaluation metric, F1 score is calculated inside each sound sample. The statistics except true negative are

defined as follows:

- true positive: a sound in the model output that has a temporal position overlapping with the temporal position of a sound with the same label in the reference.
- false positive: a sound in the model output that has no correspondence to a sound with same label in the reference.
- false negative: a sound in the reference that has no correspondence to a sound with the same label in the model output.

The reference is the deviant sounds in each instance. For example, the sound snippet one is synthesized from Rain Forest (20%), Turkey and Dog Barging. The reference of this clip is the dog barging which appeared 4 times. Then, the true positive should be the position highlighted in image indicator that have the same position of the reference. Precision, Recall and F-score (Rijsbergen, 1979) are calculated as:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F_{score} = \frac{2PR}{P + R} \tag{0.14}$$

The number of true positives (TP), the number of false positives (FP) and the number of false negatives (FN) are aggregated over the entire data, and the metrics are calculated based on the overall values.

Table 1. List of the ninety synthesized audio clips

Scene	Salient Sound	Deviant Sound
Rain Forest (20%) Rain Forest (60%)	Turkey	Animals
	Hawk	
	Small dog	
Ocean (20%) Ocean (60%)	Owl	Breakage of objects
	Speech	
	Alarm	
Street (20%) Street (60%)	Bells	Car horns
	Siren	
	Helicopter Pass	

3.4.2. Results and Analysis

Attributed to the different components of sound instances, the sound database is divided into four groups: urban scene sounds, nature scene sounds strong background scene and weak background scene. We choose MDSM to make a comparison for it shows better robustness and efficiency in environment sound saliency detection tasks then the conventional auditory saliency detection models. To some extent, the sound deviancy detection could be seen as saliency detection. Table 2, Figure 13 and Figure 14 show the results obtained in different sound sets of the proposed model and model (multi-domain saliency map, MDSM) presented in (J. Wang et al., 2015). In particular, Table 2 reports the precision, recall and F-score of each sound group of two models.

Comparing MDSM and proposed models, Table 2 clearly show that the proposed model achieves the best result in each sound group. The main idea of MDSM is to detect the most salient components in a sound clip. The deviant salient-sounds may be the most salient sound in sound snippets, however, as we illustrated in Section 3 that the deviant sound may appears more than once and should have the same attributes in both temporal domain and frequency domain. Hence, only detect the most salient sound is not sufficient. Since the proposed model applied entropy-based deviancy method, it could detect all the deviant salient-sounds in the most sound scene and shows a much better performance than sound saliency detection models.

Table 2. Sound deviancy detection performance of MDSM and proposed model

Model	MDSM			Proposed Model		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
Urban sound	79.31%	44.23%	56.79%	89.74%	88.98%	89.36%
Nature sound	67.17%	30.32%	41.78%	68.01%	62.45%	65.11%
Strong background	35.00%	10.29%	15.91%	49.89%	52.42%	51.12%
Weak background	72.14%	41.39%	52.60%	79.37%	69.30%	73.99%
Overall	61.00%	27.23%	37.65%	65.27%	62.00%	63.60%

In Figure 13 and Figure 14, the F-score of deviancy detection result in different sound groups derived from rhe two models are respectively shown. From Figure 13 we can see that the proposed model achieves the best performance than MSDM in each sound group. The proposed model shows excellent performance of deviancy detection in urban sound group which the F-score is 89.36% while the MDSM only achieve 56.79% which could be seen as no accuracy. This result further demonstrates the effectiveness of entropy-based deviancy detection method in improving the performance in different kinds of background noise. The performance of nature sound group in this experiment is notably worse than urban sound. There are several possible explanations for this. Firstly, the salient sound component and deviant sound component in nature sound are all unstable and transient sounds. Therefore, they all presented as local salient sound on GFCC curve. Secondly, as they are all intermittent appeared in the nature scene, the both of them show a high uncertainty value in entropy domain. Nevertheless, the result shows that the proposed model perform better than MDSM in this sound scene.

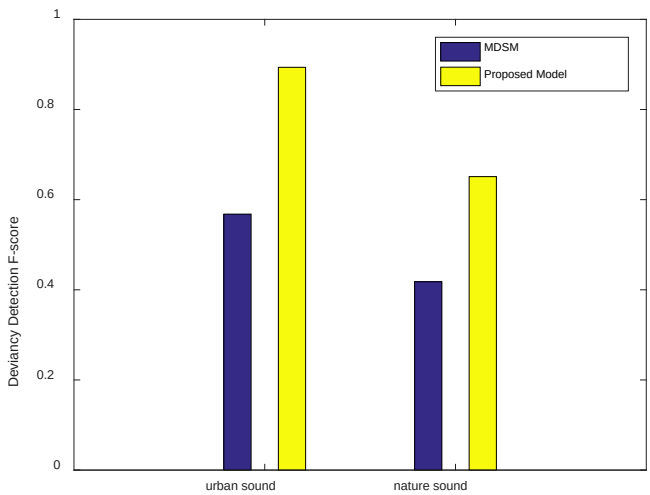


Figure 13. F-score of urban scene sound deviancy detection result and nature scene sound deviancy detection result of MDSM and proposed model

Figure 14 illustrates the deviancy detection result in strong and weak background noise scene and over all database. It is obviously that the proposed model is much

robustness and efficient than the MDSM model. Since the loudness of scene is high in strong background noise scene, the deviant salient-sounds almost masked by the background noise. This increases the difficulty of deviancy detection in such situation since the deviant salient-sounds are hard to be perceived by human acoustic. Despite this, the proposed model could still achieve a considerable result (F-score=51.12%) than the MDSM model (F-score=15.19%).

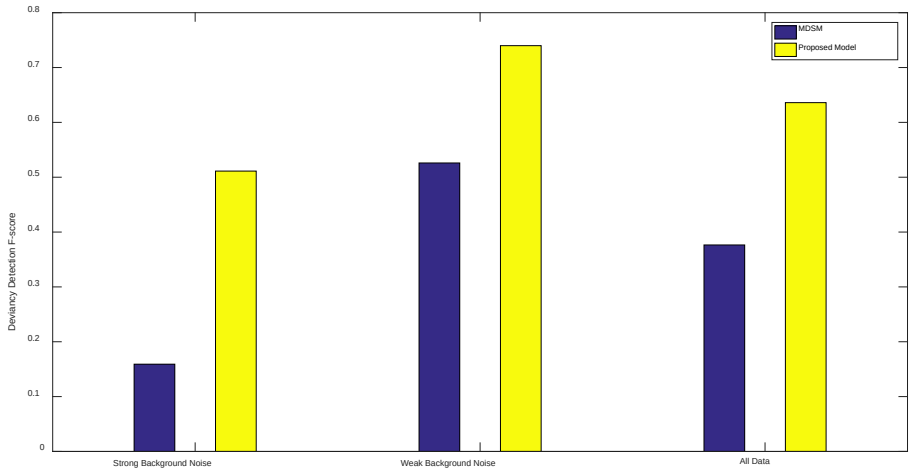


Figure 14. F-score of sound deviancy detection results under strong and weak background noise with all data deviancy detection result derived from MDSM and proposed model

3.5. Conclusion

To make artificial intelligence could have a better performance of percept the complex environment, a computational model which mimic the human auditory cognitive characteristics of auditory attention shift is proposed in this chapter. It is mainly consisting of three modules. The first module is a novel approach for detecting the temporal deviancy based on the GFCC time domain curve to detect the local saliency of a sound signal. Meanwhile, a wavelet entropy and a sample entropy-based deviancy detection method are proposed. Then, the temporal domain deviancy is

presented by the points which both possess saliency and deviancy. Second, in order to accurately detect sounds saliency, a module focus on the frequency domain saliency detection method based on the sound PSD to extract the saliency of sound in frequency domains is presented. Finally, an image indicator based on opponent color space is presented to give a better presentation of the deviant salient-sounds of sound signals. Two experiments were performed to verify the accuracy of the proposed model. The verification of the proposed model shows the details of the deviancy detection process as well as the detection results of three representative sound snippets. From the results, it can be concluded that the GFCC is a robust representation of environment sound while the entropy method is an efficient way for sound deviancy detection. The experiment further demonstrates the performance of the proposed model in a more quantitative manner and illustrates that the proposed model could mimic human auditory attention effectively.

Generally, the first step of artificial cognition (salient and deviant sounds detection) of environmental sound for machines is possible to be realized by applying the presented approach. Furthermore, the obtained result could be used as the input for the next auditory processing step, sound classification.

Chapter 4. Analysis of Multiple Aggregated Acoustic Features for Environment Sound Classification

4.1. Introduction

Environmental sound classification (ESC) is a staple component of environment auditory cognition. Although appropriate frameworks for automatic speech recognition (ASR) and music information retrieval (MIR) have been well established by multiple researchers, such as (Juang & Rabiner, 2005; Klapuri & Davy, 2007; H. Xu et al., 2018; Yakar et al., 2013), etc., the ESC research is still at the early stage. (Piczak, 2015b) has pointed out that environmental sounds are very diverse group of everyday audio events on account of considerably non-stationary characteristics that cannot be described as only speech or music. Therefore, the algorithms originally established for ASR and MIR may not be sufficient for ESC. Furthermore, the environment sounds do not have meaningful patterns or sub-structures, such as rhythm for music and phonemes for speech. Meanwhile, it is nearly impossible to identify sound mixtures from a waveform. Hence, the main idea of ESC is first applying feature extractions to map the input sound waveforms into feature space, and then using the eigenvectors to train a classifier for categorizing of environmental sounds. The frequency domain, spectrograms (time-frequency domain representations) and cepstral domain have been used in ESC for years. However, these features generally fail to precisely describe the content of environment sounds, since they cannot comprehensively represent the information in environment mixtures individually, leading to the classification accuracy of ESC failing to reach the same level as visual events categorization technologies. Hence, researchers have strived to maximize the information content with combination schemes of the three domains features in the past decades

Acoustic features developed for speech and music analysis are based on psychoacoustic properties of auditory signals such as pitch, loudness and timbre which are easy to be computed and applied generally along with other features.(Chachada &

Kuo, 2014). (Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, & Lian-Hong Cai, 2002) proposed the Octave-based Spectral Contrast features for music recognition. The experimental results indicated that these features are more efficient than MFCC for music signal classification tasks. (Xing et al., 2017) aggregated Chroma, Mel Spectrogram, MFCC, Spectral Contrast, Tonnetz and Tempogram to compose a hyper-images for CNN based music recognition. (Ghosal & Kolekar, 2018) combined MFCC with delta and double delta coefficients, Mel Spectrogram with first and second order derivation, Chroma, Constant Q Chroma, Short Time Fourier Transform, Tonnetz and Tempogram for CNN-LSTM based music recognition. This framework outperforms on the GTZAN dataset. (Zhao et al., 2012) presented a detailed demonstration and analysis of the advantages and disadvantages of MFCC and GFCC, respectively. (Burgos, 2014) combined MFCC and GFCC, and then, principal component analysis (PCA) was performed to reduce the feature dimensions. The aggregated features performed better than single features in the ASR system.

Even though the content of environmental sounds is more diverse than speech and music signals, the features established for ASR and MIR are still widely used in ESC due to their considerable performance. Several single feature-based or hybrid feature-based approaches can be found in literatures. (Piczak, 2015a) first proposed a CNN with Log-Mel spectrogram for ESC. The spectrograms are split into segments of 41 frames and combined with their deltas as a 2-channel input of the CNN. Two similar CNN-based frameworks that use Log-Mel spectrogram can be found in (Salamon & Bello, 2017) and (Takahashi et al., 2016). (H. Zhang et al., 2015) proposed a novel spectrogram image feature (SIF) for CNN based ESC system. They firstly extracted the spectrogram from a sound signal. Then, the spectrogram is smoothed in frequency, and the down-sample and de-noised of the new spectrogram are performed. At last, the time-domain energy was computed for each frame, while the maximum energy indices with the six frames around each of them were used to form the SIF. (Boddapati et al., 2017) extracted spectrogram, MFCC and Cross Recurrence Plot (CRP) from sound signals and aggregated them in to a single channel color image. Two CNN-based models, AlexNet and GoogleNet, were applied to verify the performance of this feature on the

ESC-50 and UrbanSound8K datasets. The mixed Log-Mel and Gammatone spectrograms are used in (Z. Zhang et al., 2018) with a 8-layer CNN for environment sounds classification. (S. Li et al., 2018) proposed a stacked CNN for ESC where one uses Log-Mel spectrograms with their first order derivation as input and the second CNN uses raw waveforms.

These research works indicate that, environmental sounds are different from human speeches or music signals, and the classification performance of environmental sound depends on the selection of audio features to a great extent. The conventional sound event analysis mainly addresses time-frequency features or cepstral domain features only, where some needed information is neglected. While combined acoustic features can contain more information than features extracted from a single domain. However, grounded on the fact that sometimes aggregate features from different domain may reduce classification accuracy, the feature combination strategies should be carefully designed. In view of the features developed for ASR and MIR that are well studied and have a certain effect in ESC tasks, the combinations of these features might perform well in representing the environmental sounds, while this assumption still needs further validation. Therefore, in this chapter, the performances of such acoustic feature aggregated strategies for environment sound taxonomy are ascertained.

4.2. Overview of the Approach

The ascertain work presented in this chapter consists of three general processing units, which are acoustic feature extraction unit, feature combination unit and performance of each feature sets in an environmental sound classification analysis unit.

In addition to the appropriate features, a satisfied classifier is an essential component for ESC as well. Recent research shows that deep learning models are more effective than ordinary classifiers, such as the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Support Vector Machines (SVM) (Dai, 2016). Convolutional neural networks (CNNs) is one of the outstanding structures of deep neural networks. Therefore, CNN is also exploited as classifiers for aggregated feature

analyze. The overall processing method of the investigation is graphically illustrated in Figure 15, where the composition of each step is detailed presented.

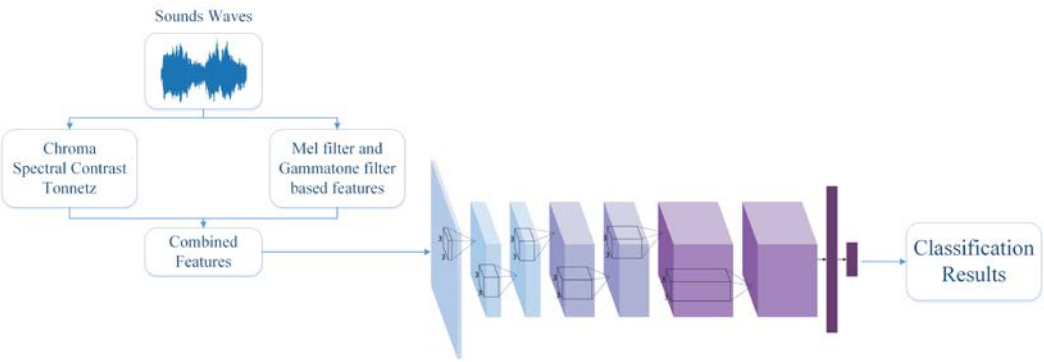


Figure 15. The framework of environmental sound classification.

From Figure 15, it can be seen that through feature extraction algorithms, two kinds of acoustic features are extracted. The first category contains Chroma, Spectra Contrast and Tonnetz, which are originally developed for music signal recognition. These three features will be used as an entirety (called CST) in the rest of the thesis, since the dimension of each of them is very small and the performance of exploit them separately for ESC is extremely bad. The second class includes MFCC, Mel Spectrogram, Log-mel Spectrogram, GFCC and Gammatone Spectrogram, which are generally applied for speech recognition. Thereafter, each feature belongs to the second feature category is combined with CST, individually.

4.3. Feature Aggregation Schemes and CNN model

The detailed introduction of each basic feature applied in our work and the aggregate schemes are presented at first. Then, the 6-layer CNN architecture with its parameter settings is introduced. Finally, the two datasets used for evaluating the performance of these features for ESC are illustrated.

4.3.1. Features

4.3.1.1. General frequency features

1. Chroma (Ewert, 2011): Chroma features are widely used in music analyze and recognition tasks (Bartsch & Wakefield, 2005; Müller, 2007). It is referred to as pitch class profiles and present to be very robust to variations in timbre and closely correlate to the musical aspect of harmony. Meanwhile, multiple results derived from research works related to music identification (Serra, Gómez, Herrera, & Serra, 2008) and audio matching (Müller, Kurth, & Clausen, 2005) indicate that chroma is a powerful mid-level feature representation in content-based audio retrieval. The details about chroma features are described as follows: assuming that the equal-tempered scale, the chromas correspond to the set $\{C, C^{\#}, D, ..., B\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. Then, a twelve-dimensional vector $x = (x_1, x_2, ..., x_{12})^T$ is presented to represent the chroma feature, where x_1 correspond to chroma C, x_2 correspond to chroma $C^{\#}$, and so on. For feature extraction, a sound waveform is converted into a sequence of chroma features, and each sequence explains how the short-time energy of the signal is spread over the twelve chroma bands.
2. Tonal centroid features (tonnetz) (Harte, Sandler, & Gasser, 2006): Tonnetz, also known as harmonic network is a representation of pitch which is first proposed by Euler (Cohn, 1998). The tonal centroid vector t_n of time frame n is the result of multiplication of the chroma vector c_n and a transformation matrix T. Then, the t_n is divided by the L_1 norm of chroma vector to prevent numerical instability, and make sure that the tonal centroid vector dimension is always six. The tonal centroid vector is given as:

$$t_n(d) = \frac{1}{\|c_n\|_1} \sum_{i=0}^{11} T(d,l) c_n(l) \qquad \begin{matrix} 0 \le d \le 5 \\ 0 \le l \le 11 \end{matrix} \tag{0.15}$$

3. Spectral Contrast: The Spectral Contrast feature represents the strength of spectral peaks, valleys and their differences. The same extraction method presented in (Dan-Ning Jiang et al., 2002) is applied to extract spectral contrast features. The sound waves are first segmented into frames of 200ms with overlapping of 100ms. Then, FFT is performed to acquire the spectrum. Afterwards, the Octave-scale filters is applied to divide the frequency into sub-bands followed by estimating the strength of spectral peaks, valleys and their differences. At last, after the estimation results are translated into the Log domain, Karhunen-loeve transform is used to map the raw spectral contrast feature to an orthogonal space and eliminate the relativity among different dimensions.

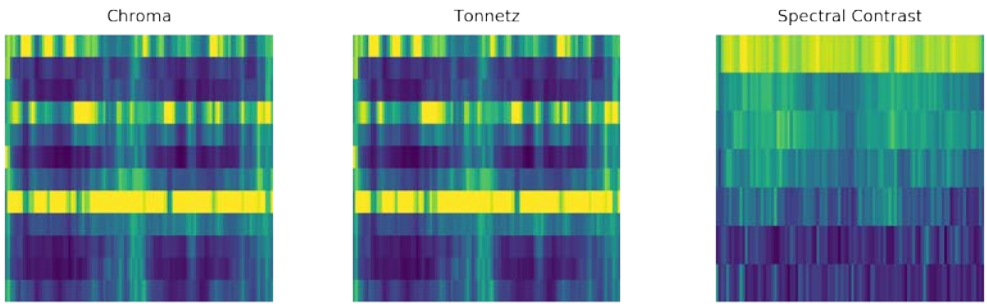


Figure 16. The Spectrograms of Chroma Tonnetz and Spectral Contrast.

4.3.1.2.Mel filter and Gammatone filter based features

The mel filterbank mimics the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstral. This characteristic makes the acoustic feature extracted based on such filterbank could be a better representation of sound. The MFCC generation process includes:

a). Signal Pre-processing,

b). Fourier transform is performed to obtain the signal spectrogram,

c). Mapping of the spectrogram into mel-spectrogram through the triangular overlapping windows which center frequencies are distributed on the mel scale (Serizel et al., 2018):

$$mel(f) = \frac{1000}{\log 2}\log(1+\frac{f}{1000}) \tag{0.16}$$

d). Taking a log calculation on the mel-spectrogram,

e). Applying DCT to the mel log power spectrogram to generate the cepstral features:

$$MFCC = \sqrt{\frac{2}{M}}\sum_{m=1}^{M} X_m(i)\cos\left(\frac{c\pi(m-\frac{1}{2})}{M_m}\right) \tag{0.17}$$

Where X_m is the log energy in m^{th} log mel spectrogram, c is the index of the cepstral coefficient.

Mel and Log-Mel Spectrograms: the same parameters for MFCC processing are used to compute the Mel and Log-Mel Spectrograms. The Mel Spectrogram is the result of Step c of the MFCC computation. The Log-Mel Spectrogram is the Mel log power spectrogram before the DCT step during the computation of MFCC.

The processing methods of GFCC can be found in Section 3.3.1. Gammatone Spectrogram s the time-frequency representation of sound signals derived from the process of GFCC step 2. The Log-Mel, Mel and Gammatone Spectrograms are shown in Figure 17.

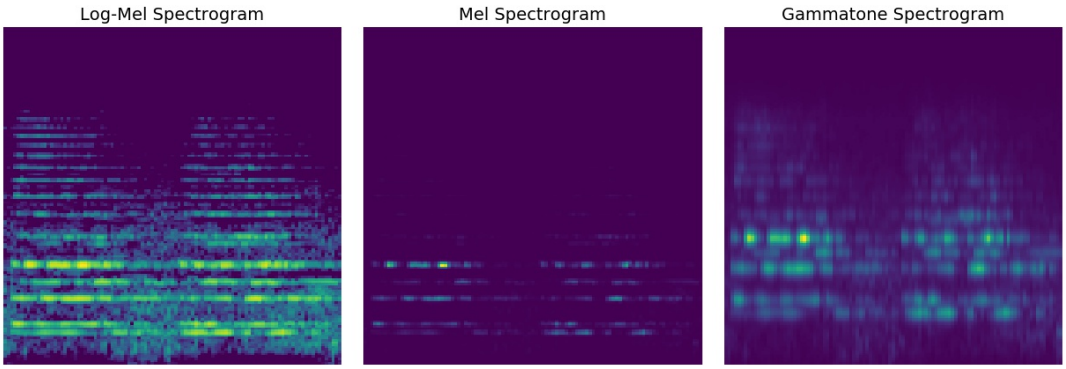


Figure 17. Log-Mel Spectrogram, Mel Spectrogram and Gammatone Spectrogram.

4.3.2. Acoustic features aggregation schemes

According to the pre-settings of Librosa, the dimension of Chroma, Spectral Contrast and Tonnetz is $7\times n$, $6\times n$ and $12\times n$, separately. Meanwhile, lower eigenvectors could not adequately characterize the environmental sounds for the neural networks-based classification tasks. Therefore, these features can be combined as an integrated feature set at first, called CST. Then, the CST is aggregated with the other features described above in a linear way, and all of the combined eigenvectors are 2-D feature vectors. Eight combination strategies for acoustic features are proposed:

1. LM-C: Log-Mel Spectrogram and CST
2. M-C: MFCC and CST
3. Mel-C: Mel Spectrogram and CST
4. M-Mel-C: MFCC, Mel Spectrogram and CST
5. M-LM-C: MFCC, Log-Mel Spectrogram and CST
6. G-C: GFCC and CST
7. GS-C: Gammatone Spectrogram and CST
8. G-GS-C: GFCC, Gammatone Spectrogram and CST

The same feature extraction method presented in (Piczak, 2015a) is performed in this work. All sound clips are converted to the monophonic wave files with 22050 Hz, and then divided into 41 frames with an overlap of 50% (each frame approximately 23 ms). The gammatone filterbank based features are extracted based on the method proposed by (Slaney, 1994). Twenty-dimensional MFCC and GFCC with their first and second order derivatives are extracted, resulting in 60-dimensional vectors for both cepstral coefficient features. The channels of Mel Spectrogram, Log-Mel Spectrogram and Gammatone Spectrogram computation are respectively set to 60. Then all the spectrograms are represented as a 41×60 matrix (corresponding to time and frequency). Meanwhile, the dimension of Chroma, Spectral Contrast and Tonnetz is $7\times n$, $6\times n$ and $12\times n$, separately. Hence, the dimension of CST is 41×25. The combination strategy of the proposed eight feature sets are linear. To be specific, each individual feature in the aggregate features is concatenated individually. Therefore, the

feature size of LMC, MC, MelC, GC and GSC is 41×85. It should be noticed that, for three acoustic features combination, the first and second-order derivations of cepstral coefficient features (MFCC and GFCC) are not used. Hence, the feature size of MMelC, MLMC and GGSC is 41×105. Image representations of each combined features are shown in Figure 18.

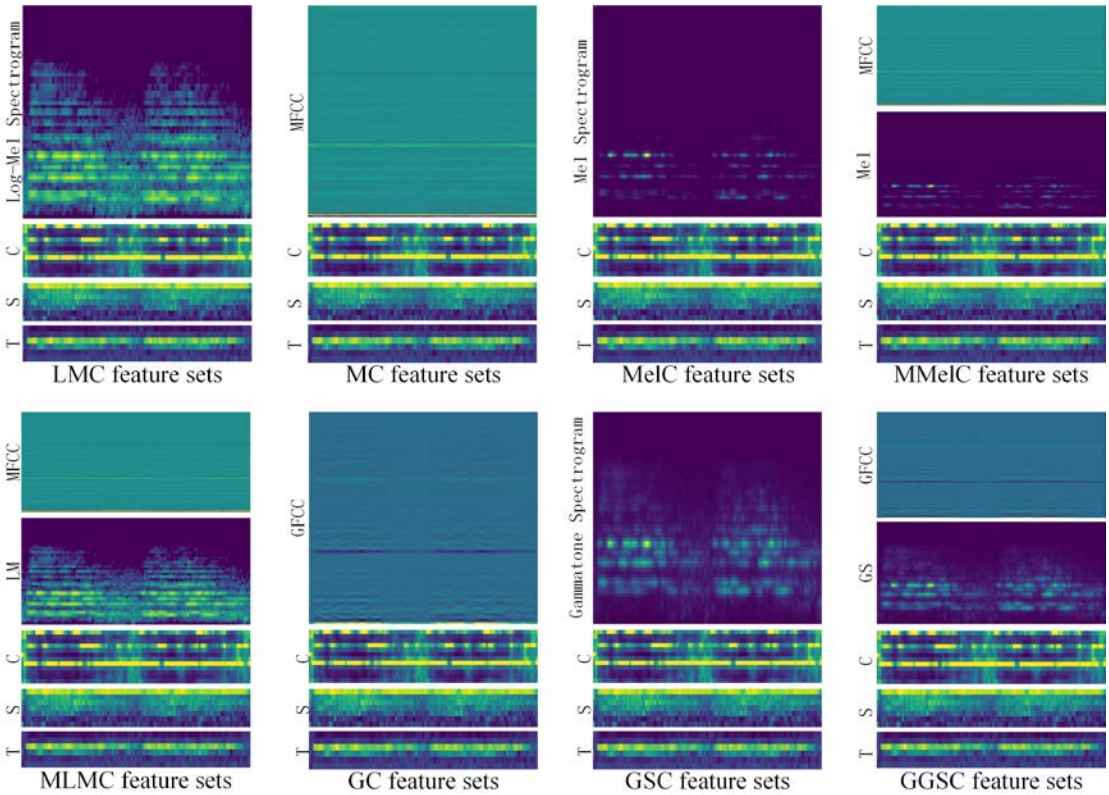


Figure 18. The image representations of eight aggregated acoustic features.

4.3.3. CNN

The convolutional neural network is one of the most famous architecture for deep learning (Gu et al., 2018). It is a type of machine learning algorithm in which can learn how to perform classification tasks with images, video, text, or sounds (LeCun, Bengio, & Hinton, 2015). CNN is a feedforward neural network with convolutional computation and deep structure (Goodfellow, Bengio, & Courville, 2016). This architecture mimics the visual perception mechanism of biological organisms for

supervised learning and unsupervised learning. The convolution kernel parameter sharing and the sparseness of the inter-layer connection in the hidden layer enable the CNN to learn the grid-like topology features (such as pixels and audio waveforms) with a small amount of computation. In addition, pooling processing and the use of many layers are the rest two foundations of CNN that take advantage of the characteristics of natural signals.

4.3.3.1. CNN components

A CNN is consisted of a series layers including: input layer, hidden layer and output layer. The input layer can process multidimensional data like: 1D for signals and sequences, 2D for images and 3D for videos. Figure 19 present how does the feature learned by LeNet-5 (LeCun et al., 1989) convolutional neural network.

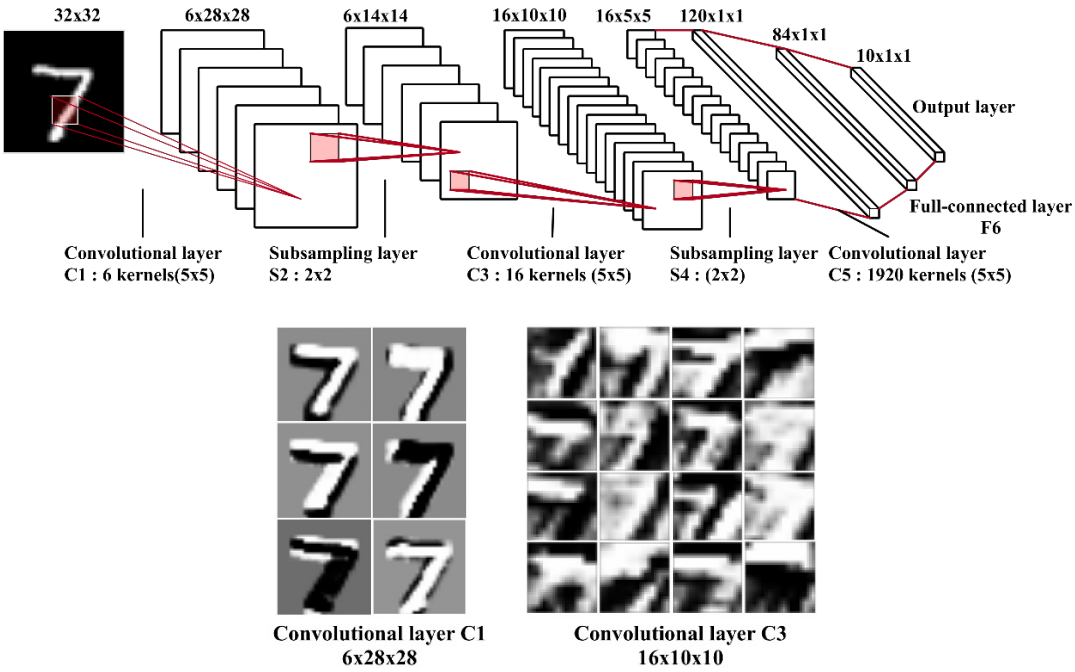


Figure 19. The architecture of the LeNet-5 CNN model works with digit classification task and the visualization of features in the model (Gu et al., 2018).

The function of the convolutional layer is to extract features from the input data, which contains multiple convolution kernels. All element that consists the convolution

kernel corresponds to a weight coefficient and a bias vector, similar to a neuron of a feedforward neural network (neuron). Neurons in a convolutional layer are organized in feature maps, each neuron of a feature map is connected to a region of neighboring neurons in the previous layer. These neighboring neurons are referred to as the neuron's receptive field in the previous layer. The new feature map can be obtained by first convolving the input with a learned kernel and then applying an element-wise nonlinear activation function on the convolved results. The complete feature maps are obtained by using multiple kernels. The mathematical formula which expresses convolution process is defined as:

$$s(i,j)=\sum_{k=1}^{n}(X_k * W_k)_{i,j}+b_k \tag{0.18}$$

Where n is the number of input feature maps from former layer, X_k is the input patch centered at location (i,j), W_k and b_k are the weight and bias vector of k-th filter. $s(i,j)$ is the value of the corresponding position element of the output matrix corresponding to the convolution kernel W_k.

In order to make CNN, which is a multi-layer neural network can have a better understanding of nonlinear features, the activation function has been applied in CNN. ReLU (Nair & Hinton, 2010) is generally used in CNN which is defined as:

$$f(x)=\max(0,x) \tag{0.19}$$

Pooling processing aims at compressing each feature maps to realize feature selection and information filtering. For example, if the pooling stride is 2×2, then, every 2×2 elements in one feature map will be turned to be one element for consisting a new feature map as the input of next convolutional layer. Average pooling (T. Wang, Wu, Coates, & Ng, 2012) and max pooling (Boureau, Ponce, & LeCun, 2010) are the typical pooling operations. The kernels in the 1st convolutional layer are used to detect low-level features, while the kernels in higher layers are learned to encode more abstract features. With several convolutional and pooling layers, higher-level feature representations could be extracted.

In a deep neural network, as the feature is transmitted step by step within the hidden layer, its mean and standard deviation will change, resulting in a covariate shift phenomenon which is an important reason for the vanishing gradient in neural networks (Ioffe & Szegedy, 2015). Performing batch-normalization is a feasible way to solve this problem. The strategy is to first normalize the features in the hidden layer, then use two linear parameters to amplify the normalized features as new inputs, and the neural network updates its parameters during the learning process.

After convolutional and pooling layers, one or more fully-connected layers which aim to perform high-level reasoning is used. They take all neurons in the previous layer and connect them to every single neuron of the current layer to generate global semantic information. Finally, the last layer of CNN is an output layer. For categorization tasks, softmax operator is widely used as output layer.

In general, the main goal of a CNN model is to find the globally optimum parameters for a specific task, which can be achieved by minimizing an appropriate loss function defined on the task. Stochastic gradient descent (Bottou, 2010) and Adam (Kingma & Ba, 2014) are common solutions for optimizing CNN.

4.3.3.2. Proposed CNN architecture

Based on the basic components of CNN which has been described above, a 6-layer CNN model for ESC tasks is established. As shown in Figure 20, the CNN consists of six convolutional layers and a fully connected layer with softmax as the output layer. Every two layers can be treated as a convolutional block since they use the same parameters. Their difference is the max-pooling and dropout, which are performed on the second convolutional layer in one convolutional block. The architecture and parameters of the neural network are as follows:

1. The first layer uses 32 kernels with a receptive field of 3×3 and stride of 2×2 and batch-normalization is applied. The activation function is Rectified Linear Units (ReLU).
2. The second layer uses the same parameters and activation function as the first layer. Then, batch-normalization is applied followed by a max-pooling layer with the

pool stride of 2×2 to reduce the dimensions of the convolutional feature maps.

3. The third layer uses 64 kernels with a receptive field of 3×3 and stride of 2×2 with batch-normalization. ReLU is applied as the activation function.
4. The fourth layer uses the same parameters and activation function as the third layer. Thus, batch-normalization and the 2×2 max-pooling processing are applied.
5. The fifth layer uses 128 kernels with a receptive field of 3×3 and stride of 2×2 with batch-normalization and consideration of ReLU as activation function.
6. The sixth layer uses the same parameters and activation function as the fifth layer, and the batch-normalization and 2×2 max-pooling processing is applied on the output of this layer.
7. The seventh layer is a fully connected layer with 1024 hidden units, and the activation function is Sigmoid. The output is 10 or 50 units according to the datasets, followed by the softmax activation function.

The CNN is trained using a variant of stochastic gradient descent, Adam (Kingma & Ba, 2014). The batch size is set to 32, while all weight parameters are subjected to L_2 regularization and learning rate are set to 0.001 with momentum of 0.9. At the training and test stages, the dropout rate is set to 0.5 and 1, respectively. Cross-entropy is performed as the loss function, which is generally applied for multi-classification tasks.

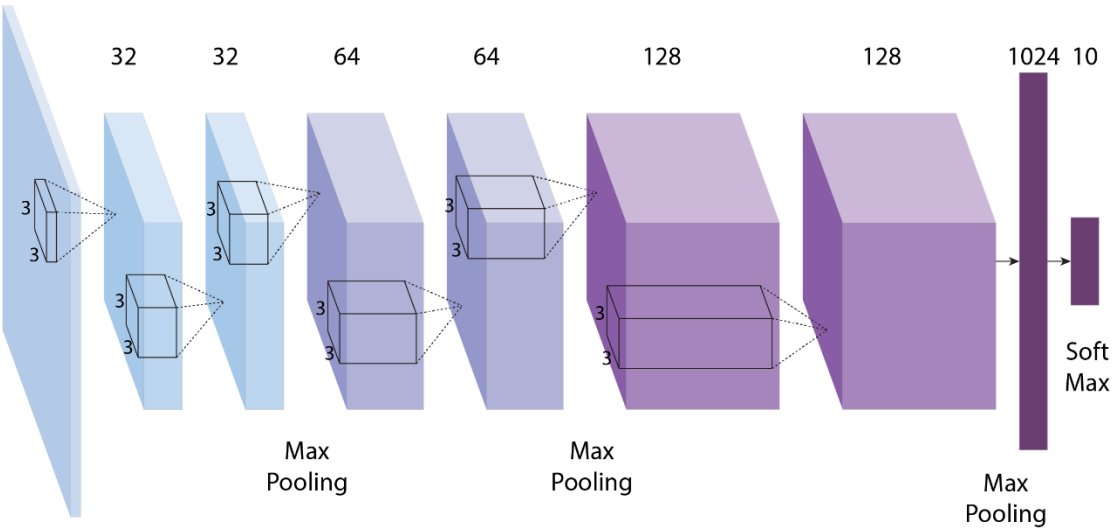


Figure 20. 6-layer CNN architecture.

4.3.4. Database

The UrbanSound8K (Salamon, Jacoby, & Bello, 2014) dataset includes 8732 labelled urban sounds (the length is less than or equal to 4 seconds) collected from the real-world, totalling 9.7 hours. The dataset is separated into 10 audio event classes: air conditioner (ac), car horn (ch), children playing (cp), dog bark (db), drilling (dr), engine idling (ei), gunshot (gs), jackhammer (jh), siren (si) and street music (sm).

The ESC-50 (Piczak, 2015b) dataset contains 2000 environmental recordings (the length is approximately 5 seconds) of 50 equally balanced categories, totalling 2.8 hours. This dataset is divided in to 5 folds. Since ESC-50 contains a large number of audio classes, hence, in the following experiments the number of each class is used to represent each class:

- No.1-10: dog, rooster, pig, cow, frog, cat, hen, insects, sheep, and crow
- No.11-20: rain, sea waves, crackling fire, crickets, chirping birds, water drops, wind, pouring water, toilet flush, and thunderstorm
- No.21-30: crying baby, sneezing, clapping, breathing, coughing, footsteps, laughing, brushing teeth, snoring, and drinking
- No.31-40: door knock, mouse click, keyboard typing, wood creaks (door), can opening, washing machine, vacuum cleaner, clock alarm, clock tick, and glass breaking
- No.41-50: helicopter, chainsaw, siren, car horn, engine, train, church bells, airplane, fireworks, and hand saw

4.4. Experiment and analyze

The features mentioned above can be divided into three categories according to their magnitude: 1) basic feature sets (B-fea), 2) two components aggregated features (2-fea), 3) three components feature combinations (3-fea). The dimension of the three classes features are 41×60, 41×85 and 41×105, separately. B-fea class includes MFCC, Mel Spectrogram, Log-Mel Spectrogram, GFCC, Gammatone Spectrogram

and CST. Since the dimension of CST feature sets are lower than others and the performance of only using CST in ESC tasks is unsatisfactory (which can be seen in Table 4). Therefore, the computational cost of CST will not be illustrated. The 2-fea class includes LMC, MC, MelC, GC and GSC, while the 3-fea class includes MMelC, MLMC and GGSC. Table 3 presents the number of parameters and the memory cost of CNN with the three categorizes features.

Table 3. Parameters and cost of memories for the 6-layer CNN with two size features.

	B-fea		2-fea		3-fea	
Layer	param	memory	param	memory	param	memory
input	0	2.5 K	0	3.5 K	0	4.3 K
Conv 3×3-32	288	78.7 K	288	111.5 K	288	137.7 K
Conv 3×3-32	9.2 K	78.7 K	9.2 K	111.5 K	9.2 K	137.7 K
Conv 3×3-64	18.4 K	40.3 K	18.4 K	57.8 K	18.4 K	71.2 K
Conv 3×3-64	36.8 K	40.3 K	36.8 K	57.8 K	36.8 K	71.2 K
Conv 3×3-128	73.7 K	21.1 K	73.7 K	31 K	73.7 K	38 K
Conv 3×3-128	147.5 K	21.1 K	147.5 K	31 K	147.5 K	38 K
Fc 1024	6.3 M	1024	8.7 M	1024	11 M	1024
Fc 10	10.2 K	10	10.2 K	10	10.2 K	10
Total	6.6 M	281.4 K	8.9 M	401.6 K	11.3 M	495 K

The 10-fold cross-validation and 5-fold cross validation are performed on UrbanSound8K and ESC-50 databases respectively to evaluate the performance of the proposed CNN model firstly. It should be noticed that random time delays, time stretching and pitch shifting are performed on the ESC-50 dataset for data augmentation. Table 4 presents a class-wise accuracy comparison of the six basic features on UrbanSound8K dataset. First, it can be noticed that the features derived from the Mel filter performe better than the Gammatone filter based features. It can be seen that, the performance of MFCC is the best and that the CST is the worst. As mentioned before,

the Librosa library pre-setting of Chroma, Spectral Contrast and Tonnetz leads to a low dimensional representation of sound signals, and thus an unsatisfied taxonomical accuracy for the CST feature set. In addition, Table 4 shows that the gunshot events are the most difficult class to classify. Only MFCC with the proposed framework can obtain an acceptable accuracy, 72.4%, whereas the classification accuracy of other features is less than 60%. However, for MFCC, successive sound (such as children playing, air conditioner, drilling, jackhammer, engine idling, siren and street music) are easier to be classified, and the categorization results of transient sounds (car horn, dog bark and gunshot) are unsatisfactory (accuracy less than 80.0%).

Table 4. UrbanSound8K class-wise accuracy of six basic acoustic features.

Class	MFCC	GFCC	LM	GS	Mel	CST
ac	91.7%	92.7%	93.7%	96.7%	94.1%	69.8%
ch	62.7%	82.1%	60.5%	79.5%	70.6%	37.6%
cp	80.8%	73.0%	79.2%	89.2%	86.5%	59.3%
db	78.2%	68.6%	78.5%	78.1%	85.0%	44.1%
dr	87.7%	83.2%	89.3%	80.6%	75.4%	60.1%
ei	93.6%	93.7%	90.2%	91.6%	94.4%	65.5%
gs	72.4%	52.1%	37.2%	21.1%	26.5%	36.6%
jh	87.0%	91.0%	92.7%	78.9%	77.3%	56.7%
si	84.8%	83.9%	95.8%	95.2%	96.9%	63.8%
sm	89.9%	68.7%	73.2%	75.5%	81.1%	42.8%
Avg.	82.9%	78.9%	79.0%	78.6%	78.8%	53.6%

MFCC achieves the best classification result (82.9%) among these features, which is 4%, 3.9%, 4.3%, 4.1 and 29.3% higher than other features separately. Meanwhile, the classification result of Log-mel spectrogram is 0.4% higher than Gammatone spectrogram and Mel spectrogram is 0.2% higher than GS. These results show that mel filterbank could be a better method than gammatone filter bank in ESC problems.

Furthermore, several research works (Dai, 2016; Juncheng Li, Dai, Metze, Qu, & Das, 2017) point out that the performance of MFCC-based or CNN-based ESC system is considerably lower than their combination for ASR tasks. However, with the proposed 6-layer CNN model, the result of MFCC is 10.2% higher than the accuracy of CNN-based ESC system proposed by (Piczak, 2015a). In addition, except for CST, all the other basic features achieve higher classifications accuracy than the method proposed by (Piczak, 2015a). This indicates that the proposed CNN is an efficient model for ESC tasks. Figure 21 shows the box plot of the comparison of class-wise classification results obtained by each basic feature.

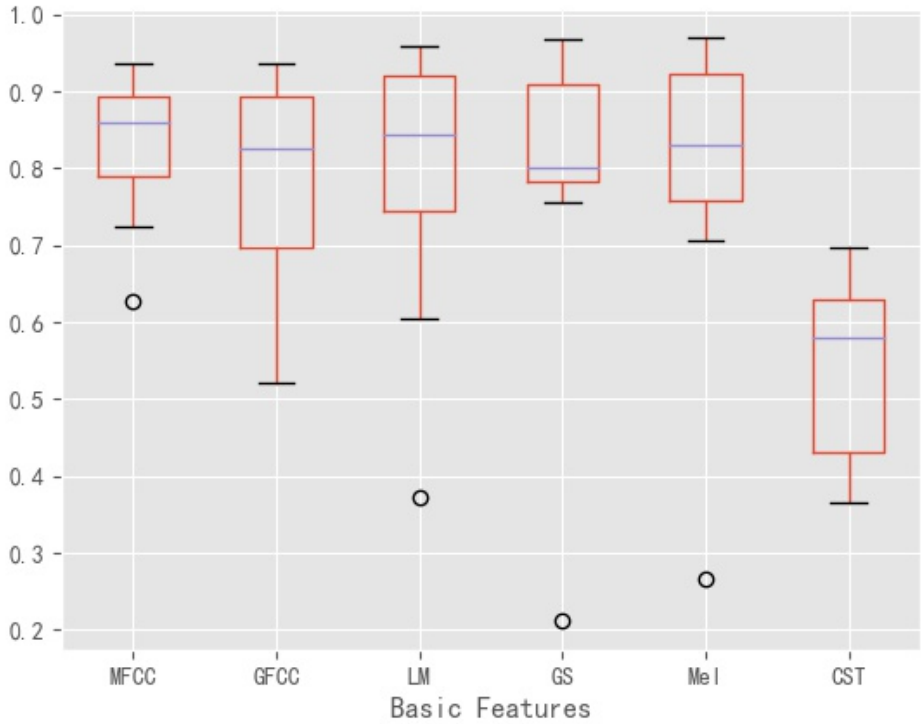


Figure 21 The box plot of the comparison of class-wise classification results obtained by each basic feature.

The class-wise classification results of the eight aggregate features on UrbanSound8K dataset are shown in Table 5. Each filter-based feature has been aggregated with the CST feature, and the cepstral coefficient features with spectrograms (derived from the same filter) are also combined with the CST feature set.

Table 5. UrbanSound8K class-wise accuracy of eight aggregate acoustic features.

Class	LM-C	M-C	Mel-C	M-Mel-C	**M-LM-C**	G-C	GS-C	G-GS-C
ac	96.4%	98.0%	98.8%	97.5%	**97.6%**	97.7%	97.3%	97.9%
ch	87.3%	72.9%	85.1%	87.7%	**90.0%**	65.1%	83.7%	84.7%
cp	94.3%	92.6%	90.6%	93.0%	**95.0%**	85.1%	91.7%	88.6%
db	91.9%	88.0%	90.0%	85.1%	**92.9%**	83.1%	82.5%	85.8%
dr	94.6%	93.2%	87.9%	94.3%	**91.8%**	91.5%	94.4%	98.2%
ei	97.8%	97.8%	96.3%	96.8%	**98.4%**	98.3%	97.3%	97.1%
gs	73.0%	77.1%	74.1%	80.3%	**83.1%**	64.2%	42.6%	41.9%
jh	94.0%	90.7%	86.4%	87.7%	**93.1%**	94.3%	83.8%	86.9%
si	98.9%	96.0%	98.0%	96.7%	**99.0%**	86.3%	93.2%	94.9%
sm	94.8%	88.6%	91.3%	87.3%	**93.4%**	80.1%	87.4%	87.8%
Avg.	92.3%	89.5%	89.8%	90.6%	**93.4%**	84.6%	85.4%	86.4%

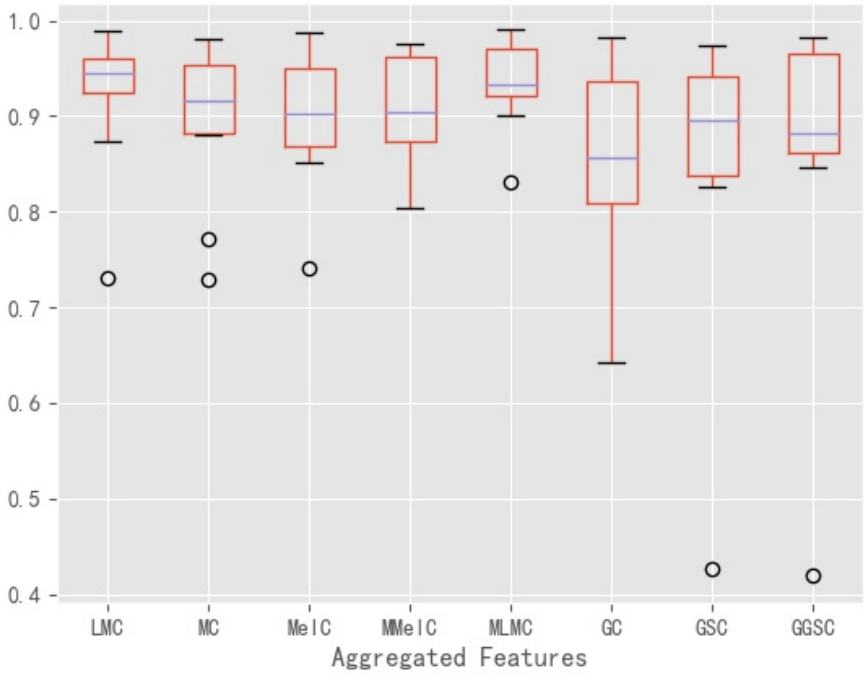


Figure 22. The comparison of class-wise classification results obtained by each aggregated feature set.

It can be demonstrated from Table 5 that with aggregation schemes, all the features have better classification results for ESC tasks than the previous single scheme. It should be noticed that, for Mel filter-based features, the Log-Mel Spectrogram performs better than Mel Spectrogram. The taxonomic accuracy of LM-C is 2.5% higher than that of Mel-C, and M-LM-C is 2.8% higher than M-Mel-C. Furthermore, it is clear that the performance of CST combined with both spectrogram and cepstral feature are better than that of CST combined with only spectrogram or cepstral feature. The M-LM-C is 1.1% higher than LM-C and 3.9% higher than M-C. The G-GS-C is 1.8% higher than G-C and 1.0% higher than GS-C. Moreover, the performance of CST aggregated with features derived from Mel filter is better than the CST combined with Gammatone filter-based features. For the strategies of CST combined with Spectrogram, the classification accuracy of LM-C and Mel-C is 92.3%, and 93.4%, which is 6.9% and 7% higher than the accuracy of GS-C. For the CST aggregated with cepstral features, the taxonomic result of M-C reaches 89.5%, which is 4.9% higher than the 84.6% of G-C. Figure 22 shows the box plot of the comparison of class-wise classification results obtained by each combined feature set.

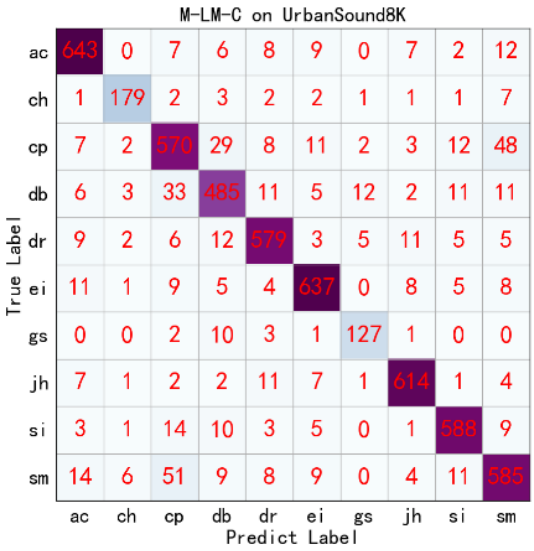


Figure 23. Confusion matrix for the M-LM-C feature with proposed CNN evaluated on the UrbanSound8K dataset.

The highest classification result is achieved by the MFCC-LM-CST (93.4%)

feature combination, and each class has outstanding classification results as well. Except for the classes of gunshot, the classification accuracy of all the other categories are higher than or equal to 90%. However, the categorization of gunshot still achieves an acceptable accuracy (83.1%). The categorization results for four classes (air conditioner, children playing, engine idling and siren) are higher than 95%. Moreover, only M-LM-C reaches 90% on the car horn taxonomy. The confusion matrix of M-LM-C with proposed CNN evaluated on the UrbanSound8K dataset is shown in Figure 23.

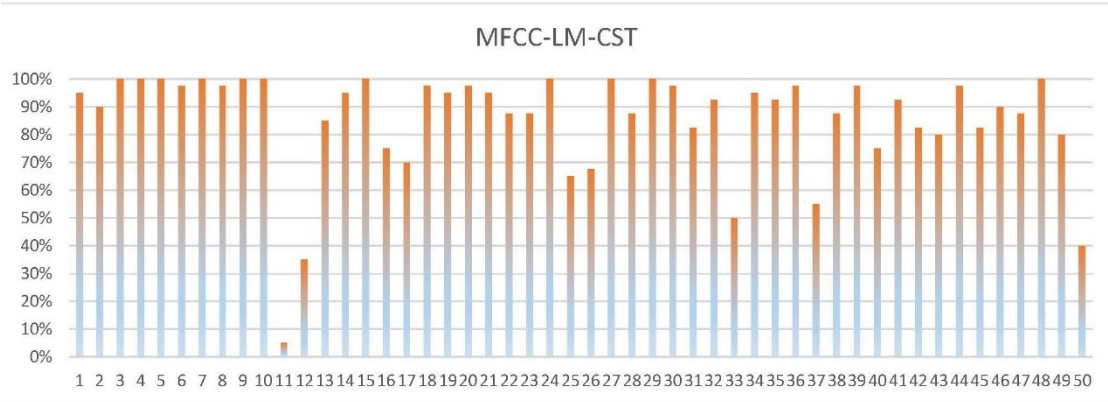


Figure 24. Classification results of M-LM-C on the ESC-50 dataset.

In Figure 24, the detailed taxonomy results obtained by M-LM-C are revealed. They illustrate that M-LM-C with the proposed CNN model can perform well on the ESC-50 dataset. For the M-LM-C feature set, 29 classes achieve a categorization accuracy higher than or equal to 90%, 11 classes reach 100%, and only 5 classes are lower than 60%. In all categories, classes No.11, No.12 and No.37, corresponding to rain, sea waves and vacuum cleaner respectively, have unsatisfactory taxonomic results. The classification of rain has the worst accuracy, only 5.0% for M-LM-C feature. The average classification accuracy for all the 50 classes is 85.6%.

The proposed CNN based ESC framework using the most efficient feature combinations is compared with several existing models proposed by (Aytar, Vondrick, & Torralba, n.d.; S. Li et al., 2018; Piczak, 2015a; Salamon & Bello, 2017; Tokozume & Harada, 2017; X. Zhang, Zou, & Shi, 2017; Z. Zhang et al., 2018; Zhu et al., 2018), the comparison result is presented in Table 6.

With the ESC-50 dataset, the proposed framework can reach 85.6% for M-LM-C feature sets, which is 20.7% higher than the 64.9% of the (Piczak, 2015a) model. Moreover, our taxonomy result is higher than the 83.1% of the (S. Li et al., 2018) model, which has been the state-of-the-art classification result with the ESC-50 dataset in recent years. Furthermore, the proposed algorithm performance is also higher than human recognition accuracy, 81.3%. The confusion matrix of M-LM-C with proposed CNN evaluated on the ESC-50 dataset is shown in Figure 25.

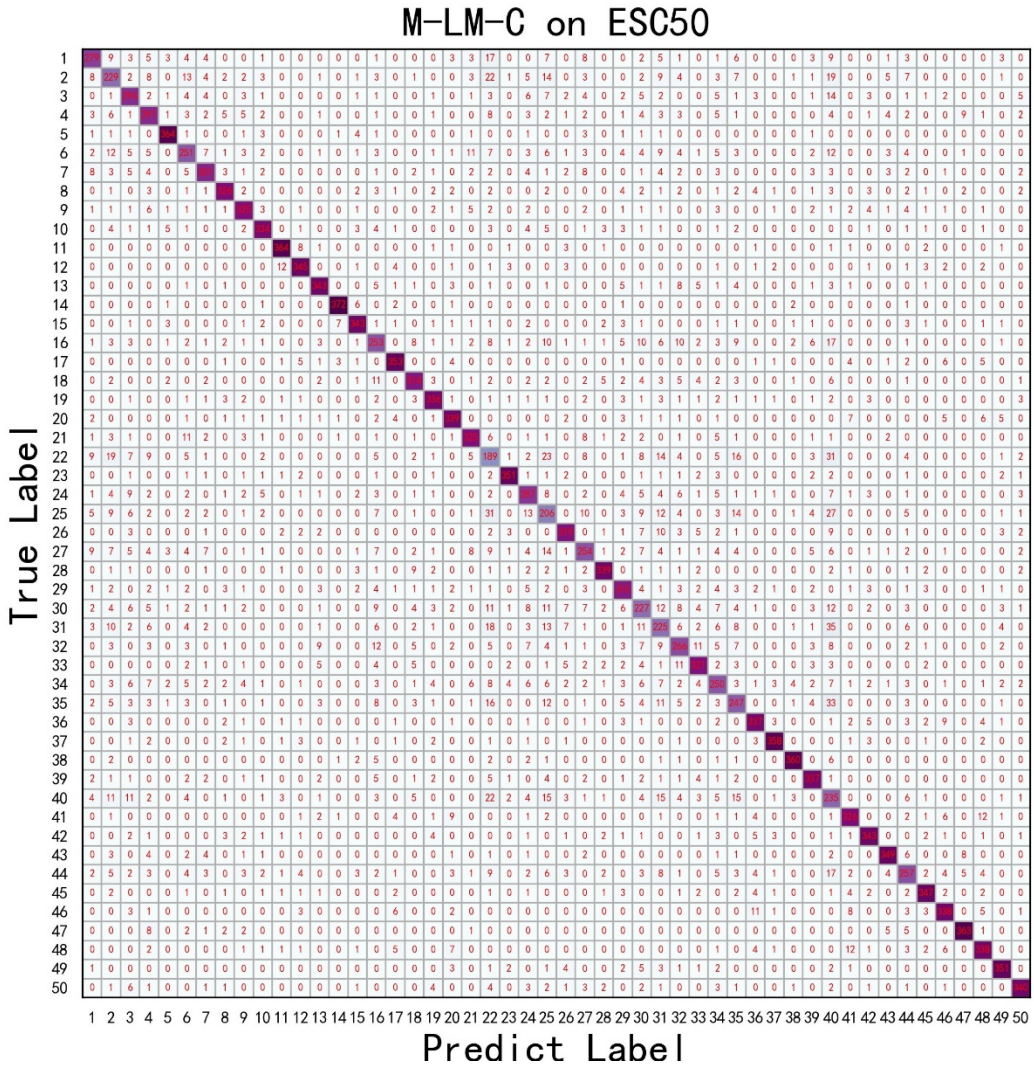


Figure 25. Confusion matrix for the M-LM-C with proposed CNN evaluated on the ESC-50 dataset.

With the UrbanSound8K dataset, the proposed M-LM-C feature sets reached

93.4%, which is 20.7% higher than the (Piczak, 2015a) model. Moreover, the result derived from the proposed method is also higher than the recent works presented in Table 3. These results indicate that, the aggregated features (a combination of features developed for music signals and speech signals) have achieved significant enhancement in environmental sound classification. To our knowledge, the proposed feature combination strategy is currently one of the most efficient manually selected features for environmental sound taxonomy.

Table 6. Comparison of classification accuracy with other models.

Model	Feature	Mean Accuracy	
		ESC-50	UrbanSound8K
(Piczak, 2015a)	LM	64.9%	72.7%
(Salamon & Bello, 2017)	-	-	73.0%
(Tokozume & Harada, 2017)	Raw Data	71.0%	78.3%
(X. Zhang et al., 2017)	Mel	68.1%	81.9%
(Aytar et al., n.d.)	Raw Data	74.2%	-
(Z. Zhang et al., 2018)	LM-GS	83.9%	83.7%
(Zhu et al., 2018)	Raw Data	79.1%	-
(S. Li et al., 2018).	Raw Data-LM	83.1%	92.2%
Our Model With M-LM-C	**MFCC-LM-CST**	**85.6%**	**93.4%**
Human Performance	-	81.3%	-

4.5. Conclusion

In this chapter, the performances of several aggregated features for ESC tasks are evaluated. Since the conventional sound event analysis mainly addresses time-frequency features or cepstral domain features only, and grounded on the fact that sometimes aggregate features from different domain may reduce classification accuracy. Meanwhile, the classification performance of CNN as the classifier is sensitive to the hyperparameters. Minor changes in parameters can lead to a large difference in

classification results. Hence, features that comprehensively represent environment sounds and an appropriate CNN model should be carefully designed for ESC. The efficiency of the 6-layer CNN is evaluated at first, six basic acoustic features (Log-Mel Spectrogram, Mel Spectrogram, MFCC, Gammatone Spectrogram and GFCC) are used as features with the CNN on UrbanSound8K dataset. The results indicated that features such as MFCC which performed unsatisfactorily in other models (Dai, 2016; J. Li et al., 2017) could reach 82.9% with the 6-layer convolutional neural network. These results illustrate that the proposed CNN is sufficient for ESC tasks. Then, eight feature aggregate schemes that combined Chroma, Spectral Contrast and Tonnetz (CST) with the six basic features are presented. The performances of these feature combinations are tested on ESC-50 and UrbanSound8K datasets and the classification accuracy of each class include in these datasets is presented. These results indicate that the feature combination methods and 6-layer CNN can significantly improve the classification accuracy of environmental sounds.

In general, the proposed feature of aggregation strategies can represent more environmental sound information than isolated features. Meanwhile, CNN is proved to be powerful in ESC problems as well. These features could be exploited with more efficient CNN architectures in ESC tasks for achieving higher taxonomic results, and can also provide a better judgment foundation for artificial auditory cognition.

Chapter 5. Modeling Auditory Cognition

5.1. Introduction

Computational modelling the acoustic cognition to make artificial intelligence can percept surrounding environment has long bedeviled researchers. This is because that, the artificial acoustic cognition system is a composite system, which at least consists of three sub-modules: 1) auditory attention module, 2) sound recognizing module, 3) response module. The first component is used to detect the salient or deviant sound events among tremendous stimulus in the environment. Then, the detected signals can be recognized by the second module. At last, based on the recognition result, the third sub-system should judge whether such sound events are needed to attend or not. Each module should give accurate results to ensure that the whole system can simulate human auditory cognition. Compared with those methods which partly mimic the human auditory system, the difficulty of computational modelling auditory cognition system is significantly increased.

Attention is a bi-direct process (Driver, 2001), it is composed of 'bottom-up' stimulus-driven factors and 'top-down' task-specific goals (Kaya & Elhilali, 2017). A number of conceptions have been proposed concerning neural models for understanding auditory attention. Most of these works are closely related to visual theories. In one perspective of view, the auditory attention is regarded as a filtering or selecting mechanism. This concept is directly related to the findings of receptive field characteristics in the cortex, where neurons are viewed as filters (S. Shamma & Fritz, 2014). Another perspective of view is that, the attention is an integration mechanism, where attentional feedback acts as a prior to bias processing of certain stimuli of interest (Kaya & Elhilali, 2017). This conception is also widely accepted in many theories of auditory cognition, in which attention aggregate elements belonging to same sound event.

Based on these theories and conceptions, multiple acoustic attention models have

been proposed in the past decades. (C. Kayser et al., 2005) initially proposed a bottom-up auditory attention model for salient sound detection. Then, two innovative saliency-driven attention models are presented by (Kalinli & Narayanan, 2007) and (Duangudom & Anderson, 2007). These models are built on the tradition in the visual modality, and they neglect some acoustic characteristics, and exploit visual domain-based attention model inherently limits the ability of auditory attention model. Therefore, a considerable number of models have been proposed to address these problems (Kaya & Elhilali, 2012, 2013; Principi et al., 2015; Tsuchida & Cottrell, 2012; J. Wang et al., 2015). However, most of these works are attention models concerning auditory saliency detection, only a few studies realize deviancy detection mechanism for attention models.

For sound recognition module, deep learning-based techniques have been proved to be more efficient than the conventional methods on solving complex classification problems in many domains. Multiple scientists choose deep learning model, such as CNN, in sound classification problems. The advantages of CNN have been illustrated in chapter 4 that CNN can solve the limitations of conventional classifiers in multiple learning and classification problems. However, there is still a long way to go when compared with CNN based image classification algorithms. For example, the longer temporal context information still cannot be captured by original CNN. Hence, many works propose to use merged neural networks to address the above-mentioned shortcomings through integrating information from the earlier steps (Adavanne, Pertilä, et al., 2017; Adavanne & Virtanen, 2017; S. Li et al., 2018; Parascandolo et al., 2017). In these methods, one or more CNNs are used to extract the spatial information with different acoustic features firstly. Then, the outputs are merged by concatenation and feed to recurrent neural network (RNN) layers or another CNN layers for temporal information extraction.

Several research works exploit CNN models which originally developed for image recognition tasks, and achieve outstanding performance in ESC as well. (Boddapati et al., 2017) the environment sound classification accuracy of AlexNet (Iandola et al., 2016) and GoogLeNet (Szegedy et al., 2015) were evaluated on UrbanSound8K, ESC-10 and ESC-50 (Piczak, 2015b) datasets. (Tokozume & Harada, 2017) proposed an

end-to-end ESC system using a convolutional neural network. In this model, raw waveforms were used as inputs and two convolution layers are applied to extract features. A VGGNet (Simonyan & Zisserman, 2014) based ESC system was presented by (Z. Zhang et al., 2018), where the convolution filters were set to 1-D for learning frequency patterns and temporal patterns respectively. (Zhu et al., 2018) proposed a CNN based model called WaveNet, which uses multi-scale features to make CNN learns comprehensively information of environment sounds.

Multiple works apply decision-level fusion in ESC tasks. The main idea of decision level fusion method is to fuse the softmax values acquired from different neural networks through mean calculation, or uncertainty reasoning algorithms such as Dempster-Shafer evidence theory (DS theory) and Bayesian Theory (S. Li et al., 2018; H. Ye et al., 2015). The experiment results indicate that merged neural networks with decision level fusion outperform single deep architectures in taxonomic tasks (Jing Li, Qiu, Wen, Xie, & Wen, 2018; S. Li et al., 2018; Y. Li, Chen, Ye, & Liu, 2016; H. Ye et al., 2015). Although these works have greatly improved the performance of ESC systems. However, from the classification accuracy derived from these recently published works, it is clearly that the CNN-based ESC systems still have great potentials for making further progress.

Recent research works have shown that long-term life experiences affect the ability to hear in background noise (Anderson et al., 2013). To be specific, compared with the unconsciously detected salient sound events, the sounds which have been heard can attract our attention more easily. This result closely parallels theories from ‘top-down’ attention mechanism, which points out that subjective consciousness also has a great influence on attention. For example, listeners can easily attend to one speaker in a multi-speaker environment (O’sullivan et al., 2014), this phenomenon is also known as cocktail party problem. This result pointed out that prior knowledge should be regarded as a crucial component of realizing artificial auditory cognition. Inspired by this, (J. Xu, Shi, Liu, Chen, & Xu, 2018) propose a model about auditory selection with attention and memory, where the top-down task-specific attention and the bottom-up stimulus-driven attention are all realized for speech identification. In addition, decision making

is another key component of cognition system of selecting an action or an event within a series of more alternatives (X.-J. Wang, 2008). (Romo & Salinas, 2001) conducted a study on what components of the neural activity evoked by a stimulus are directly related to decision making, and how are they related. (Heekeren et al., 2008) conducted a review work on conjunction with data analysis methods that can directly link decisions and signals in the human brain on a trial-by-trial basis. Through observing the monkey's responses, (Tsunada et al., 2016) learned the specific and causal contributions of different brain regions in the ventral auditory pathway to auditory decisions.

In the exploration of surrounding environment, artificial intelligence will definitely expose to tremendous sound events while response to all the detected sound events will cost much computational resources. Hence, it is essential to only identify the valuable events and response to them while neglect the rest. Although the above-discussed theoretical works have pointed out that prior knowledge and decision making are essential and crucial components of auditory cognition. However, it is hard to see related researches on establishing such models for environmental sound signals. Hence, novel approach should be researched for better recognizing sounds occurred in the complex environment and make response to it.

5.2. Overview of the Approach

Motivated by the mentioned shortcomings of current approaches and the practical requirement of intelligent environment auditory cognition, an artificial auditory cognition system is proposed. The deviancy detection model presented in chapter 3 is used as the attention module, which is applied to detect the salient or deviant sounds in the environment. To be specific, when there is only one salient sound exist in the environment, this algorithm could be regarded as a saliency detection model. While, if there is more than one sound exist with saliency or an abnormal sound event occurs, it could be applied to detect the deviant sound in the environment. After deviant sound signals are detected, it should be further processed to identify their categorizations.

In order to precisely categorize the detected salient or deviant sound events, a novel four-layer stacked CNN architecture based on two combined auditory features and DS theory-based information fusion method is proposed. The proposed system consists of three steps: sound deviancy detection, sound identification and DS theory-based decision-level fusion. Two combined features (i.e. LMC and MC feature sets) presented in chapter 4 have been used here to train CNN models. The outputs derived from the softmax layer of these two CNNs are fused by DS theory.

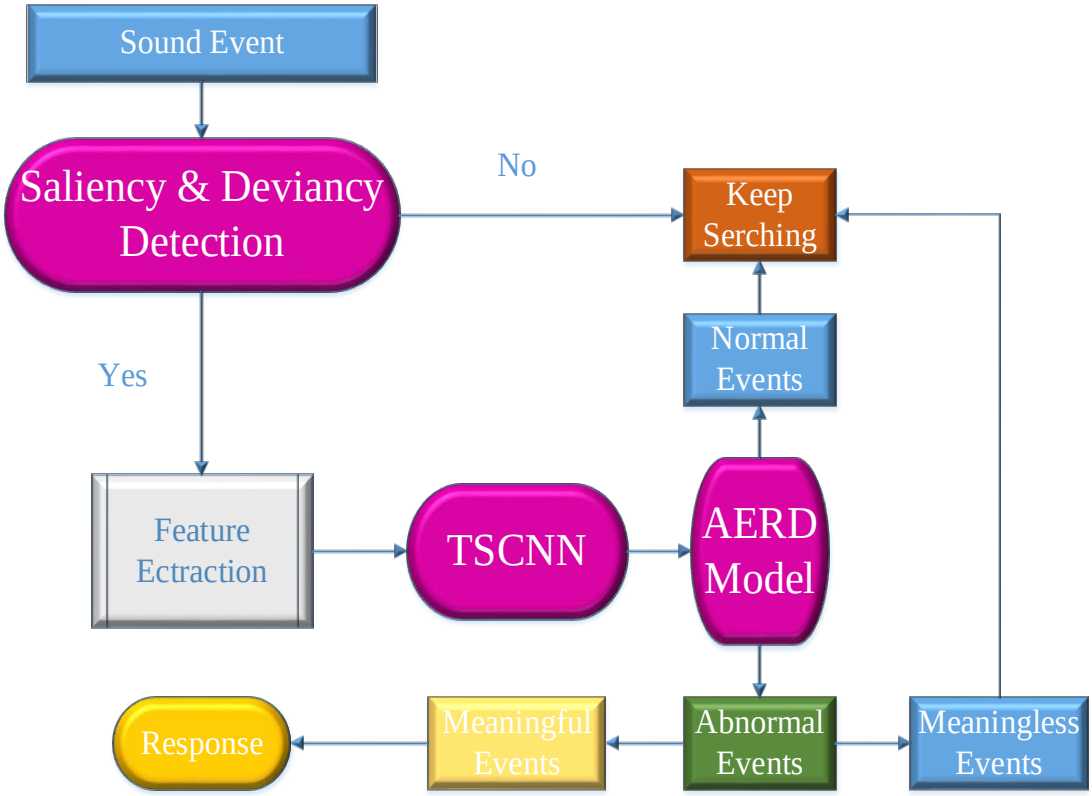


Figure 26. The architecture of proposed artificial acoustic cognition model.

Finally, a knowledge-based system inspired auditory events response decision (AERD) is presented to demonstrate the relationships between sound scenes and occurred sound events. Inspired by the perceptual process of human cognition mechanism, the proposed method is performed by comparing the prior knowledge-based significance of detected salient or deviant sounds with sound scenes information to determine whether the system needs to respond to the abnormal sound events. To be

specific, it is assumed that the surrounding sound scene or environment of the artificial machine is determined, and each normal and abnormal sound event which experimentally would occur in an environment is assigned a level of significance under prior knowledge. Then, the abnormal sounds will be further categorized into meaningful and meaningless events, which means that meaningful deviant sounds need to respond and meaningless events do not need to respond. Finally, the detected sound will be recognized by TSCNN and then, the AERD model will determine whether machines needs to respond to it. The diagram of proposed auditory cognition system is presented in Figure 26.

5.3. DS Evidence based Two-Stream CNN Fusion Method

In this section, the combined features used here are described at first. Then, the structure of 4-layer convolutional neural network model and DS theory-based information fusion algorithm will be presented. Several experiments are conducted to evaluate the efficiency of the proposed ESC module.

5.3.1. Feature aggregation

Selecting a series number of acoustic features to represent the characteristics of environmental sound signals is one of the main obstacles of ESC problems. A comparison of the performance of multiple auditory features in ESC tasks is presented in chapter 4. The experiment results derived from chapter 4 shows that the MLMC feature sets obtained the best classification results among eight aggregated features. Meanwhile, LMC and MC features also present outstanding efficiency in ESC tasks, and the performance is just slightly lower than MLMC. To be specific, MLMC could be regarded as a linear combination of LMC and MC where Chroma, Spectral Contrast and Tonnetz are only used once. Therefore, in order to take advantages of each acoustic feature, LMC and MC are chosen to train two CNN models, separately. Detailed descriptions of feature combination strategies can be found in chapter 4 section 3. The image representation of these two features is shown in Figure 27.

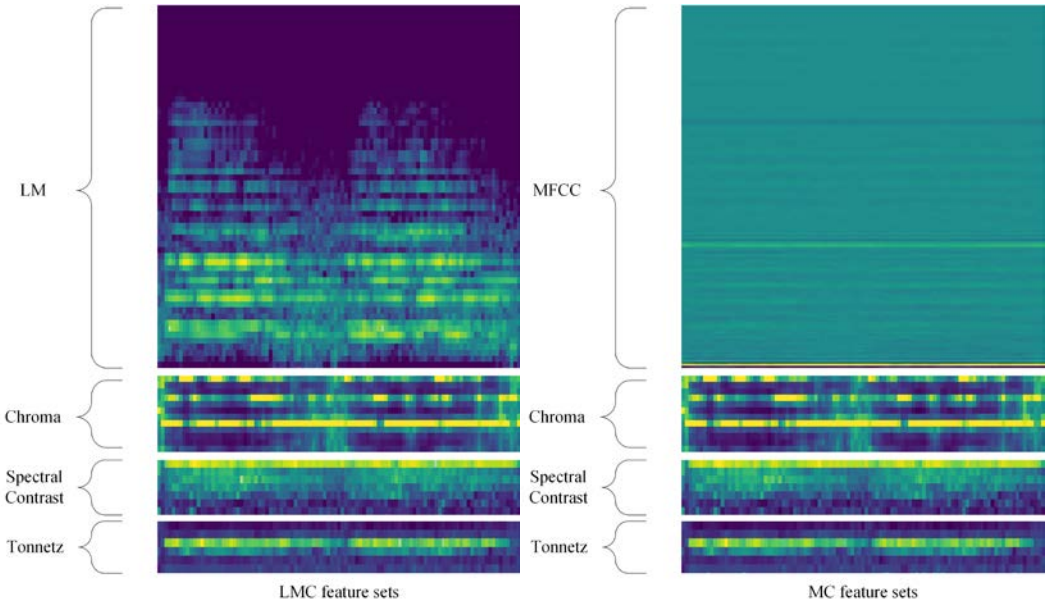


Figure 27. The image representations of LMC and MC feature sets.

5.3.2. Structure of the MCNet and LMCNet

The two networks of TSCNN both contain four convolution layers and one fully connected layer. The framework of the proposed four-layer CNN is shown in Figure 28, the architecture of the model is as follows:

1) The first layer uses 32 kernels with 3×3 receptive field and the stride step is set to 2×2 and batch-normalization is performed. The Rectified Linear Units (ReLU) is exploited as the activation function.
2) The second layer uses the same settings as the first layer, where 32 convolution kernels with receptive filed of 3×3 and stride step of 2×2. The batch-normalization is performed and activation function is ReLU as well. The difference is that the second layer applies max-pooling for dimensionality reduction of feature maps.
3) The third layer uses 64 convolution kernels with a receptive field of 3×3 and the stride step is also 2×2, where batch-normalization is used. Followed by

the activation function, ReLU.

4) The fourth layer 64 convolution kernels with receptive filed of 3×3 and stride step of 2×2. The batch-normalization is performed and activation function is ReLU.
5) The fifth layer is the fully connected layer with 1024 hidden units and the activation function is Sigmoid.
6) The output is 10 units based on the datasets, followed by the softmax activation function.

At the training stage, we use a 0.5 dropout probability for the second layer, fourth layer and the fully connected layer to prevent overfitting. The CNN is trained through a variant of stochastic gradient descent, Adam (Kingma & Ba, 2014). The batch size is set to 32, while all weight parameters are subjected to L_2 regularization and learning rate are set to 0.001 with the momentum of 0.9. The cross-entropy is applied as the loss function. At the testing stage, all parameters are the same as the training stage, while the dropout will not be implemented.

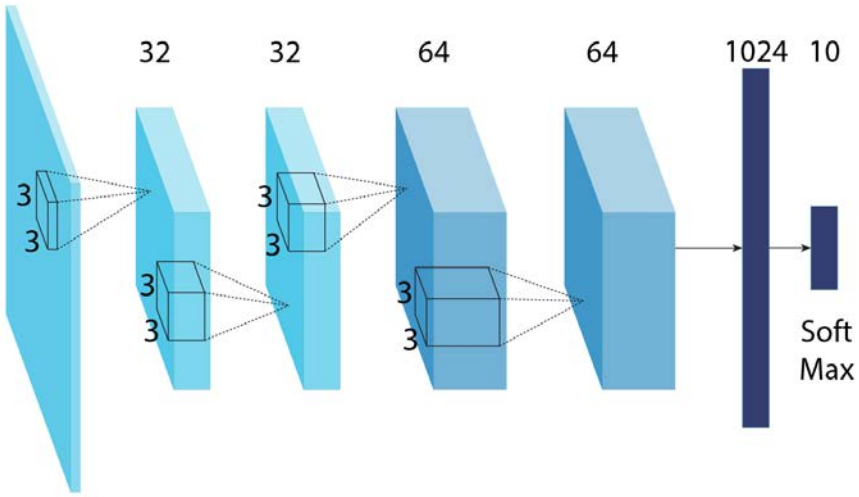


Figure 28. The architecture of proposed 4-layer CNN.

5.3.3. Dempster-Shafer evidence theory-based information fusion

Dempster-Shafer evidence theory (DS theory) is originally established by (Shafer, 1976), it is also known as belief function theory. The DS theory is mainly about quantified beliefs like Bayesian probability. The main idea of this theory is the notion of evidence and how different pieces of evidence should be combined in order to make inferences (Reineking, 2014).

The basis of DS theory is to establish a frame of discernment Θ and a subset of hypothesis $\{\mathrm{A}_1, \mathrm{A}_2 \ldots \mathrm{A}_n\} \subseteq \Theta$, where n is the number of hypothesis. A_i is an element of the power set $P(\Theta)$. Mass function or basic probability assignment M is a mapping: $P(\Theta) \rightarrow [0,1]$ distribute a mass value to each hypothesis $\mathrm{A}_i \subseteq \Theta$. The mass function represents the trust level of each element itself. There are two constraints of mass function:

1) $\sum_{A \subseteq \Theta} M(A) = 1$, which means the sum of each probability in subset A is 1.
2) $M(\varnothing) = 0$, this indicate that the mass function cannot allocate any value to an empty set. Meanwhile, a mass function with this characteristic is called normalized mass function.

In this work, the category of sounds in the dataset can be treated as an element in subset A under the frame of discernment Θ. Here, $n = 10$ according to the classes number of UrbanSound 8K and each element are independent. For solving reasoning problems, the mass function representing different part of evidence must be combined in a meaningful way. Here, we use Dempster's rule to combine the two mass functions derived from each CNN. This combination rule allows combining normalized mass function that are obtained over the same frame of discernment.

The outputs of softmax of LMCNet and MCNet are used as the mass function $M_1(B)$ and $M_2(C)$. The combination of mass function ($M_{1\oplus 2} = M_1 \oplus M_2$) based on Dempster's rule $\oplus$ is defined as:

$$M_{1\oplus 2}(A) = \alpha_D \sum_{B\cap C=\varnothing} M_1(B_i)M_2(C_i), \qquad \forall A \subseteq \Theta, A \neq \varnothing \tag{0.20}$$

$$M_{1\oplus 2}(\varnothing)=0 \tag{0.21}$$

$$\alpha_D = \frac{1}{\sum_{B\cap C=\varnothing} M_1(B_i)M_2(C_i)} \tag{0.22}$$

Where, α_D is a normalization constant indicating the mass function is normalized. $M_{1\oplus 2}(A)$ is a mass function as well and satisfied $\sum_{A\subseteq\Theta} M_{1\oplus 2}(A) = 1$, which is the final probability assignment of $M_1(B)$ and $M_2(C)$, it is also the result of the fusion process of LMCNet and MCNet.

With the LMCNet, MCNet and the DS theory-based information fusion method, we propose the TSCNN. The overall framework of the this ISR system is shown in Figure 29.

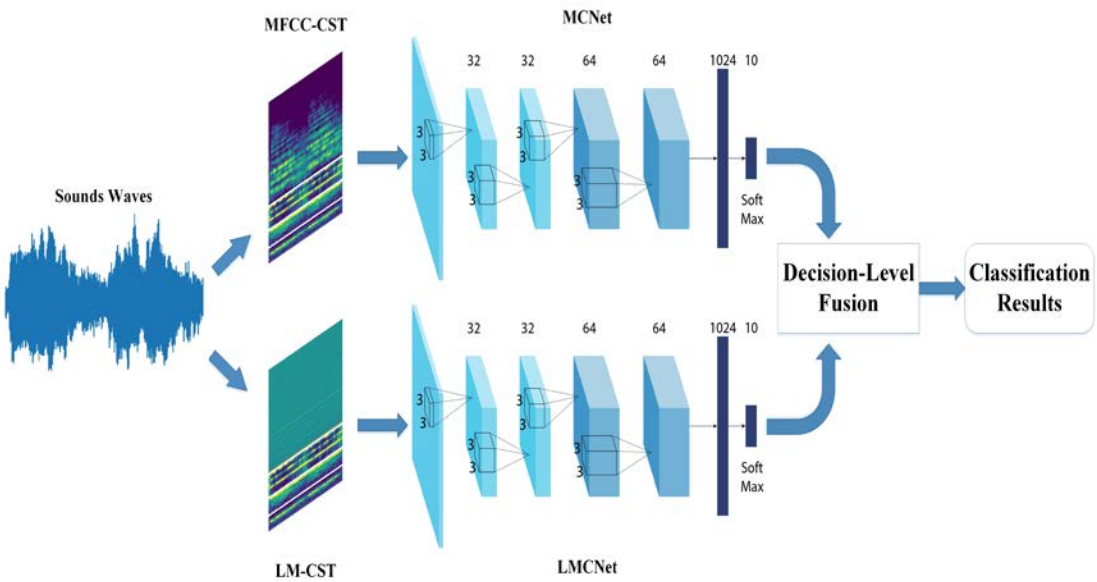


Figure 29. The overall framework of the DS theory based ISR system.

From Figure 29, it can be seen that, MFCC, Log-Mel Spectrogram, Chroma, Spectral Contrast and Tonnetz features are extracted from sound waveforms at first. Then, MFCC and Log-Mel Spectrogram is combined with the rest three features, separately. The MFCC-CST feature set is used to train the MCNet and LM-CST is used

to train the LMCNet. Finally, the softmax value derived from each neural network are fused through DS evidence theory to form the sound classification results.

5.3.4. Experiment

The UrbanSound8K (Salamon et al., 2014) dataset includes 8732 labeled urban sounds (the length is less than or equal to 4 seconds) collected from the real-world, totaling 9.7 hours. The dataset is separated into 10 audio event classes: air conditioner (ac), car horn (ch), children playing (cp), dog bark (db), drilling (dr), engine idling (ei), gunshot (gs), jackhammer (jh), siren (si) and street music (sm). Waveform and spectrogram of each audio class are shown in Figure 30.

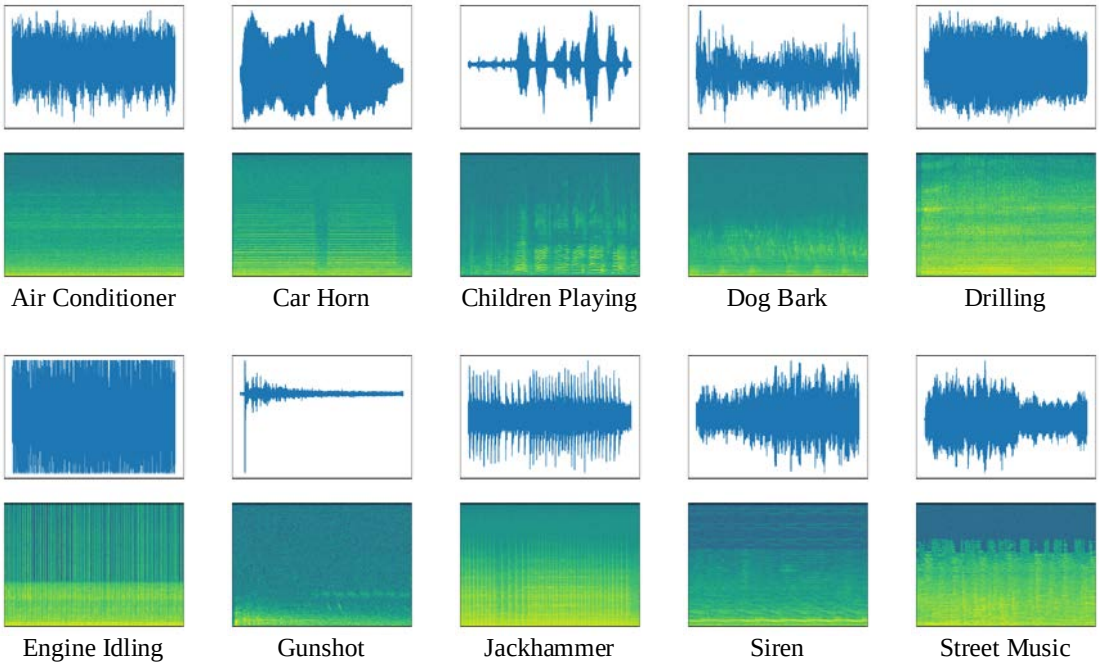


Figure 30. Waveform and Spectrogram of each audio class

5.3.4.1. Experiment Setup

The same feature extraction method presented by (Piczak, 2015a) is used in this work. All sound clips are converted to the single channel wave files with the frequency

of 22050 Hz. Then, they are divided into 41 frames with an overlap of 50% (each frame is about 23 ms). We use the pre-setting channels of Librosa to extract the Chroma, Spectral Contrast and Tonnetz features. For the MFCC extraction, the values of first twenty channels with their first and second order derivatives are used, resulting in 60-dimensional feature vectors. The channels of Log-Mel Spectrogram are set to 60, in order to make the dimension to be equal with MFCC. Then all the spectrograms are represented as a matrix with a size of 41×60. The feature size of chroma, tonnetz and spectral contrast is 41×7, 41×6 and 41×12, separately. Therefore, the size of LMC and MC are all 41×85. Figure 31 shows the graphically representation of how does the feature learned by the proposed 4-layer CNN.

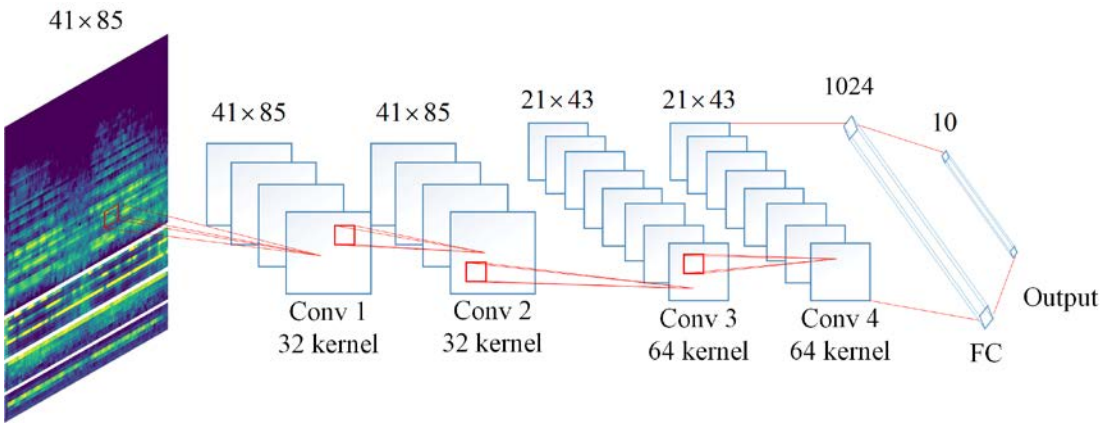


Figure 31. The architecture and size of feature maps in each convolutional layer.

It can be seen from Figure 31 that, the feature maps derived from first and second convolutional layer have the same size as input feature. After 2×2 max pooling processing, the size of input feature maps for third convolutional layer is 21×43. Since the max pooling is not performed after convolutional layer 3, so that the size of input features for 4th convolutional layer is 21×43 as well. Then, features with a size of 11×22 are derived from the last hidden layer and feed to the fully-connected layer which has 1024 hidden units. The output is a 1×10 tensor according to the number of classed of UrbanSound8K dataset is 10.

For each experiment, the 10-fold cross-validation is performed to evaluate the

proposed ISR model on UrbanSound8K dataset. The combined features and 4-layer CNN architecture are two main contributions of this work. Hence, we first analyze the efficiency of the CNN model train with combined features. Meanwhile, the influence of the different number of convolution layers (six and eight) on CNN-based ESC system is also investigated. The additional convolution layers in the CNNs for comparison use the same receptive fields of 3×3 and stride step of 2×2, batch-normalization is performed on each layer with ReLU as the activation function. Dropout with a rate of 0.5 is exploited for the sixth and eighth convolution layer in the two additional CNN models respectively. Table 7 presents the number of parameters and the memory cost of CNN with different number of convolutional layers.

Table 7. Parameters and memory of CNN with different convolution layers.

	4-layer		6-layer		8-layer	
Layer	param	memory	param	memory	param	memory
input	0	3.5 K	0	3.5 K	0	3.5 K
Conv 3×3-32	288	111.5 K	288	111.5 K	288	111.5 K
Conv 3×3-32	9.2 K	111.5 K	9.2 K	111.5 K	9.2 K	111.5 K
Conv 3×3-64	18.4 K	57.8 K	18.4 K	57.8 K	18.4 K	57.8 K
Conv 3×3-64	36.8 K	57.8 K	36.8 K	57.8 K	36.8 K	57.8 K
Conv 3×3-128	0	0	73.7 K	31 K	73.7 K	31 K
Conv 3×3-128	0	0	147.5 K	31 K	147.5 K	31 K
Conv 3×3-256	0	0	0	0	294.9 K	4.6 K
Conv 3×3-256	0	0	0	0	589.8 K	4.6 K
Fc 1024	15.9 M	1024	8.7 M	1024	4.7 M	1024
Fc 10	10.2 K	10	10.2 K	10	10.2 K	10
Total	15.9 M	339.6 K	8.9 M	401.6 K	5.9 M	413.4 K

5.3.4.2.Experiment Result

The classification performance of the feature level fusion method is also considered. Since MLMC can be regarded as the linear combination of LMC and MC features, hence, MLMC is employed as feature-level fusion method to make a further investigation of the influence of various feature combination strategies in 4-layer CNN based ESC system. The detailed combination method and image representation of MLMC is shown in chapter 4 section 4, the feature size of MLMC is 41×145. The class-wise classification accuracy and the average accuracy of 10-fold cross-validation obtained by LMCNet, MCNet and MLMC-CNN contains different number of convolutional layers and the proposed TSCNN model on UrbanSound8K dataset is presented in each table.

Table 8. Class-wise classification accuracy of four models with 4-layer CNN.

Class	LMC (LMCNet)	MC (MCNet)	MLMC	**TSCNN**
ac	98.6%	99.9%	99.2%	**99.9%**
ch	93.9%	91.4%	93.2%	**94.2%**
cp	97.3%	93.9%	96.1%	**97.5%**
db	92.6%	90.4%	94.2%	**95.3%**
dr	94.8%	95.0%	95.7%	**97.2%**
ei	98.9%	99.6%	98.5%	**99.6%**
gs	88.6%	91.1%	85.9%	**95.4%**
jh	93.2%	95.9%	91.1%	**97.1%**
si	98.6%	98.3%	98.5%	**98.9%**
sm	95.0%	97.4%	94.1%	**96.9%**
Avg.	95.2%	95.3%	94.6%	**97.2%**

Table 8 describes the experiment results of each method with 4-layer CNN models. We can find that the feature combination of LMC and MC performs well in the 4-layer CNN based ISR system. Five classes taxonomic accuracy of LMCNet and six classes

taxonomic accuracy of MCNet are higher than 95%. It can be seen that the MLMC which aggregated of all feature sets cannot improve the performance, the taxonomic result derived from MLMC-CNN is 0.6% and 0.7% worse than LMCNet and MCNet. LMCNet and MCNet achieve 95.2% and 95.3%, which is 22.5% and 22.6% higher than the model presented in (Piczak, 2015a), respectively. In addition, although MLMC-CNN has the worst performance among the four models, however, it is still 21.9% higher than the 72.7% of Piczak's model. It can be seen that for both methods, the classification accuracy of all categories is higher than 90% except for gunshot of LMC and MLMC. The proposed TSCNN model reaches 97.2% which is 24.5% higher than Piczak's work, and it significantly improved the classification accuracy of gunshot (95.4%). The box plot of comparison between four models with 4-layer CNN on UrbanSound8K is presented in Figure 32.

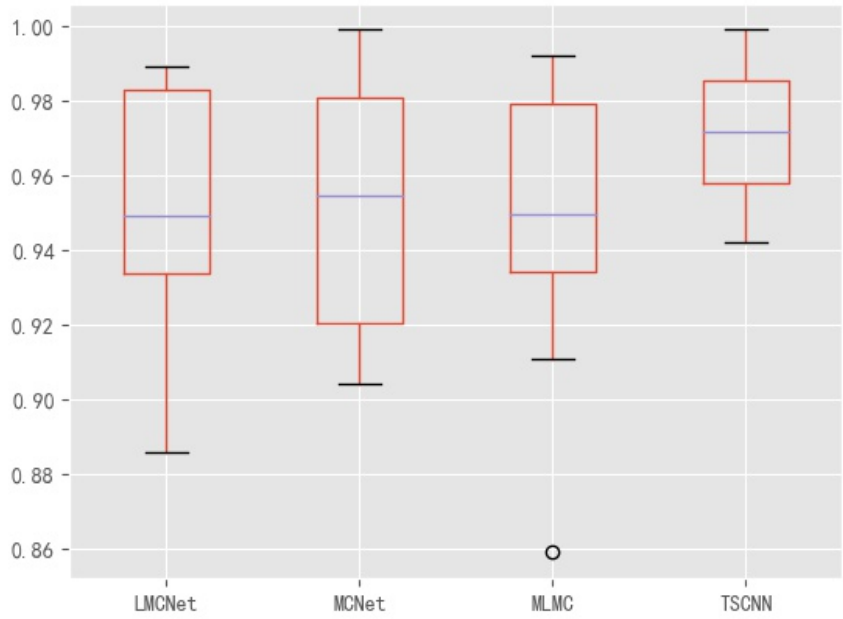


Figure 32. Comparison of four models with 4-layer CNN on UrbanSound8K.

In order to further illustrate whether the proposed TSCNN model outperform LMCNet, MCNet and 4-layer CNN using MLMC feature sets, we show the standard deviation and time cost of each model in Table 9. The classification accuracy obtained by TSCNN is 2% and 1.9% higher than LMCNet and MCNet. It is also shown in Table

3 that the standard deviation of TSCNN is much less than three other methods, which further demonstrate that the fusion model outperforms three other single models. The mean time cost for LMCNet, MCNet, MLMC and TSCNN is 0.023s, 0.024s, 0.028s and 0.077s, separately. It should be pointed out that the time consuming is the single sound classification time in the test stage, and the model loading time is not considered. The test is conducted in Python under Microsoft Windows 10 x64 OS on a computer with Intel Core i7-8700 CPU, two GTX 1080 GPU (the memory of each GPU is 8 GB) and 32 GB RAM. Although the time cost of the proposed model is almost three times longer than single neural networks, the computational cost of TSCNN is still well acceptable for ESC tasks in real time.

Table 9. Statistics analyze and time cost of 4-layer CNN based models

	Mean	N	Std Deviation	Time cost
LMCNet	0.9515	10	0.03121	0.023
MCNet	0.9529	10	0.03352	0.024
MLMC	0.9465	10	0.03812	0.028
TSCNN	0.9720	10	0.01788	0.077

It can be seen in Table 10 that, the 6-layer CNN based models performs slightly worse than the methods use 4-layer CNN. The LMCNet, MCNet, MLMC-CNN and TSCNN is 2.2%, 6.0%, 1.9% and 2.3% worse when compared with the 4-layer CNN based models. The categorization accuracy of gunshot for both methods is less than 90% and it is less than 80% for LMC and MC feature sets. Classification accuracy of dog barking with MCNet failed to reach 90%, and taxonomic accuracy on children playing of MCNet dramatically reduced to 69.4%. The MLMC feature cannot improve the classification performance as well, where the accuracy of children playing and gunshot failed to reach 90%. The same situation also appear on TSCNN model. Nevertheless, the proposed TSCNN model still achieves the best classification result (94.9%) among the four models. The box plot of comparison between four models with 6-layer CNN

on UrbanSound8K is shown in Figure 33.

Table 10. Class-wise classification accuracy of four models based on 6-layer CNN.

Class	LMC (LMCNet)	MC (MCNet)	MLMC	TSCNN
ac	98.9%	98.9%	97.5%	99.9%
ch	90.2%	69.4%	87.9%	89.2%
cp	94.8%	91.1%	93.6%	96.4%
db	91.3%	88.0%	91.6%	93.1%
dr	93.8%	90.9%	91.5%	95.5%
ei	98.2%	97.7%	98.1%	99.1%
gs	77.2%	77.2%	81.7%	85.1%
jh	92.6%	91.6%	93.4%	97.1%
si	99.0%	96.1%	99.0%	98.9%
sm	94.3%	92.1%	92.9%	94.7%
Avg.	93.0%	89.3%	92.7%	94.9%

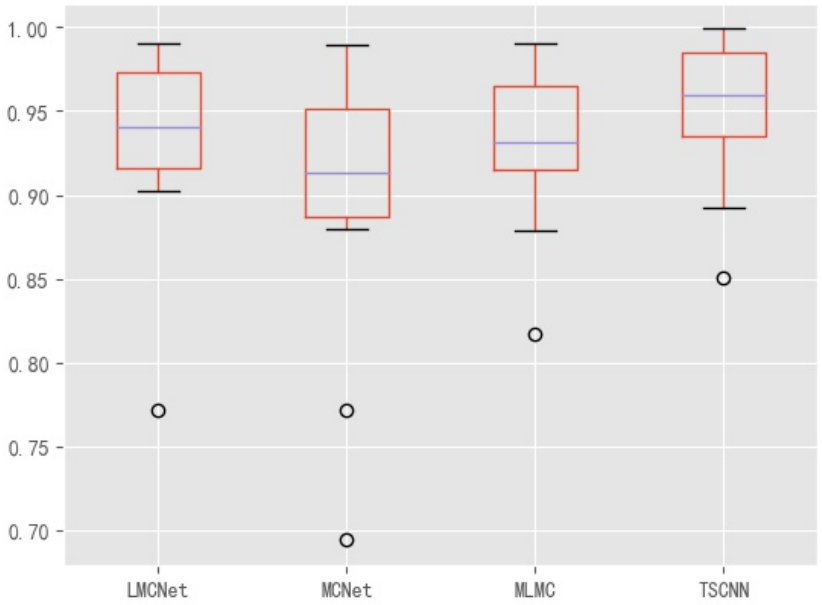


Figure 33. Comparison of four models with 6-layer CNN on UrbanSound8K.

From Table 11 we can find that the performance of all methods is unsatisfactory

with the 8-layer CNN. Most of the categories and all methods obtain a taxonomic result that less than 90%. This indicates that using deeper layers may not give a better result for deep architectures, while appropriate layers and suitable parameter settings are the most important components of deep learning architectures. The box plot of comparison between four models with 8-layer CNN on UrbanSound8K is shown in Figure 34.

Table 11. Class-wise classification accuracy of four models based on 8-layer CNN.

Class	LMC (LMCNet)	MC (MCNet)	MLMC	TSCNN
ac	94.8%	91.5%	93.2%	98.2%
ch	76.1%	47.3%	88.1%	69.9%
cp	84.0%	80.9%	87.9%	88.0%
db	79.9%	73.3%	86.8%	80.8%
dr	87.8%	87.4%	87.0%	91.6%
ei	96.8%	94.8%	95.3%	97.4%
gs	57.2%	63.4%	45.4%	67.8%
jh	89.8%	74.7%	85.9%	87.6%
si	97.8%	88.3%	96.5%	96.3%
sm	85.3%	71.8%	90.3%	80.3%
Avg.	84.9%	77.3%	85.7 %	85.8%

In general, we can find out that the applied LMC and MC features present to be efficiency with the proposed ESC system, which clarifies the advantage of the proposed feature combination strategies in ESC tasks. The TSCNN model outperforming other models for both CNN architectures with different convolution layers. Then, the four-layer CNN achieves the best taxonomic accuracy when compared with other CNN architectures. Meanwhile, the taxonomic accuracy of both methods with the proposed 4-layer CNN are higher than existing models. These results demonstrate the efficiency of the proposed 4-layer CNN and DS theory fusion method based TSCNN model.

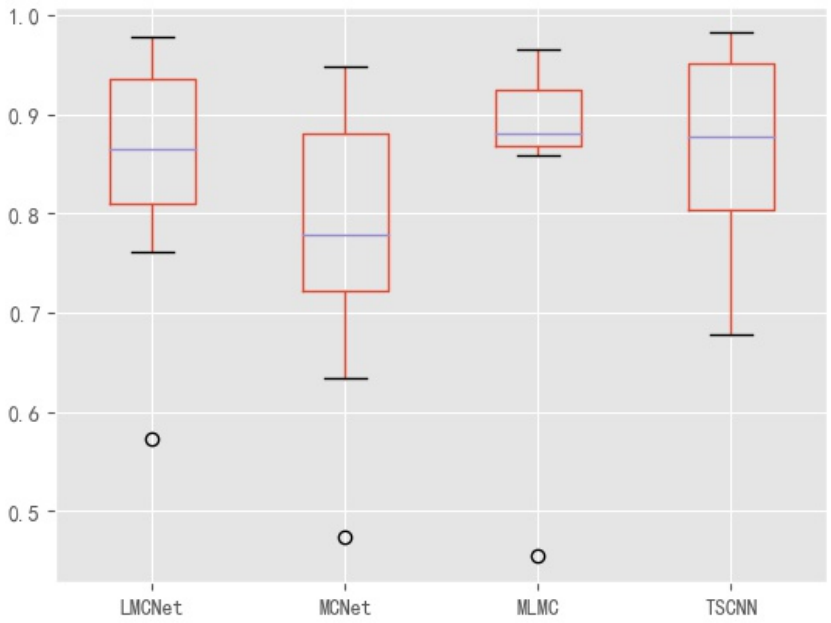


Figure 34. Comparison of four models with 8-layer CNN on UrbanSound8K.

In order to make a comprehensively comparison, we also investigate the two-stream CNN with layer stack method. This model combined the outputs of the second convolution layer of both CNN and the concatenate feature maps are than used as inputs for the next convolution layers. We test this stacked CNN with 4, 6 and 8 layers as well. The parameter settings of each convolution layers and fully connected layers are equal to the 4-, 6- and 8-layer CNN described above. The classification accuracy of these stacked CNNs on UrbanSound8K dataset are shown in Table 12.

Table 12. The ESC results of stacked CNNs with 4, 6 and 8 convolution layers.

Model	Accuracy
Stacked 4-layer CNN	86.4%
Stacked 6-layer CNN	79.8%
Stacked 8-layer CNN	80.1%

It is clearly that the stacked 4-layer CNN models achieve the highest (86.4%) classification accuracy among the three models. Which is 6.6% and 6.3% higher than stacked six- and eight-layer CNN respectively. This result further proves that the proper

number of layers and parameters is the key to the deep learning model based ISR system, where the advantage of the proposed 4-layer CNN is further proved as well.

At last, we compare our TSCNN model with several existing CNN based ISR models as presented by (Piczak, 2015a), (Tokozume & Harada, 2017), (X. Zhang et al., 2017), (Z. Zhang et al., 2018), Li(S. Li et al., 2018) and (Boddapati et al., 2017). The results are shown in Table 13.

Table 13. Comparison of performance with other models on UrbanSound8K datasets.

Model	Feature	Accuracy
Piczak **(Piczak, 2015a)**	LM	72.7%
Tokozume**(Tokozume & Harada, 2017)**	Raw Data	78.3%
Zhang X**(X. Zhang et al., 2017)**	Mel	81.9%
Zhang Z**(Z. Zhang et al., 2018)**	LM-GS	83.7%
Li**(S. Li et al., 2018)**.	Raw Data-LM	92.2%
Boddapati**(Boddapati et al., 2017)**	Spec -MFCC-CRP	93%
LMCNet	LM-C	95.2%
MCNet	M-C	95.3%
TSCNN	**MC & LMC**	**97.2%**

The LMCNet use LMC feature sets achieve an accuracy of 95.2%, which is 22.5% higher than the (Piczak, 2015a) model use LM features. Meanwhile, it is 11.5% higher than the (X. Zhang et al., 2017) model use LM and Gammatone Spectrogram combined feature. Furthermore, the performance of LMCNet is slightly higher (3%) than the model presented by (S. Li et al., 2018), which also applies DS theory as fusion method to fuse two CNN models. The classification accuracy of MCNet is 95.3%, which is much higher than the 72.7% of the model proposed by (Piczak, 2015a), and is 2.3% higher than the (Boddapati et al., 2017) model which also use MFCC based aggregated features. Finally, the proposed DS theory based TSCNN model obtains the highest

taxonomic accuracy (97.2%) among all the ESC models. The performance of our algorithm is much higher than the (Piczak, 2015a) model and is also 5% higher than the (S. Li et al., 2018) model which uses same fusion strategy. As far as I know, this is the first time that the categorization accuracy has reached over 95% on UrbanSound8K dataset and is the highest accuracy compared with existing models.

5.4. Knowledge based System for Auditory Cognition

5.4.1. Auditory Perception

Auditory perception is the ability to understand the information contained in the sounds. A sound begins as a physical vibration in the atmosphere which propagates to the ear. Then, the sound will be transduced into neural stimuli followed by analyzation, categorization and selection into events with meaningful characteristics. Constant interaction exists among top-down attention, bottom-up attention and perception. The selection and filtering operation take lots of constantly variable event and compress them into relative number of events according to categorizes.

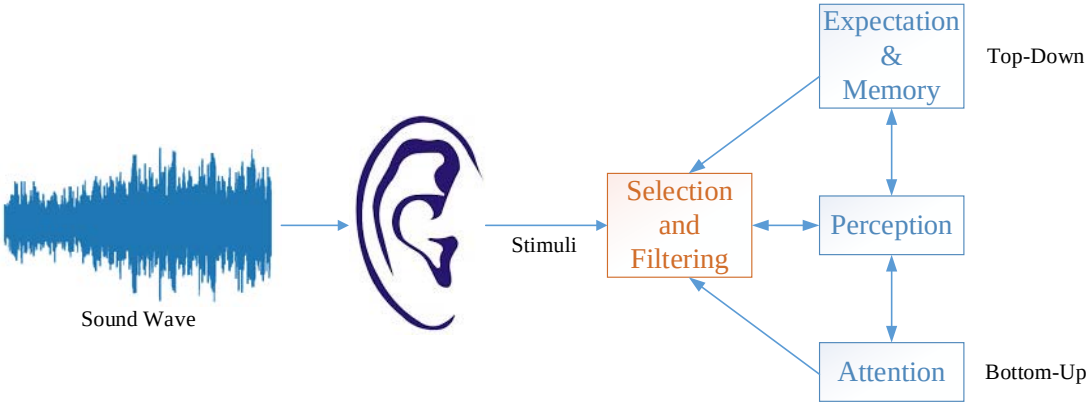


Figure 35. The brief illustration of sound processing in auditory system.

Expectations and memories can influence the formation of sound patterns, and they continuously interact with selection and filtering process, which will affect the class of sound events forward to perception. For example, people are more likely to perceive the voice of their acquaintances in a crowd. This is because that the top-down

and bottom-up mechanism to the selection and filtering operation is unidirectional processing since neither the conscious nor the memory access sound waveform directly.

Bottom-up perception is the mechanism to detect targets and target-triggered attentional processing by the salience or deviancy of the targets, and their ability to trigger attention through exploiting cortical areas in a bottom-up pathway (Sarter, Givens, & Bruno, 2001). However, previous work has not supported a direct role for salience detection regions and processes in the enhancement of memory for salient stimuli (Santangelo & Macaluso, 2013). A fMRI study conducted by (Wills et al., 2016) pointed out that salience detection regions in the human brain have not been activated during the encoding of contingently salient stimuli. While activation in frontoparietal regions has been found which thought to enhance task representations, trigger cognitive control and task goals to prioritize information in memory. This result explained that the bottom-up manner could be regarded as the enhancement to top-down attention.

Top-down attention underlies our ability to concentrate on relevant stimuli and neglect irrelevant conspicuous events. The widely accepted opinion is that top-down, or goal-directed attention is undeniably important in volitionally selecting stimuli that match current task demands (Awh et al., 2006). Top-down modulation of sensory processing is not an intrinsic property of sensory cortices, but rather relies on long-range inputs from and interactions with a network of ‘control’ regions in our brain (Gazzaley & Nobre, 2012). To be specific, the life experience and memories can influence auditory cognition processing directly. Hence, both of these attention pathways should be considered in modeling artificial auditory cognition system.

5.4.2. Knowledge Based System

Knowledge based system (KBS) has been an important theme in information systems research for decades (Giboney, Brown, Lowry, & Nunamaker Jr, 2015). It is a computer application of Artificial Intelligence which simulates the performance of a human expert in a specific filed. KBS could be regarded as a computer-based technique that facilitate managerial decision-making by presenting various effective alternatives.

This algorithm has been applied in many domains, such as medical diagnosis (Naser & ALmursheidi, 2016; Nilashi, Ibrahim, Ahmadi, & Shahmoradi, 2017), credibility assessment (Jensen, Lowry, Burgoon, & Nunamaker, 2010) and recommendation (Vijayakumar, Vairavasundaram, Logesh, & Sivapathi, 2019).

The core components of knowledge-based systems are knowledge-database and inference/reasoning mechanisms (Huang, 2009). Such a problem processing system which aims at retrieving information from a knowledge database and use associated information to present answers for assisting humans in decisions making. (Dhaliwal & Benbasat, 1996) defined four main elements of KBS: 1) knowledge-base, 2) inference engine, 3) knowledge engineering tool, and 4) specific user interface. Subsequently, (Chau & Albermani, 2002) compress the components of KBS to three: 1) knowledge-base, 2) context and 3) inference mechanism. The most widely used method to realize a knowledge-based system is the "if (condition) – then (action)" rule. The diagram of basic KBS is shown in Figure 36.

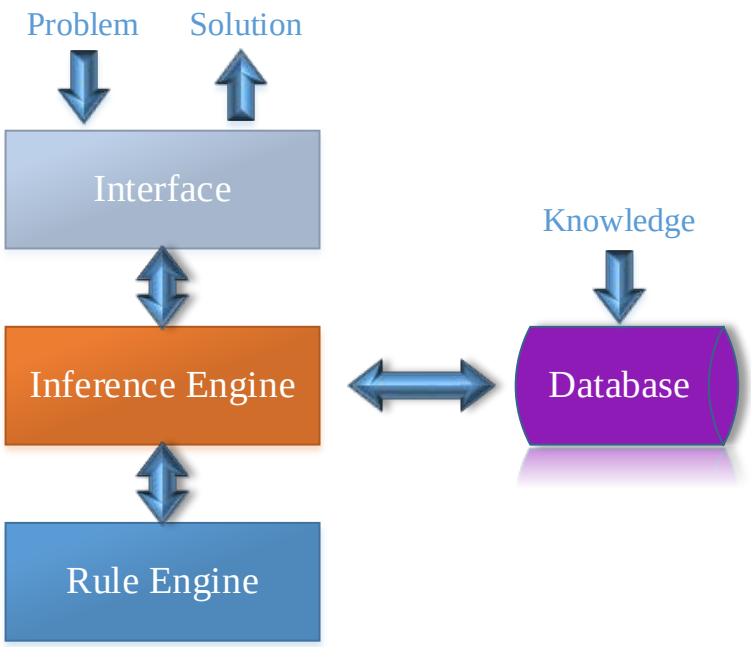


Figure 36. The schematic diagram of knowledge-based system

5.4.3. Auditory Events Response Decision Model

Generally speaking, bottom-up and top-down perception represent overlapping organizational principles rather than dichotomous constructs, and these two processes interact with each other to optimize attentional performance (Egeth & Yantis, 1997). In line with the above-presented findings and results, the knowledge-based system for auditory events response decision (AERD) which take advantages of both top-down and bottom-up mechanism is proposed to simulate the selection and filtering operation in auditory cognition processing. The diagram of proposed AERD model is presented in Figure 37.

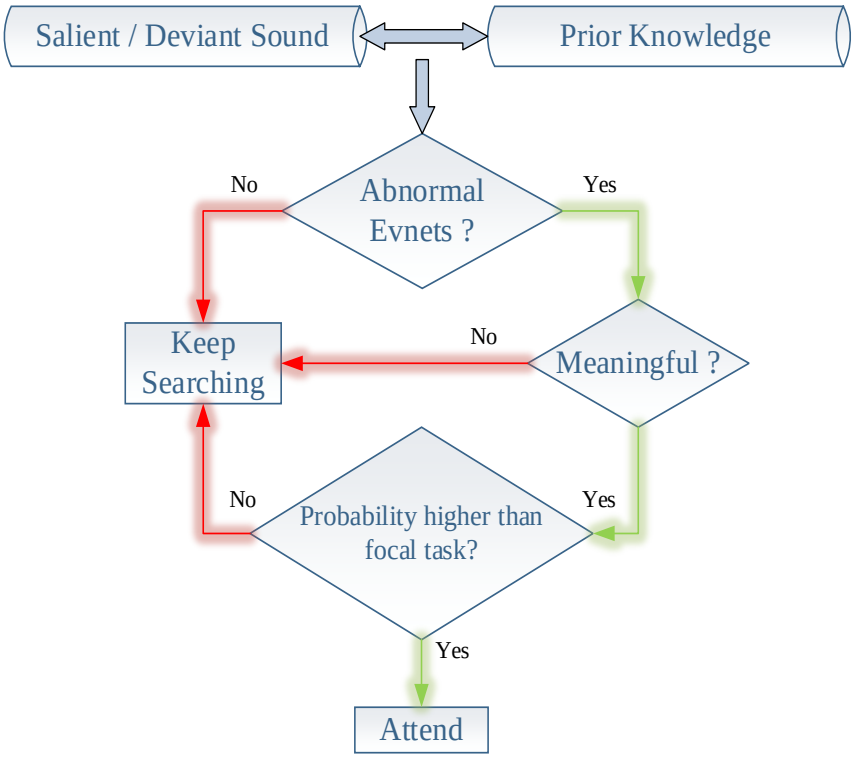


Figure 37. The diagram of AERD model.

From Figure 37, it can be seen that there are three “if - then” judgment steps and two solutions. The salient or deviant sound events will be compared with the knowledge to judge whether it is a normal or abnormal event at first. If it is a normal event, the

AERD model would decide to keep searching new events. Otherwise, the input will be further judged that if this sound event is meaningful or not according to the prior set significance and attention threshold. If the abnormal sound event is meaningless, the model will decide to launch keep searching operation. Otherwise, the significance judgment rule will be performed to judge whether the significance of meaningful abnormal events is higher than the prominence of the focal task. If the answer is yes, the AERD model will suggest cognition system pay attention to such sound events, otherwise, the system will turn to search new abnormal sound events. The mathematical description of the operation mode of AERD model is presented as follows:

Assuming that the scenario $Scene_j$ is already known, each normal sound events $(N_1^j, N_2^j, N_3^j, ..., N_n^j)$ and abnormal sound events $(AN_1^j, AN_2^j, AN_3^j, ..., AN_m^j)$ are distributed a probability, where N_n^j represents the n_{th} normal events in j_{th} scene and AN_m^j represents the m_{th} abnormal events in $Scene_j$. Letting x_i^j denote the recognized sound events in j_{th} scene, the first judgment rule can be elaborate as:

$$\begin{cases} x_i^j \ is\ normal\ event, & if\ \ x_i^j \in (N_1^j, N_2^j, N_3^j, ..., N_n^j) \\ x_i^j \ is\ abnormal\ event, & if\ \ x_i^j \in (AN_1^j, AN_2^j, AN_3^j, ..., AN_m^j) \end{cases} \tag{0.23}$$

If x_i^j is determined an abnormal sound event, the second judgment rule will be triggered to decide whether x_i^j is meaningful or meaningless:

$$\begin{cases} x_i^j \ is\ meaningful, & if\ \ P\left(x_i^j\right) \geq \alpha_P \\ x_i^j \ is\ meaningless, & if\ \ P\left(x_i^j\right) < \alpha_P \end{cases} \tag{0.24}$$

Where P denote the level of significance of a sound event, α_P is the attention threshold. Since one sound can be normal or abnormal in different scenarios, and its saliency or deviancy can vary with the scene, which means that the significance of same sound event might be different in different scenes. Hence, $P\left(x_i^j\right)$ represents the significance of sound event x_i^j in j_{th} environment. Finally, the meaningful abnormal

sound event x_i^j will be determined if its probabilistic is higher than previous sound event:

$$\begin{cases} if\ P\left(x_i^j\right) \geq P\left(y_i^j\right),\ attend \\ if\ P\left(x_i^j\right) < P\left(y_i^j\right),\ keep\ searching \end{cases} \tag{0.25}$$

Where $P\left(y_i^j\right)$ is the significance of previously salient sound which attract attention in $Scene_j$.

5.4.4. Experiment Validation

5.4.4.1. Experiment setup

To validate the effectiveness of proposed KBS based auditory events response decision model as well as the artificial auditory cognition system (AAC), typical perception tasks are built to cover the characteristics of classical environment scenes. An office scene is considered in the experiment with four sound scene conditions, which is: 1) only one sound event exists, 2) meaningless abnormal sound events occurred, 3) meaningful abnormal sound events occurred while the significance is lower than focal task, and 4) meaningful abnormal sound events occur with higher significance than focal task.

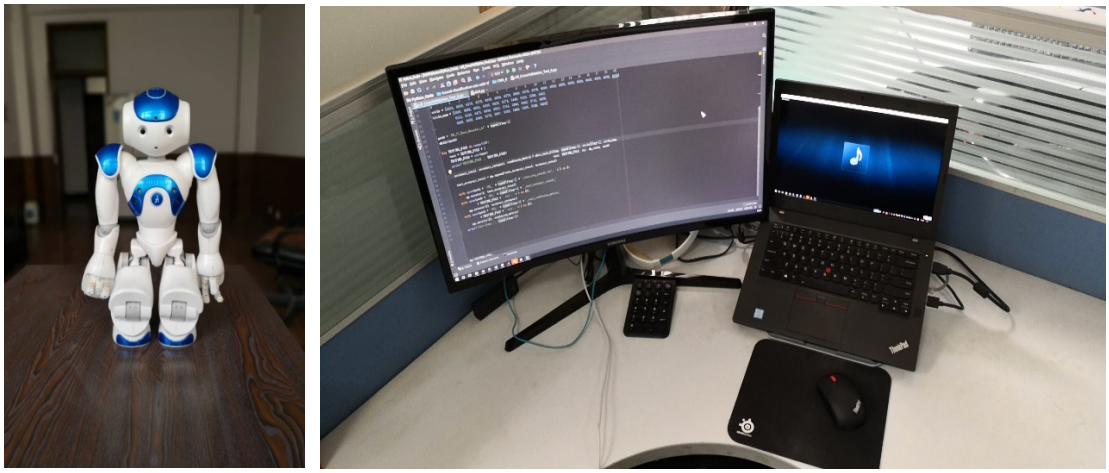


Figure 38. The Nao robot and data processing equipment.

Nao robot is exploited as the platform to perceive the surrounding environment on account of such robot posse four microphones. It should be noticed that, according to the storage condition and processing ability, it is nearly impossible to embed all modules that belong to AAC to Nao robot. Therefore, the robot is mainly applied as a sound events observer and most of the computation works will be conducted on computers.

Table 14. The composition of normal and abnormal sound events in an office

Environment	Office	
	Normal Sound Events	Abnormal Sound Events
Sound Events	talking (0.9), knocking (0.6) keyboard tapping (0.3) footstep (0.8), pouring water (0.3) air conditioner (0.2), …	siren (1.0), car horn (0.3) crackling fire (0.9) dog (0.2), crickets (0.3) thunderstorm (0.2), …

The office scenario is chosen to verify the efficiency of the proposed system, where Table 14 presents portion of normal and abnormal sound events in office environments. The values in parentheses of objects denote the level of significance of the sound event in office scenario. For example, “talking (0.9)” means that the sound event of “talking” is a normal sound event in office environment, and the level of significance of such a signal is 0.9. The value “1.0” represents the most meaningful sound while the most meaningless sound is represented by “0.1”. Meanwhile, the ellipsis symbols in Table 5 represent that more objects can be considered as probable candidates which exist in the scene and the presented objects are limited examples in the scene. The attention threshold α is set to 0.5 in the following experiments.

5.4.4.2.Experiment Results

For the purpose of validate the effectiveness of the proposed artificial cognition framework, simulated perception tasks correlate to each sound situation in an office

environment scene will be designed. The NAO robot platform uses Python programming language to obtain sound data by calling the interface of the microphone module, and the collected sound information is stored in a computer and Python is used to process them. Meanwhile, the computer is used to display each normal and abnormal sound event, since several abnormal events rarely appear under normal conditions. Consequently, in order to verify the efficiency of the proposed system in different conditions, it is needed to simulate the generation of some events that rarely appear. Furthermore, in order to reduce the interference caused by non-human factors in the experiments, the background noise of the experimental environment is controlled at a low level, so that the sound signals apply in the experiment are significant signals.

Experiment 1

The first experiment aims at testing AAC system when only one sound event exists in the office. Footstep sound signal is displayed at first and Nao robot records this signal and stored it in the processing computer. Then, the deviancy detection module of AAC system is triggered. The results indicate that the deviancy detection module successfully detected the footstep, the spectrogram and image indicator of deviancy detection result are shown in Figure 39.

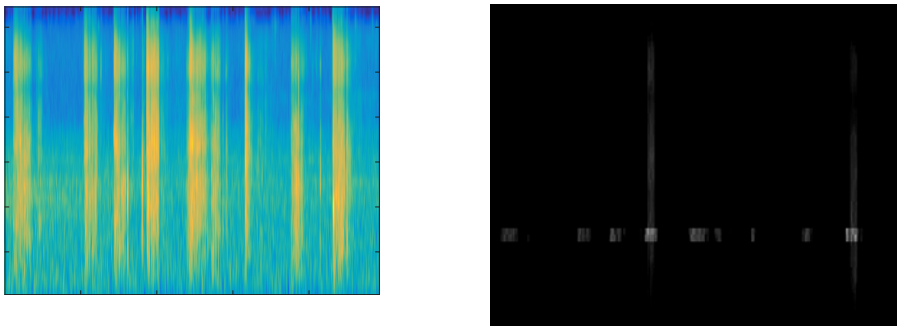


Figure 39. The deviancy detection results in experiment 1.

Thereafter, the sound is processed through TSCNN model to recognized the class of the deviant sound, followed by response judgment through AERD model. The results

are shown in Figure 40. It is clearly shown in Figure 15 that the ESC module can precisely identify the categorize of the displayed sound event. In the first experiment, since the footstep sound is a normal sound event in the office scene according to Table 14, hence, the AERD model gives the result of “keep searching”.

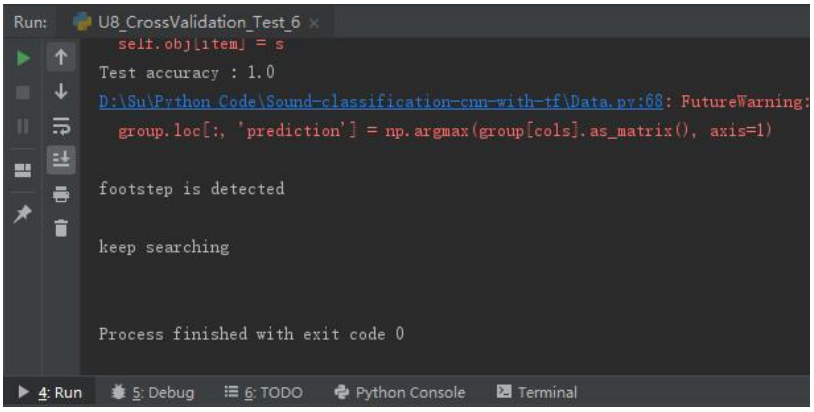


Figure 40. Environmental sound event cognition results under first scene condition.

Experiment 2

The second experiment aims at testing AAC system when meaningless abnormal sound events occurred in the office. Dog barking sound event is used in this experiment. When the sound is displayed, the robot recorded this event and the deviancy detection module is performed to analyze the signal. The results are shown in Figure 41, in which it can be seen that the dog bark is detected.

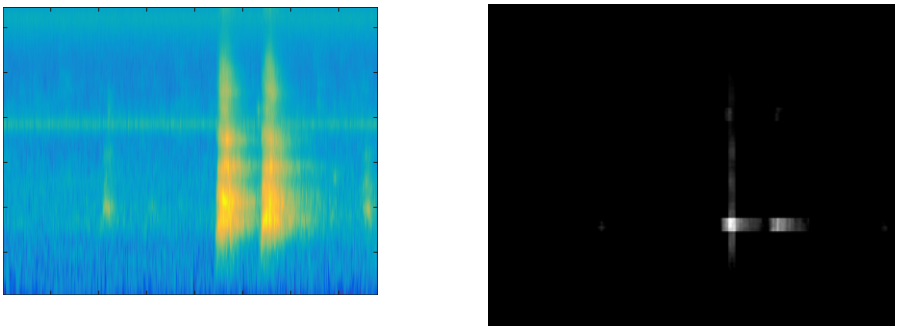


Figure 41. The deviancy detection results in experiment 2.

Subsequently, TSCNN model is triggered to identify the class of the sound event. TSCNN can accurately identify the signal's category, which is dog barking. On account of the level of significance presented in Table 14, and according to Eq. (5.5), the significance of dog barking in office environment is:

$$P\left(x_{dog}^{office}\right) < 0.5 \tag{0.26}$$

Therefore, the AERD model suggests "keep searching" in such scene. The classification and judgment results are shown in Figure 42.

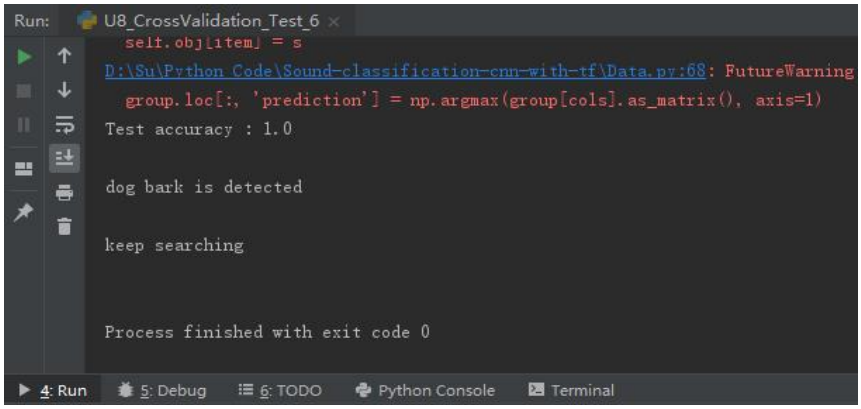


Figure 42. Environmental sound event cognition results under second scene condition.

Experiment 3

The third experiment aims at testing AAC system when meaningful abnormal sound events occurred in the office while the probability is lower than the focal task.

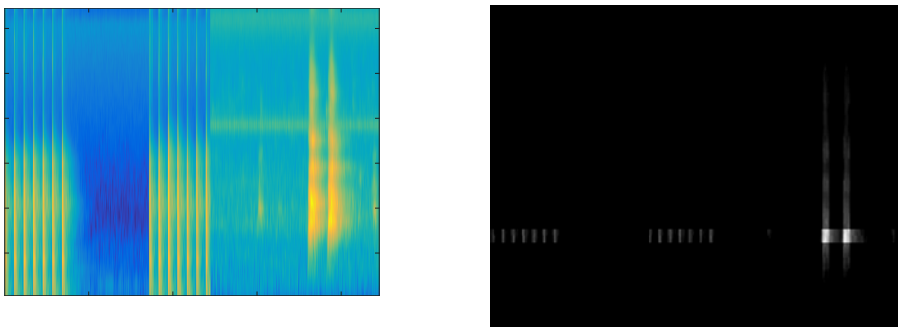


Figure 43. The deviancy detection results in experiment 3.

Two sound events are played one by one, the first is knocking and the second is dog barking. These sounds are analyzed through deviancy detection module simultaneously and the results are shown in Figure 43. It can be seen that the dog barking event is identified as a deviant sound event among these two events. Thereafter, the sound is processed through TSCNN model to recognized the class followed by response judgment through AERD model. From Table 14 it can be noticed that, although dog barking is the deviant sound in this condition, however, the significance level of such sound event is lower than previous event:

$$P\left(x_{dog}^{office}\right) < P\left(x_{knocking}^{office}\right) \tag{0.27}$$

Hence, the AERD model give the solution of “keep searching” in this scene. The classification results derived from TSCNN and judgment result presented by AERD module through Python programming language is shown in Figure 44.

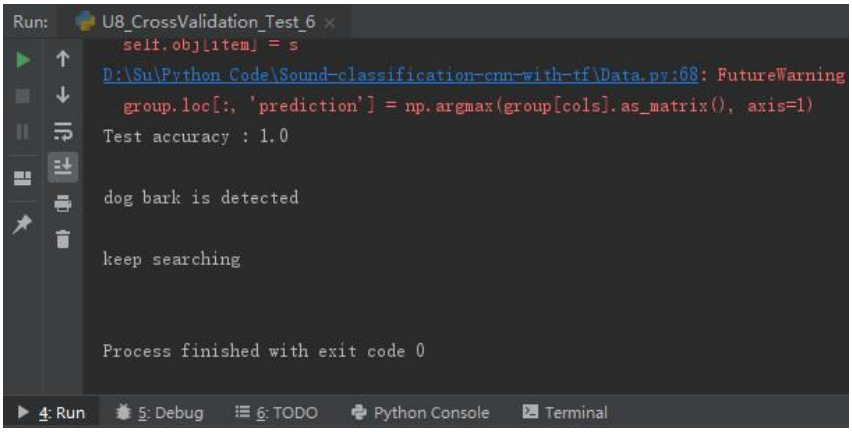


Figure 44. Environmental sound event cognition results under third scene condition.

Experiment 4

The fourth experiment aims at testing AAC system when meaningful abnormal sound events appeared in the office with a higher probability than the focal task. In this assumed condition, air conditioner and siren sound events are exploited. The air conditioner is displayed at first followed by siren. These sound events are analyzed through deviancy detection module simultaneously, where Figure 45 presents the

deviancy detection results.

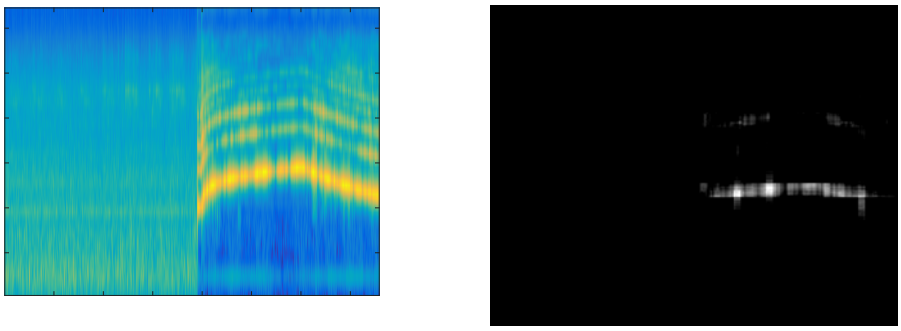


Figure 45. The deviancy detection results in experiment 4.

It is clearly shown in Figure 45 that the second sound event is the deviant sound in this scene. Then, the class the sound is processed through TSCNN model to recognized the class followed by response judgment through AERD model. The results are shown in Figure 46. It can be noticed that the siren is precisely identified. Finally, according to Table 14 and Eq. (5.6), the relation of the significance of both sounds is:

$$P\left(x_{siren}^{office}\right) > P\left(x_{air\ conditioner}^{office}\right) \tag{0.28}$$

Therefore, the AERD model suggests “please be aware” to the siren sound events in such scene. The classification and judgment results are shown in Figure 46.

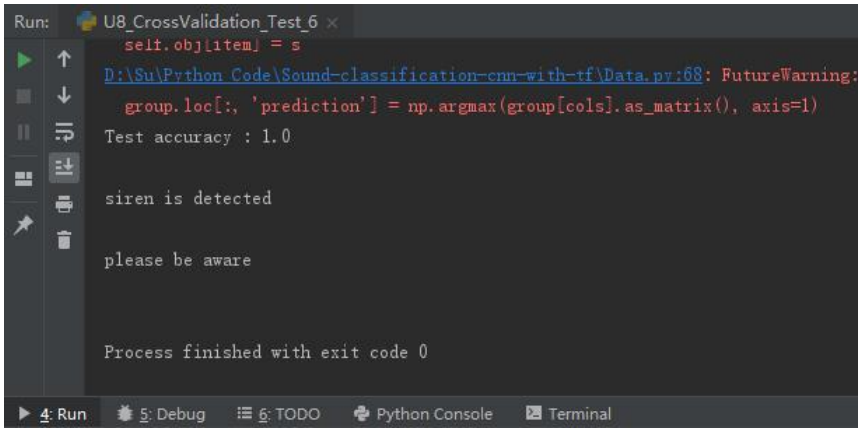


Figure 46. Environmental sound event cognition results under fourth scene condition.

Experiment 5

In order to test the validity of the AAC system in a more quantitative manner, comprehensively perception tasks correlate to each sound situation of the office scene are conducted. As the for experiments presented above, in this experiment, there are also four sound situations: 1) The first one is only one sound event exists in the office, and the sound clips of air conditioner are used here. 2) The second situation is only meaningless abnormal sound events occurred, and the sound of jackhammer is applied as the deviant sound. 3) The third one is the meaningful abnormal sound events occurred in the office while the probability is lower than the focal task, and the knocking is focal task while the dog barging is the meaningful abnormal sound events. 4) The fourth situation is when meaningful abnormal sound events appeared in the office with a higher probability than the focal task. A series acoustic segments of siren select from the UrbanSound8K dataset are chosen to test the accuracy of the proposed system in the 4th situation. The office scene is chosen as soundscape in this experiment and the sound of air conditioner is selected as normal sound events. The details including sound class, number of total segments, number of correct and incorrect detected segments and the incorrect sound classes of each situation are shown in Table 15. The abbreviations in Table 15 refer to: air conditioner (ac), children playing (cp), dog bark (db), drilling (dr), jackhammer (jh), siren (si) and street music (sm). In the column “Incorrect class”, the superscript indicates the number of occurrences of this class of sound.

Table 15. The details of each sound situation in experiment 5

	Sound class	Total segments	Correct segments	Incorrect segments	Incorrect class
Situation 1	ac	100	100	0	
Situation 2	jh	96	91	5	cp^1, dr^1, sm^3
Situation 3	db	100	92	8	ac^7, cp^1
Situation 4	si	71	68	3	cp^2, db^1

Table 15 shows the results obtained from the preliminary analysis of experiment 5. It is apparent from this table that very few incorrect detection results occur in each sound situation. In sound situation 1, it can be seen that all the tested 100 sounds (air conditioner) have been correctly detected in the office scene. In second situation, 96 meaningless abnormal sound events of jackhammer have been used in this test. 91 clips are correctly recognized and only 5 segments are incorrectly classified. Among these errors, one clip is recognized as children playing and one is detected as drilling, three segments are classified as street music. In sound situation 3, 100 sound segments of dog barging are applied, and 92 of them have been correctly detected, 8 are not detected correctly. One of the incorrectly classified sound events is considered to be children playing while the rest is recognized as the sound of air conditioner. In the last situation, 71 segments of siren are used as meaningful abnormal sound events that appeared in the office with a higher probability than the focal task. Only 3 of them are incorrectly detected, two of them are classified as children playing while the other one is considered to be dog barging. These results indicate that the proposed AAC system is very robustness in different sound detection tasks. The statistical detection results are shown in Figure 47.

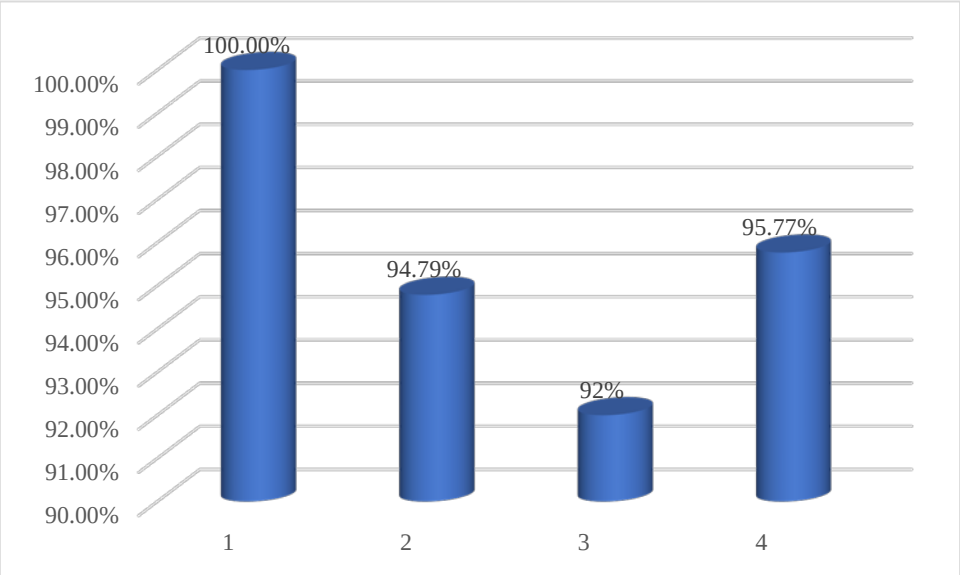


Figure 47. The statistical detection results of four sound situations in the office scene.

In Figure 47 we can see that in first sound case, where only one normal sound

event occurred, the sound clips of air conditioner could be detected accurately in the office scene. In the second case, where only meaningless abnormal sound events (jackhammer) occurred, the AAC system achieves a detection rate, which is 94.79%. In the third soundscape, where meaningful abnormal sound events occurred in the office while the probability is lower than the focal task. It can be seen from Figure 47 that the recognition rate of deviant sound events of dog barging achieves 92% in such condition. In the last soundscape, the recognition performance of deviant sound of siren, which arouses the auditory attention shift, can achieve a high accuracy of 95.77%. These results clearly point out that the proposed AAC system could get a considerable performance in different auditory artificial cognition tasks.

5.5. Conclusion

In this chapter, the artificial auditory cognition system which contains three modules including deviancy detection module, environmental sound classification module and acoustic event response module is initially proposed to achieve complex perception tasks. To be specific, in order to further improve the performance of the CNN-based ESC model, the TSCNN model is proposed to precisely identify the class of environmental sounds. It consists of two 4-layer convolutional neural networks, the LMCNet and MCNet trained by two combined features, LMC and MC feature sets, respectively. Then, the outputs of the softmax layer of both networks are fused through DS evidence theory, the result is the predicted categorize of an environmental sound. The performance of two CNN with the novel combined feature sets and the entire framework is tested on the UrbanSound8K dataset and compared with existing models published in recent years. Both LMCNet and MCNet can obtain better classification accuracy when compared with existing methods that use the same features (LM or MFCC) to form a combined eigenvector. These results indicate that the proposed CNN architecture is more effective for environment sounds classification tasks according to the appropriate parameter settings and comprehensive representation of sound recordings through the combined feature sets. Finally, TSCNN achieves 97.2% on the

UrbanSound8K dataset, which is 4.2% higher than the state-of-art methods (the (Boddapati et al., 2017) model), and is 5% higher than the (S. Li et al., 2018) model where the same fusion algorithm is exploited in this work. These results indicate that the proposed TSCNN model present to be more efficient and robustness than existing models in ESC tasks.

Thereafter, a knowledge-based system inspired auditory events response decision model is originally proposed to better describe the significant characteristic of acoustic information obtained from the environment. Inspired by the perceptual process of human cognition mechanism, the proposed method is performed by comparing the prior knowledge-based significance of detected salient or deviant sounds with sound scenes information to determine whether the system needs to respond to the abnormal sound events. Thus, abnormal sounds will be further categorized into meaningful and meaningless events, which means that meaningful deviant sounds need to respond and meaningless events do not need to respond. Meanwhile, the meaningful events need to be judged whether their significance is higher than focal tasks. If so, such events should be focused on. Otherwise, they will be neglected. By using the AERD model, the detected sound events can be judged whether they are valuable focused or not.

At last, four major perception tasks are designed to verify the performance of the proposed framework. As objects can be subjectively characterized into normal and abnormal according to the environment, the abnormal events can be effectively perceived and recognized through deviancy detection module and TSCNN. Due to the usage of the auditory events response decision model, various kinds of situations that could happen during perception can be correctly processed. The experimental results of simulated perception tasks have shown that the proposed artificial auditory cognition system can efficiently aware of the surrounding environment with prior scene knowledge.

It can be also supported by the experiments that the proposed framework could cover most of the perception requirements. Particularly, the real deviant sound could be distinguished among multiple environmental events by applying deviancy detection model and could be precisely identified through TSCNN model. Therefore, the

proposed approach is considered to be promising for achieving intelligent perception ability in complex environments.

Chapter 6. General Conclusion

Conclusion

Cognition of the surrounding environment using auditory information should be an important function of intelligent machines. Considering that the realization of bio-inspired auditory cognition is a complex systematic research work, it will be quite difficult to model such a mechanism directly. Consequently, the artificial auditory cognition modeling work is divided into three steps in this thesis. Which is modeling the human auditory attention mechanism-based information acquisition module at first. Then, the realization of the sound event recognition method simulates the transformation of low-level information to high-level information in the human brain. Thus, establish the auditory response decision model to judge the significance of sound events. Finally, these modules are combined to realize the artificial auditory cognition.

According to the comprehensive review works of the state-of-art studies presented in Chapter 2, it can be seen that auditory saliency and deviancy detection mechanism can be used as the most efficient principle in obtaining the novel sound events. The deviancy detection mechanism could be regarded as a supplement to saliency detection, a bottom-up selection mechanism made up of both helping us to perceive the environment more precisely. However, the research issue of auditory deviancy detection is more complex than auditory saliency detection, because a sound event should be salient at first, then, it could be deviant. Hence, novel bio-inspired attention models aim at detecting the sound deviancy should be proposed. Moreover, the research work of environmental sound classification is still at an early stage. This is because that the environmental sounds are a very diverse group of everyday audio events on account of the considerably non-stationary characteristics that cannot be described as only speech or music, leading to the classification accuracy of existing models that cannot reach a satisfactory level. In order to comprehensively simulate human auditory cognition, the capability of responding to an auditory event should be considered as

well. However, the main achievements are almost all focused on shed light on the theoretical basis of such ability. Only a few published works present applicable computational models with response mechanism, yet it can be found that these models are insufficient for environment perception.

For conquering the above-mentioned obstacles and realizing the artificial auditory cognition, the solutions and novel models are presented in Chapters 3 to 5. In Chapter 3, a computational model is proposed to mimic such a human auditory attention mechanism, where saliency principle and deviancy principle are used as the theoretical basis. The prosed model consists of two modules: temporal deviancy detection and frequency saliency detection. Combining the information issued from each of the aforementioned modules, the proposed model generates the image indicator that identifies the deviant salient-sound which elicit auditory attention shifts. The sounds recorded from the real environment have been used for verifying the advantages of the proposed model. The results show that the proposed model is able to point out the deviant salient-sound in a mixture of a sound clip and shows acceptable robustness and accuracy. Furthermore, a more comprehensive experiment is performed and illustrates that the proposed model could effectively simulate the human auditory attention mechanism.

In Chapter 4, considering that accurate classification of acoustic events is one of the foundations of environment acoustic awareness that has a strong correlation with the selected features. Therefore, a performance analysis work of different acoustic features aggregation schemes in ESC tasks is presented. This work aims at finding the best feature aggregate strategies to overcome the challenging problem of elevating the classification accuracy of environmental sounds. Six basic acoustic features derived from the frequency domain and two kinds of perceptually motivated acoustic features with a 6-layer convolutional neural network (CNN) model. Then, eight feature aggregate schemes were presented and evaluated on the proposed model, where the best classification accuracy is acquired by the MFCC-Log-Mel Spectrogram-CST (M-LM-C) feature sets. The categorizing accuracy of the proposed aggregate feature M-LM-C feature with CNN can reach 85.6% on ESC-50 and 93.4% on UrbanSound8K,

respectively, and is 19.7% and 20.7% higher than the (Piczak, 2015a) model.

In Chapter 5, the TSCNN model is proposed at first to precisely identify the class of environmental sounds with two aggregated features. The TSCNN is consists of two identical 4-layer CNN use LMC and MC as features, separately. CNN uses LMC features is the LMCNet and the other is the MCNet. Softmax outputs of both CNNs are fused through DS evidence theory, the fusion result is the predicted categorize of an environmental sound. he performance of two CNN with the novel combined feature sets and the entire framework is tested on the UrbanSound8K dataset and compared with existing models published in recent years. Both LMCNet and MCNet can obtain better classification accuracy when compared with existing methods that use the same features (LM or MFCC) to form a combined eigenvector. These results indicate that the proposed CNN architecture is more effective for environment sounds classification tasks according to the appropriate parameter settings and a comprehensive representation of sound recordings through the combined feature sets. Finally, TSCNN achieves 97.2% on the UrbanSound8K dataset, which is 4.2% higher than the state-of-art methods ((Boddapati et al., 2017) model), and is 5% higher than (S. Li et al., 2018) model where same fusion algorithm is exploited in this work. These results indicate that the proposed TSCNN model present to be more efficient and robustness than existing models in ESC tasks.

An auditory events response decision model is proposed to judge the significant characteristic of acoustic information obtained from the environment is proposed in Chapter 5 as well. Inspired by the perceptual process of human cognition mechanism, the proposed method is performed by comparing the detected salient or deviant sounds with sound scenes information which has previous distributed significance value to determine whether the system needs to respond to the abnormal sound events. Thus, abnormal sounds will be further categorized into meaningful and meaningless events, which means that meaningful deviant sounds need to respond and meaningless events do not need to respond. The meaningful events need to be judged whether their significance is higher than focal tasks to further determine if it is worth attention or not. At last, these proposed modules are combined to yield the final artificial auditory

cognition system. In order to verify the efficiency of the framework, the simulated perception task correlates to each sound situation in an office environment scene are designed. The experimental results of these perception tasks have shown that the proposed system can efficiently aware of the surrounding environment.

Perspectives

In this thesis, an artificial auditory cognition system consists of a deviancy detection based auditory attention model, TSCNN sound classification model and auditory events response decision model is presented for complex environmental auditory cognition. Although this bio-inspired system has achieved competitive results for the intelligent machine, the following perspectives could be considered in the future.

As for the deviancy detection model, though various acoustic features and sample entropy-based sound event deviancy detection method have been proposed for deviant environmental sounds detection, the proposed method present to be unsatisfactory when the level of intensity of background noise is high. Thus, noise reduction techniques can be introduced in the pre-processing stage to reduce the impact of noise. Moreover, dynamic information of sounds could also be exploited in deviancy detection. Since it has been described in Chapter 1 that a novel event is determined with deviancy should satisfy that such event breaks the existing status of the current environment which it appears. However, the deviant sound might turn out to be a normal sound event in the environment as time goes by. Thus, considering this characteristic can make this model take advantage of both short-term information and long-term information in precepting the real environment.

Regarding the environmental sound classification tasks, the DS theory could substantially improve the taxonomic performance of a single CNN model in ESC problems. However, it can be seen that the accuracy of repeated discrete sounds (car horn, dog barging and gunshot) is worse than other sound classes. This is may cause by the number of convolutional layers, which make the model cannot extract enough feature maps to comprehensively represent important information of sound signals.

Another probability is the feature (LC and MC) may neglect some needed information for representing such discrete sound signals. To improve the categorization accuracy on these kinds of sounds with the TSCNN-DS model will be the future works. Both of new feature extraction methods and novel CNN architectures should be established for conquering these problems and improve the classification performance. Meanwhile, the computation cost should also be considered to make the ISR model can be applied in real-time.

The ultimate goal of this thesis is to establish a practical artificial auditory cognition system for the intelligent machine to aware of the environment with auditory information. Even though the experiments designed in Chapter 5 have illustrated the effectiveness of the proposed fusion framework in dealing with different simulated scenes. The system should be tested in the real environment on intelligent machines or robots and should have the ability to deal with acoustic information on time.

Publications

Su, Y., Wang, J., Zhang, K., & Madani, K. Computational modeling of environment deviant sound detection based on human auditory cognitive mechanism. Biologically Inspired Cognitive Architectures, 2018, 24, 87–97.

Su, Y., Wang, J., Zhang, K., & Madani, K. and Wang, X. (2018). Computational Modelling Auditory Awareness. In Proceedings of the 10th International Joint Conference on Computational Intelligence - Volume 1: IJCCI, ISBN 978-989-758-327-8, pages 160-167.

Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream cnn based on decision-level fusion. Sensors, 19(7), 1733.

Su, Y., Zhang, K., Wang, J., Zhou, D., & Madani, K. (2020). Performance analysis of multiple aggregated acoustic features for environment sound classification. Applied Acoustics, 158, 107050.

Bibliography

Adavanne, S., Parascandolo, G., Pertilä, P., Heittola, T., & Virtanen, T. (2017). Sound event detection in multichannel audio using spatial and harmonic features. *ArXiv Preprint ArXiv:1706.02293*.

Adavanne, S., Pertilä, P., & Virtanen, T. (2017). Sound event detection using spatial features and convolutional recurrent neural network. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On*, 771–775. IEEE.

Adavanne, S., & Virtanen, T. (2017). Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. *ArXiv Preprint ArXiv:1710.02998*.

Adiga, A., Magimai, M., & Seelamantula, C. S. (2013). Gammatone wavelet Cepstral Coefficients for robust speech recognition. *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*, 1–4. https://doi.org/10.1109/TENCON.2013.6718948

Aleksander, I. (2001). The Self'out there'. *Nature*, *413*(6851), 23.

Ali, H., Tran, S. N., Benetos, E., & d'Avila Garcez, A. S. (2018). Speaker recognition with hybrid features from a deep belief network. *Neural Computing and Applications*, *29*(6), 13–19. https://doi.org/10.1007/s00521-016-2501-7

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults.

Hearing Research, *300*, 18–32. https://doi.org/10.1016/j.heares.2013.03.006

Anwer, R. M., Vázquez, D., & López, A. M. (2011). Opponent colors for human detection. *Iberian Conference on Pattern Recognition and Image Analysis*, 363–370. Springer.

Awh, E., Vogel, E. K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208.

Aytar, Y., Vondrick, C., & Torralba, A. (n.d.). *SoundNet: Learning Sound Representations from Unlabeled Video*. 9.

Baars, B. J., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, *1*(01), 23–32.

Bartsch, M. A., & Wakefield, G. H. (2005). Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, *7*(1), 96–104.

Baum, E., Harper, M., Alicea, R., & Ordonez, C. (2018). Sound Identification for Fire-Fighting Mobile Robots. *2018 Second IEEE International Conference on Robotic Computing (IRC)*, 79–86. https://doi.org/10.1109/IRC.2018.00020

Bayne, T., Cleeremans, A., & Wilken, P. (2014). *The Oxford companion to consciousness*. Oxford University Press.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, *7*(3), 295–301. https://doi.org/10.1038/nn1198

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and*

Brain Sciences, *18*(2), 227–247.

Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Computer Science*, *112*, 2048–2056. https://doi.org/10.1016/j.procs.2017.08.250

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

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 111–118.

Burgos, W. (2014). *Gammatone and MFCC Features in Speaker Recognition* (PhD Thesis). Florida Institute of Technology, Melbourne, Florida.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860–1862.

Cakir, E., Heittola, T., Huttunen, H., & Virtanen, T. (2015). Polyphonic sound event detection using multi label deep neural networks. *Neural Networks (IJCNN), 2015 International Joint Conference On*, 1–7. IEEE.

Chachada, S., & Kuo, C.-C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, *3*. https://doi.org/10.1017/ATSIP.2014.12

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford university press.

Chau, K.-W., & Albermani, F. (2002). Expert system application on preliminary design of water retaining structures. *Expert Systems with Applications*, *22*(2), 169–178.

Chen, Y., Guo, Q., Liang, X., Wang, J., & Qian, Y. (2019). Environmental sound classification with dilated convolutions. *Applied Acoustics*, *148*, 123–132. https://doi.org/10.1016/j.apacoust.2018.12.019

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Chia Ai, O., Hariharan, M., Yaacob, S., & Sin Chee, L. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*, *39*(2), 2157–2165. https://doi.org/10.1016/j.eswa.2011.07.065

Chu, S., Narayanan, S., & Kuo, C.-C. J. (2009). Environmental Sound Recognition With Time–Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1142–1158.

Churchland, P. M. (1984). *Matter and consciousness*. MIT press.

Cohn, R. (1998). Introduction to neo-riemannian theory: A survey and a historical perspective. *Journal of Music Theory*, 167–180.

Cowling, M., & Renate, S. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, *24*(15), 2895–2907. https://doi.org/10.1016/S0167-8655(03)00147-8

Crick, F., & Clark, J. (1994). The astonishing hypothesis. *Journal of Consciousness Studies*, *1*(1), 10–16.

Dai, W. (2016). *Acoustic Scene Recognition with Deep Learning*. Carnegie Mellon.

Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2017). Very deep convolutional neural networks for raw waveforms. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On*, 421–425. IEEE.

Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, & Lian-Hong Cai. (2002). Music type classification by spectral contrast feature. *Proceedings. IEEE International Conference on Multimedia and Expo*, 113–116. https://doi.org/10.1109/ICME.2002.1035731

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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222.

Dhaliwal, J. S., & Benbasat, I. (1996). The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation. *Information Systems Research*, *7*(3), 342–362.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, *92*(1), 53–78.

Duangudom, V., & Anderson, D. V. (2007). Using auditory saliency to understand complex auditory scenes. *Signal Processing Conference, 2007 15th European*,

1206–1210. IEEE.

Edelman, G. M., & Tononi, G. (2003). A universe of consciousness: How matter becomes imagination. *Contemporary Psychology*, *48*(1), 92.

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology*, *48*(1), 269–297.

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience*, *10*(5), 590–604.

Escera, C., Leung, S., & Grimm, S. (2014). Deviance detection based on regularity encoding along the auditory hierarchy: Electrophysiological evidence in humans. *Brain Topography*, *27*(4), 527–538. https://doi.org/10.1007/s10548-013-0328-4

Escera, C., & Malmierca, M. S. (2014). The auditory novelty system: An attempt to integrate human and animal research: The auditory novelty system. *Psychophysiology*, *51*(2), 111–123. https://doi.org/10.1111/psyp.12156

Espi, M., Fujimoto, M., Kinoshita, K., & Nakatani, T. (2015). Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1). https://doi.org/10.1186/s13636-015-0069-2

Evangelopoulos, G., Rapantzikos, K., Maragos, P., Avrithis, Y., & Potamianos, A. (2008). Audiovisual attention modeling and salient event detection. In *Multimodal Processing and Interaction* (pp. 1–21). Springer.

Ewert, S. (2011). Computer Science III University of Bonn. *Proc. ISMIR.*, 6.

Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and Cognition*, *20*(3), 807–827.

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, *7*(1), 6.

Fuertes, C. T., & Russ, G. (2002). Unification of perception sources for perceptive awareness automatic systems. *IEEE AFRICON. 6th Africon Conference in Africa*, *1*, 283–286. IEEE.

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135.

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

Ghosal, D., & Kolekar, M. H. (2018). Music Genre Recognition Using Deep Neural Networks and Transfer Learning. *Interspeech 2018*, 2087–2091. https://doi.org/10.21437/Interspeech.2018-2045

Giboney, J. S., Brown, S. A., Lowry, P. B., & Nunamaker Jr, J. F. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, *72*, 1–10.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138.

https://doi.org/10.1016/0378-5955(90)90170-T

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gouyon, F., Pachet, F., & Delerue, O. (2000). *ON THE USE OF ZERO-CROSSING RATE FOR AN APPLICATION OF CLASSIFICATION OF PERCUSSIVE SOUNDS*. 7.

Gray, J. A., Buhusi, C. V., & Schmajuk, N. (1997). The transition from automatic to controlled processing. *Neural Networks*, *10*(7), 1257–1268.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., … Cai, J. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia - AMCMM '06*, 21. https://doi.org/10.1145/1178723.1178727

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*(6), 467.

Higashi, Y., Kim, K.-S., Jeon, H.-G., & Ichikawa, M. (2010). Enhancing spectral contrast in organic red-light photodetectors based on a light-absorbing and exciton-blocking layered system. *Journal of Applied Physics*, *108*(3), 034502. https://doi.org/10.1063/1.3466766

Holland, O. (2009). *Machine consciousness*. Oxford University Press.

Hu, J., Mitchell, J. E., & Pang, J.-S. (2012). An LPCC approach to nonconvex quadratic programs. *Mathematical Programming*, *133*(1–2), 243–277.

https://doi.org/10.1007/s10107-010-0426-y

Huang, H.-C. (2009). Designing a knowledge-based system for strategic planning: A balanced scorecard perspective. *Expert Systems with Applications*, *36*(1), 209–218. https://doi.org/10.1016/j.eswa.2007.09.046

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *ArXiv:1602.07360 [Cs]*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv Preprint ArXiv:1502.03167*.

Itti, L., Koch, C., & 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.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech Recognition using MFCC. *Simulation and Modeling*, 4.

Jensen, M. L., Lowry, P. B., Burgoon, J. K., & Nunamaker, J. F. (2010). Technology dominance in complex decision making: The case of aided credibility assessment. *Journal of Management Information Systems*, *27*(1), 175–202.

Juang, B. H., & Rabiner, L. R. (2005). Automatic Speech Recognition – A Brief History of the Technology Development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

Kalinli, O., & Narayanan, S. S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.

INTERSPEECH, 1941–1944.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature Reviews Neuroscience*, *1*(2), 91.

Kaya, E. M., & Elhilali, M. (2012). A temporal saliency map for modeling auditory attention. *Information Sciences and Systems (CISS), 2012 46th Annual Conference On*, 1–6. IEEE.

Kaya, E. M., & Elhilali, M. (2013). A model of auditory deviance detection. *Information Sciences and Systems (CISS), 2013 47th Annual Conference On*, 1–6. IEEE.

Kaya, E. M., & Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160101. https://doi.org/10.1098/rstb.2016.0101

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map. *Current Biology*, *15*(21), 1943–1947. https://doi.org/10.1016/j.cub.2005.09.040

Kayser, S. J., McNair, S. W., & Kayser, C. (2016). Prestimulus influences on auditory perception from sensory representations and decision processes. *Proceedings of the National Academy of Sciences*, *113*(17), 4842–4847. https://doi.org/10.1073/pnas.1524087113

Kim, K., Lin, K.-H., Walther, D. B., Hasegawa-Johnson, M. A., & Huang, T. S. (2014). Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, *38*, 78–85.

https://doi.org/10.1016/j.patrec.2013.11.010

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*.

Klapuri, A., & Davy, M. (2007). *Signal Processing Methods for Music Transcription*. Springer Science & Business Media.

Koch, C., & Ullman, S. (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In L. M. Vaina (Ed.), *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience* (pp. 115–141). https://doi.org/10.1007/978-94-009-3833-5_5

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kuipers, B. (2005). Consciousness: Drinking from the firehose of experience. *AAAI*, 1298–1305.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Li, Jing, Qiu, T., Wen, C., Xie, K., & Wen, F.-Q. (2018). Robust Face Recognition Using the Deep C2D-CNN Model Based on Decision-Level Fusion. *Sensors*, *18*(7), 2080. https://doi.org/10.3390/s18072080

Li, Juncheng, Dai, W., Metze, F., Qu, S., & Das, S. (2017). A comparison of deep learning methods for environmental sound detection. *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference On*, 126–130. IEEE.

Li, S., Yao, Y., Hu, J., Liu, G., Yao, X., & Hu, J. (2018). An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition. *Applied Sciences*, *8*(7), 1152. https://doi.org/10.3390/app8071152

Li, Y., Chen, J., Ye, F., & Liu, D. (2016). The Improvement of DS Evidence Theory and Its Application in IR/MMW Target Recognition. *Journal of Sensors*, *2016*, 1–15. https://doi.org/10.1155/2016/1903792

Lian, Z., Xu, K., Wan, J., & Li, G. (2017). Underwater acoustic target classification based on modified GFCC features. *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 258–262. https://doi.org/10.1109/IAEAC.2017.8054017

Liao, H.-I., Yoneya, M., Kidani, S., Kashino, M., & Furukawa, S. (2016). Human Pupillary Dilation Response to Deviant Auditory Stimuli: Effects of Stimulus Properties and Voluntary Attention. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00043

Liu, C.-C., Doong, J.-L., Hsu, C.-C., Huang, W.-S., & Jeng, M.-C. (2009). Evidence for the selective attention mechanism and dual-task interference. *Applied Ergonomics*, *40*(3), 341–347.

Liu, J.-M., You, M., Li, G.-Z., Wang, Z., Xu, X., Qiu, Z., … Chen, S. (2013). Cough

signal recognition with Gammatone Cepstral Coefficients. *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference On*, 160–164. IEEE.

López-Caballero, F., Zarnowiec, K., & Escera, C. (2016). Differential deviant probability effects on two hierarchical levels of the auditory novelty system. *Biological Psychology*, *120*, 1–9. https://doi.org/10.1016/j.biopsycho.2016.08.001

Lotto, A., & Holt, L. (2011). Psychology of auditory perception: Psychology of auditory perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 479–489. https://doi.org/10.1002/wcs.123

Marchi, E., Vesperini, F., Squartini, S., & Schuller, B. (2017). Deep Recurrent Neural Network-Based Autoencoders for Acoustic Novelty Detection. *Computational Intelligence and Neuroscience*, *2017*, 1–14. https://doi.org/10.1155/2017/4694860

McDermott, D. (2007). Artificial intelligence and consciousness. *The Cambridge Handbook of Consciousness*, 117–150.

McGinn, C. (2004). *Consciousness and its objects*. Oxford University Press on Demand.

McLoughlin, I., Zhang, H., Xie, Z., Song, Y., & Xiao, W. (2015). Robust Sound Event Classification Using Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(3), 540–552. https://doi.org/10.1109/TASLP.2015.2389618

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network

model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for Polyphonic Sound Event Detection. *Applied Sciences*, *6*(12), 162. https://doi.org/10.3390/app6060162

Meyer, M., Cavigelli, L., & Thiele, L. (2017). Efficient Convolutional Neural Network For Audio Event Detection. *ArXiv Preprint ArXiv:1709.09888*.

Mohamed, A., Dahl, G. E., & 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., & Sandini, G. (2015, September). *Saliency based sensor fusion of broadband sound localizer for humanoids*. 362–367. https://doi.org/10.1109/MFI.2015.7295835

Müller, M. (2007). *Information retrieval for music and motion* (Vol. 2). Springer.

Müller, M., Kurth, F., & Clausen, M. (2005). Audio Matching via Chroma-Based Statistical Features. *ISMIR*, *2005*, 6th.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 807–814.

Naser, S. S. A., & ALmursheidi, S. H. (2016). A Knowledge Based System for Neck Pain Diagnosis. *World Wide Journal of Multidisciplinary Research and Development (WWJMRD)*, *2*(4), 12–18.

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based

system for breast cancer classification using fuzzy logic method. *Telematics and Informatics*, *34*(4), 133–144.

Oldoni, D., De Coensel, B., Boes, M., Rademaker, M., De Baets, B., Van Renterghem, T., & Botteldooren, D. (2013). A computational model of auditory attention for use in soundscape research. *The Journal of the Acoustical Society of America*, *134*(1), 852–861.

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.

Palaz, D. (2015). *Analysis of CNN-based Speech Recognition System using Raw Speech as Input*. 5.

Pan, Y., Long, R., Cheng, X., & Chen, Y. (2013). Saliency-Based Auditory Detection Method Using Energy Linear Superposition. *Pacific-Rim Conference on Multimedia*, 289–298. Springer.

Parascandolo, G., Heittola, T., Huttunen, H., & 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.

Petit, C., El-Amraoui, A., & Avan, P. (2013). Audition: Hearing and Deafness. In *Neuroscience in the 21st Century* (pp. 675–741). https://doi.org/10.1007/978-1-4614-1997-6_26

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural

networks. *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop On*, 1–6. IEEE.

Piczak, K. J. (2015b). *ESC: Dataset for Environmental Sound Classification*. 1015–1018. https://doi.org/10.1145/2733373.2806390

Pons, J., & Serra, X. (2018). Randomly weighted CNNs for (music) audio classification. *ArXiv:1805.00237 [Cs, Eess]*.

Posner, M. I., Rafal, R. D., Choate, L. S., & Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, *2*(3), 211–228. https://doi.org/10.1080/02643298508252866

Principi, E., Squartini, S., Bonfigli, R., Ferroni, G., & 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. https://doi.org/10.1016/j.eswa.2015.02.036

Rakotomamonjy, A., & Gasso, G. (2015). Histogram of gradients of time–frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 142–153.

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131. https://doi.org/10.1016/j.neunet.2013.03.011

Reineking, T. (2014). *Belief functions: Theory and algorithms* (PhD Thesis). Staats-und Universitätsbibliothek Bremen.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing.

Annu. Rev. Neurosci., *27*, 611–647.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, *278*(6), H2039–H2049.

Rijsbergen, C. J. V. (1979). *Information Retrieval* (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *22*(21), 9475–9489.

Romo, R., & Salinas, E. (2001). Touch and go: Decision-making mechanisms in somatosensation. *Annual Review of Neuroscience*, *24*, 107–137. https://doi.org/10.1146/annurev.neuro.24.1.107

Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381

Salamon, J., Jacoby, C., & Bello, J. P. (2014). *A Dataset and Taxonomy for Urban Sound Research*. 1041–1044. https://doi.org/10.1145/2647868.2655045

Santangelo, V., & Macaluso, E. (2013). Visual salience improves spatial working memory via enhanced parieto-temporal functional connectivity. *Journal of Neuroscience*, *33*(9), 4110–4117.

Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained

attention: Where top-down meets bottom-up. *Brain Research Reviews*, *35*(2), 146–160.

Schafer, R. W. (2008). Homomorphic Systems and Cepstrum Analysis of Speech. *Springer Handbook of Speech Processing*, 161–180. https://doi.org/10.1007/978-3-540-49127-9_9

Schauerte, B., & Stiefelhagen, R. (2013). "Wow!" Bayesian surprise for salient acoustic event detection. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference On*, 6402–6406. IEEE.

Schröger, E. (1996). A Neural Mechanism for Involuntary Attention Shifts to Changes in Auditory Stimulation. *Journal of Cognitive Neuroscience*, *8*(6), 527–539. https://doi.org/10.1162/jocn.1996.8.6.527

Serizel, R., Bisot, V., Essid, S., & Richard, G. (2018). Acoustic Features for Environmental Sound Analysis. In T. Virtanen, M. D. Plumbley, & D. Ellis (Eds.), *Computational Analysis of Sound Scenes and Events* (pp. 71–101). https://doi.org/10.1007/978-3-319-63450-0_4

Serra, J., Gómez, E., Herrera, P., & Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(6), 1138–1151.

Seth, A. (2009). The strength of weak artificial consciousness. *International Journal of Machine Consciousness*, *1*(01), 71–82.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 42). Princeton university press.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences*, *5*(8), 340–348.

Shamma, S. A., & Micheyl, C. (2010). Behind the scenes of auditory perception. *Current Opinion in Neurobiology*, *20*(3), 361–366. https://doi.org/10.1016/j.conb.2010.03.009

Shamma, S., & Fritz, J. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology*, *25*, 164–168.

Shannon, C. E. (2001). A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, *5*(1), 3–55. https://doi.org/10.1145/584091.584093

Shao, Y., Jin, Z., Wang, D., & Srinivasan, S. (2009). An auditory-based feature for robust speech recognition. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference On*, 4625–4628. IEEE.

Shao, Y., & Wang, D. (2008). Robust speaker identification using auditory features and computational auditory scene analysis. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference On*, 1589–1592. IEEE.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.

Slaney, M. (1994). Auditory toolbox: A MATLAB Toolbox for auditory modeling work. *URL: Http://Web. Interval. Com/Papers/1998-010*.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Phil.*

Trans. R. Soc. B, *372*(1714), 20160105.

Swee, T. T., Salleh, S. H. S., & Jamaludin, M. R. (2010). Speech pitch detection using short-time energy. *International Conference on Computer and Communication Engineering (ICCCE'10)*, 1–6. https://doi.org/10.1109/ICCCE.2010.5556836

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Takahashi, N., Gygli, M., Pfister, B., & Van Gool, L. (2016). Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Detection. *ArXiv:1604.07160 [Cs]*.

Takahashi, N., Gygli, M., & Van Gool, L. (2017). Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*.

Tokozume, Y., & Harada, T. (2017). Learning environmental sounds with end-to-end convolutional neural network. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2721–2725. https://doi.org/10.1109/ICASSP.2017.7952651

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*(4), 242–248.

Tsuchida, T., & Cottrell, G. (2012). Auditory saliency using natural statistics. *Proceedings of the Cognitive Science Society*, *34*.

Tsunada, J., Liu, A. S. K., Gold, J. I., & Cohen, Y. E. (2016). Causal contribution of

primate auditory cortex to auditory perceptual decision-making. *Nature Neuroscience*, *19*(1), 135–142. https://doi.org/10.1038/nn.4195

Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVMs with a new set of feature parameters to improve recognition rates. *Int. J. Innov. Comput. Inf. Control*, *8*(5), 3511–3524.

Vachon, F., Labonté, K., & Marsh, J. E. (2017). Attentional capture by deviant sounds: A noncontingent form of auditory distraction? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 622–634. https://doi.org/10.1037/xlm0000330

Vijayakumar, V., Vairavasundaram, S., Logesh, R., & Sivapathi, A. (2019). Effective Knowledge Based Recommender System for Tailored Multiple Point of Interest Recommendation. *International Journal of Web Portals (IJWP)*, *11*(1), 1–18.

Wang, J. (2015). *Contribution to study and implementation of a bio-inspired perception system based on visual and auditory attention* (PhD Thesis). Paris Est.

Wang, J., Zhang, K., Madani, K., & Sabourin, C. (2015). Salient environmental sound detection framework for machine awareness. *Neurocomputing*, *152*, 444–454. https://doi.org/10.1016/j.neucom.2014.09.046

Wang, J.-C., Wang, J.-F., He, K. W., & Hsu, C.-S. (2006). Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 1731–1735. https://doi.org/10.1109/IJCNN.2006.246644

Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012). End-to-end text recognition with convolutional neural networks. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 3304–3308. IEEE.

Wang, X.-J. (2008). Decision Making in Recurrent Neuronal Circuits. *Neuron*, *60*(2), 215–234. https://doi.org/10.1016/j.neuron.2008.09.034

Wills, K. M., Liu, J., Hakun, J., Zhu, D. C., Hazeltine, E., & Ravizza, S. M. (2016). Neural Mechanisms for the Benefits of Stimulus-Driven Attention. *Cerebral Cortex*, *27*(11), 5294–5302.

Winkler, I., & Schröger, E. (2015). Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and Language*, *148*, 1–22. https://doi.org/10.1016/j.bandl.2015.05.003

Xing, Z., Baik, E., Jiao, Y., Kulkarni, N., Li, C., Muralidhar, G., … Pouliot, C. (2017). Modeling of the Latent Embedding of Music using Deep Neural Network. *ArXiv:1705.05229 [Cs]*. Retrieved from http://arxiv.org/abs/1705.05229

Xu, H., Chen, T., Gao, D., Wang, Y., Li, K., Goel, N., … Khudanpur, S. (2018). A PRUNED RNNLM LATTICE-RESCORING ALGORITHM FOR AUTOMATIC SPEECH RECOGNITION. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5929–5933. IEEE.

Xu, J., Shi, J., Liu, G., Chen, X., & Xu, B. (2018). Modeling attention and memory for auditory selection in a cocktail party environment. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yakar, T. B., Litman, R., Sprechmann, P., Bronstein, A., & Sapiro, G. (2013). BILEVEL SPARSE MODELS FOR POLYPHONIC MUSIC TRANSCRIPTION. *ISMIR*, 65–70.

Ye, H., Wu, Z., Zhao, R.-W., Wang, X., Jiang, Y.-G., & Xue, X. (2015). Evaluating Two-Stream CNN for Video Classification. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*, 435–442. https://doi.org/10.1145/2671188.2749406

Ye, J., Kobayashi, T., & Murakawa, M. (2017). Urban sound event classification based on local and global features aggregation. *Applied Acoustics*, *117*, 246–256. https://doi.org/10.1016/j.apacoust.2016.08.002

Zhang, H., McLoughlin, I., & Song, Y. (2015, April). *Robust sound event recognition using convolutional neural networks*. 559–563. https://doi.org/10.1109/ICASSP.2015.7178031

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 32–32. https://doi.org/10.1167/8.7.32

Zhang, X., Zou, Y., & Shi, W. (2017). Dilated convolution neural network with LeakyReLU for environmental sound classification. *2017 22nd International Conference on Digital Signal Processing (DSP)*, 1–5. https://doi.org/10.1109/ICDSP.2017.8096153

Zhang, Z., Xu, S., Cao, S., & Zhang, S. (2018). Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. *ArXiv:1808.08405 [Cs,*

Eess].

Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(5), 1608–1616.

Zhao, X., & Wang, D. (2013). Analyzing noise robustness of MFCC and GFCC features in speaker identification. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference On*, 7204–7208. IEEE.

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589. https://doi.org/10.1007/BF02943243

Zhu, B., Wang, C., Liu, F., Lei, J., Lu, Z., & Peng, Y. (2018). Learning Environmental Sounds with Multi-scale Convolutional Neural Network. *ArXiv:1803.10219 [Cs, Eess]*.

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