University of Salford-Manchester
School of Computing, Science & Engineering

Binaural Sound Source Localization Using Machine Learning with Spiking Neural Networks Features Extraction

By

Hanaa Mohsin Ali Al- Abboodi

Supervised By

Dr Paul Kendrick **Dr Bruno Fazenda**

A thesis Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

May 2019

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

All Praise is due to Allah Lord of the Universe, the most merciful.

Firstly, I would like to express my gratitude to the Government of the Republic of Iraq including the Ministry of Higher Education and Scientific Research and The Iraqi Cultural Attaché-London for their help and support throughout my studies in the United Kingdom.

Immeasurable gratitude goes toward my supervisors, **Dr Paul Kendrick** and **Dr Bruno Fazenda,** for their continuous support, unforgettable patience and motivation, which made this work possible. It was an honour to have them as my supervisors.

Special thanks must also go to all my colleagues at the acoustics research centre. The last four years as a PhD student would certainly have been more difficult without you all. But special thanks should go to Will Bailey, Manish Varma and James Massaglia for their advice and guidance. I will never forget their help and support along the PhD trip.

I really should thank my dearest Dad and Mum for their big love and support, my beloved husband Alaa, my children Hussein and Zainalabdeen for all unlimited love and support.

Finally, a special thank for my Brothers Ali and Hussain my three sisters Rajaa, Khamaal and Nidal, and my all the encouragement and supports through the time of this research. I dedicate this work to my late father, who has always wished for me to pursue a PhD, and I hope it makes him proud. Without their love, encouragement, and prayers I would certainly not be in the position I am today, and for that I am eternally grateful.

AUTHOR PUBLICATIONS

- Al-Abboodi, H.M, Li, F.F., 2016 **Deep belief spiking neural network for Sound sources localization with HRTFs**. Salford Postgraduate Annual Research Conference (SPARC), University of Salford.

- Al-Abboodi, H.M, Kendrick, P & Fazenda, B, 2018**Automatic sound source localization in hearing aids**. Salford Postgraduate Annual Research Conference (SPARC), University of Salford.

ABSTRACT

Human and animal binaural hearing systems are able take advantage of a variety of cues to localise sound-sources in a 3D space using only two sensors. This work presents a bionic system that utilises aspects of binaural hearing in an automated source localisation task. A head and torso emulator (KEMAR) are used to acquire binaural signals and a spiking neural network is used to compare signals from the two sensors.

The firing rates of coincidence-neurons in the spiking neural network model provide information as to the location of a sound source. Previous methods have used a winner-takes-all approach, where the location of the coincidence-neuron with the maximum firing rate is used to indicate the likely azimuth and elevation. This was shown to be accurate for single sources, but when multiple sources are present the accuracy significantly reduces.

To improve the robustness of the methodology, an alternative approach is developed where the spiking neural network is used as a feature pre-processor. The firing rates of all coincidence-neurons are then used as inputs to a Machine Learning model which is trained to predict source location for both single and multiple sources.

A novel approach that applied spiking neural networks as a binaural feature extraction method was presented. These features were processed using deep neural networks to localize multi-source sound signals that were emitted from different locations. Results show that the proposed bionic binaural emulator can accurately localise sources including multiple and complex sources to 99% correctly predicted angles from single-source localization model and 91% from multi-source localization model.

The impact of background noise on localisation performance has also been investigated and shows significant degradation of performance. The multisource localization model was trained with multi-condition background noise at SNRs of 10dB, 0dB, and -10dB and tested at controlled SNRs. The findings demonstrate an enhancement in the model performance in compared with noise free training data.

LIST OF ABBREVIATIONS

AM	Amplitude Modulation
ANN	Artificial Neural Network
Azim L. Acc	Azimuth Localization Accuracy
Elve L. Acc	Elevation Localization Accuracy
A/D	Analog to Digital
BEM	Boundary Elements Method
CC	Cross-Correlation
CCA	Cross-correlation classification algorithm
CIs	Cochlear Implants
CNNs	Convolution Neural Networks
D/A	Digital to Analog
dB	decibels
DCNN	Deep Convolution Neural Networks
DNN	Deep Neural Network
DOA	Direction of Arrival
DBN	Deep Belief Network
ERB	Equivalent Rectangular Bandwidth
FB	Front-Back
FIR	Finite Impulse Response
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
HH	Hodgkin–Huxley
ICA	Independent Component Analysis
IIR	Infinite Impulse Response
ITD	Interaural Time Difference
ILD	Interaural Level Differences
IIF	Infinite Impulse Response Filter
IRCAM	Institute for Research and Coordination in Acoustics/Music
GA	Genetic Algorithm
GCC	Generalised Cross-Correlation
GF	Gammatone Features
GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
GWN	Gaussian White Noise
KEMAR	Knowles Electronic Manikin for Acoustic Research
KNN	K-nearest-neighbour algorithm
LIF	Leaky integrator and Fire
LS-SVMs	Least Squares Support Vector Machines
M-SVM	Multiclass Support Vector Machines
MEKA	Multiple Extended Kalman Algorithm
MLP	Multilayer Perceptron
MSO	Medial Superior Olive
MUSIC	Multiple Signal Classification
OAA	One-Against-All
OAO	One-Against-One
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
RF	Random Forest
RDNN	Recurrent Deep Neural Network
ReLU	Rectified Linear Unit
RBMs	Restricted Boltzmann Machines
SALU-AC	Salford University – Anechoic Chamber

SALU-CR	Salford University –Clean Room
SGM	Spatial Grid Matching
SMW	Sinewave Modulated White Noise
SNN	Spiking Neural Network
SNN/LP	Spiking Neural Network/Limited Precision
SNR	Signal to Noise Ratio
SOC	Superior Olivary Complex
SRM	Spike Response Model
SSL	Sound Source Localization
STDP	Spike-Timing Dependent Plasticity
STRF	Spectro-Temporal Receptive Field
SVM	Support Vector Machine
TDE	Time Delay Estimation
TDOA	Time Delay of Arrival
TOA	Time of Arrival
UWN	Uniform White Noise

CHAPTER 1
INTRODUCTION

1.1 Binaural source localisation

Binaural source localisation has attracted increasing attention in recent years over a broad range of applications (Talagala et al. 2014, Andéol et al. 2013). The auditory systems of humans and many other animals are capable of localising sound sources to survive in environments relying on their sense organs, i.e. ears and the information processing power of the brain (Talagala et al. 2014, Jindong et al. 2008). Ears are highly sophisticated: through their complex directivity patterns, information about sources from various locations is encoded in the signals from the two ears. This feature enables 3D localisation from only two channels. Three-dimensional source localisation has many critical applications. Arguably, advanced 3D spatial audio would need no more and no less information than the binaural signals in a listener's position. This can be viewed as a sufficient and necessary condition for human perception of 3D sounds (Ziegelwanger et al. 2015a, So et al. 2006). For robotics and security systems, sound source localisation can assist in sensing of specific events, taking advantage of the fact that the sources do not require straight line of sight, and sensing is not restricted to operational camera angles. For example, a domestic robot might be able to hear what is happening in the next room by perceiving the sound transmitted through the wall, whereas cameras do not have such an ability (Murray et al. 2004, Valin et al. 2003).

Source localisation is also an important and sometimes integrated step for source separation and signal cleaning (Taddese 2006). Many efforts to localise sources accurately are based on the use of large multi-channel arrays, e.g. (Wang and Kaveh 1985, Pavlidi et al. 2012). These Methods have many limitations. The sensitivity and accuracy are dependent on the size of the arrays and the number of microphones used. Logistical constraints can prohibit the use of large arrays in certain situations, e.g. if a home care robot is to adopt a 1-metre circle array, it would make it difficult for the robot to deliver its expected function. Calibration and channel matching for large arrays are particularly burdensome tasks. Multichannel signal processing is also not straightforward. Inspired by the binaural hearing of humans and animals, the use of a dummy head or two microphones with fine-tuned directivities and advanced signal processing

techniques to achieve source localisation has been proposed by some authors, who have all achieved promising results (Woodruff and Wang 2012).

Spiking neural networks, which deploy third-generation neurone models, behave similarly to real neurones in the brain and been used by neuroscientists to study and emulate lower level brain functions(Baladhandapani and Nachimuthu, 2015). These models have been particularly successful in the study of binaural source localisation of animal brains (Goodman and Brette 2011). Spiking neurone models have a built-in ability to handle time delays, which is a key feature of third generation models compared with previous generations (Yu et al. 2016, Diaz et al. 2016). This feature is essential in sound localisation prediction, as much of the information is encoded in the interaural time difference and interaural phase shifts of different frequency components from a given incidence angle.

This work attempts to explore the suitability of the spiking neural network model as signal processing engine to resolve source locations from binaural signals. The experiments result of this model appeared high performance for SNN model in analysing the binaural information and detecting the sound source with a variety range of sound signals (speech, noise and tones). However, the model showed a weak performance to localize two sound sources and separate between them. Deep neural network was applied to manipulate the SNN frequency-timing outputs features. This novel combination from two neural learning levels provide an important idea to solve multisource localization problem.

A training dataset is generated. The HRTF datasets, which have different azimuth and elevation angles, were convolved with different instances of speech sample (500 ms duration). The response of a spiking neural network (embedded with the same HRTF dataset) to each of these white speech bursts is analysed and the firing rate of each coincidence-neuron calculated. The generated data from different spiking output points was used an input feature to a deep neural network. This network was trained to localise multiple sources simultaneously.

Our robustness Localization model benefit from the powerful computation features of two advanced machine learning networks; spiking neural networks and deep neural networks. This integration presents as a multi-dimensional processing unit start from processing the binaural inputs by employing the temporal features of spiking neural networks to generate millions of

output points that represent the firing rate of coincidence detection of spiking neurons. Therefore, spiking neurons firing rate that included all the spatial information of input sound will be the raw input for novel structure of deep neural networks. Then, the deep neural model will have trained for learning from recurrence occurrences to detect the patterns similarity in its raw inputs features to analyse the binaural information in it and separate and predict its compounds. The model is giving a broadly chance to investigate the correspondence between the spiking neural model as unsupervised method and the deep learning of deep neural network as supervised algorithm. The solidity of this intelligent combination summarizes by its effectiveness in solve all the challenges concern with multisource sound localization.

1.2 Review of State of the Art

Research studies in sound source localization have been carried out for more than five decades, with steady interest in a variety of methodologies (Knapp & Carter, 1976, Ward et al., 1998). The major motivation of such research was to investigate the human hearing mechanisms and to try to mimic the human ability to localize different sound sources by using only two sensors, the ears (Goodman and Britte 2010, May et al. 2011, Roman & Wang 2008). A number of techniques achieve high localisation accuracy in the presence of environmental noise using only two sensors(May et al.2011, Roman & Wang 2008). The main development in binaural sound source localization is relatedwith the significant development of information technology and computing power and, in particular, machine learning systems applied to signal processing and localization. One of the most significant research studies in single sound source localization has combined binaural hearing and spiking neural networks to analyse a binaural signal and identify its location (Goodman and Britte 2011). Furthermore, advent of machine learning methods (Chen and Ser 2009), neural networks (Sun et al. 2018) and deep neural networks (Yalta and Ogata 2017) have been applied to solve sound source localization challenges.

1.3 Sound Localization Challenges

Despite the increasing research into spatial hearing and sound source localization, there are many limitations and challenges in the understanding of the neural representation of the auditory processing of mammalian brains. In spatial hearing, developing a realistic model of the human auditory system’s source localisation is a significant achievement. One of its features is the ability to specify the dimensions and characteristics of any bounded space (e.g. room) by using

acoustical cues. Sound source localization can contribute to the understanding of the cocktail party effect, whereby a listener is able to distinguish between voices in a crowded listening space (Grothe et al. 2010). Generally, the more considerable challenges in sound localization fields can be classified into three main categories:

1. Two sensors for hearing: Various methods have been applied to localize sound sources based on correlation analysis, beamforming, and signal subspace techniques, where sensor arrays are capable of source localisation in free space (Knapp and Carter 1976, Ward et al. 1998, Valin et al. , 2003). The challenge, inspired by the binaural hearing of human and animals, is to build a more realistic model that emulates human sound source localization by using only two sensors (ears). Most of these methods achieved an important level of accurecy in predicting one sound source from binaural signal. A greater challenge is to carry out accurate multisource localization by using only binaural information.
2. Non-indivituality (HRTFs Mismatch): this refers to the variation between HRTFs that are measured under different conditions and with different subjects. HRTFs play a considerable role in sound source localization tasks, and their characteristics are highly individual and related to the geometry of the head, pinna, and torso (Parseihian and Katz 2012). Non-individuality can increase the localization error because when the actual HRTF do not matched the those used to train the system (Wenzel et al. 1993, MENDONÇA et al. 2014). Spatial sound applications have broadly depended on non-individual HRTFs. Therefore, the challenge here is to find generic model that can work with multiple HRTF data sets
3. Noisy environments: One of basic challenges related to sound source localization is any undesirable change in signal-to-noise ratio due to environmental background noise. This kind of variation could have a significant effect on localization model performance and lead to a decrease in accuracy. Environmental noise is one of the main challenges in spatial hearing and sound source localization especially when dealing with complex sound signals (e.g. multiple talkers). Increased background noise is detrimental to signal detection and recognition (Recio-Spinoso and Cooper 2013). Furthermore, many questions have been raised to describe the relationship between the cochlea and noisy signals (Recio-Spinoso et al. 2009); cochlear processing depends on the response of the basilar membrane, which handles noisy signals in ways that are still not clear. The effect of background noise on

neuronal coding of the interaural level difference ILD and on the sound source localization performances is further discussed by (Mokri et al. 2015).

1.4 Research Motivation

Sound source localization using only two ears brings many challenges. One challenge is (cocktail party) which refer to complex situation when sound signals such as speech from different speakers in various positions at the same time (Macpherson and Middlebrooks 2002). From the human standpoint, we find it difficult when there are multiple sources to locate them (Drullman and Bronkhorst 2000). Moreover, the challenge of separation of sources with two sensors, that linear separation is limited by number of sensors and sources (Shoko et al. 2007). Another key issue is environmental noise. In this work a multisource sound localization model is developed using the signals captured from the two microphones on a head and torso simulator as inputs. The performance of the model is investigated with diverse types of sound signals and HRTFs, and in various background noise conditions.

1.5 **The Aims of the Research**

The main aim of this study is developing an automatic localization model for multisource localization. In other words, the aim is to investigate the capabilities of the integrated Spiking neural networks SNN with Deep The main aim of this study is developing an automatic localization model for multisource localization. In other words, the aim is to investigate the capabilities of the integrated Spiking neural networks SNN with Deep neural network DNN in solving the multisource localization challenges. “Can SNN be effectively used as a features extraction method and the DNN as multiclass classier to processing the binaural signals in order to estimate the directions of two sound signals that emitted from two different locations at the same time?” is the research question.

1.6 **The Objectives of the Research**

There are many limitations and challenges related to simulating abiological neural network and emulating its ability in executing a various of synchronous functions in an important level of accuracy. To address this challenge, a spiking neural network is developed with the following objectives:

- Perform reviewing for the related works and background study.
- Generate and prepare the appropriate sets of data to carry out the experimental study of the proposed localization models in this thesis.
- Explore models that emulate behaviour of both ears and brain to localisesound emitted from multiple sources with no prior knowledge of the scene.
- Investigate the localisation accuracy of the model with a range of diverse types of sound signals (speech, white noise, tones, tone modulated white noise).
- Test the impact of noise on localisation performance.
- Examine the importance of different binaural cues (interaural time differences (ITD), interaural level differences (ILD) and spectral cues) on localization performance
- Compare the suitable of spiking neural networks and traditional neural networks for binaural sound localisation.
- Improve the localization model by applying a deep learning mechanism with spiking neural networks as a method for features extraction to construct a more realistic neural model, with the potential for addressing known issues in spatial hearing and sound perception for robotic sound localisation and engineering applications.

1.7 Research Methodology

The main steps of the methodology are shown in Figure 1.1.

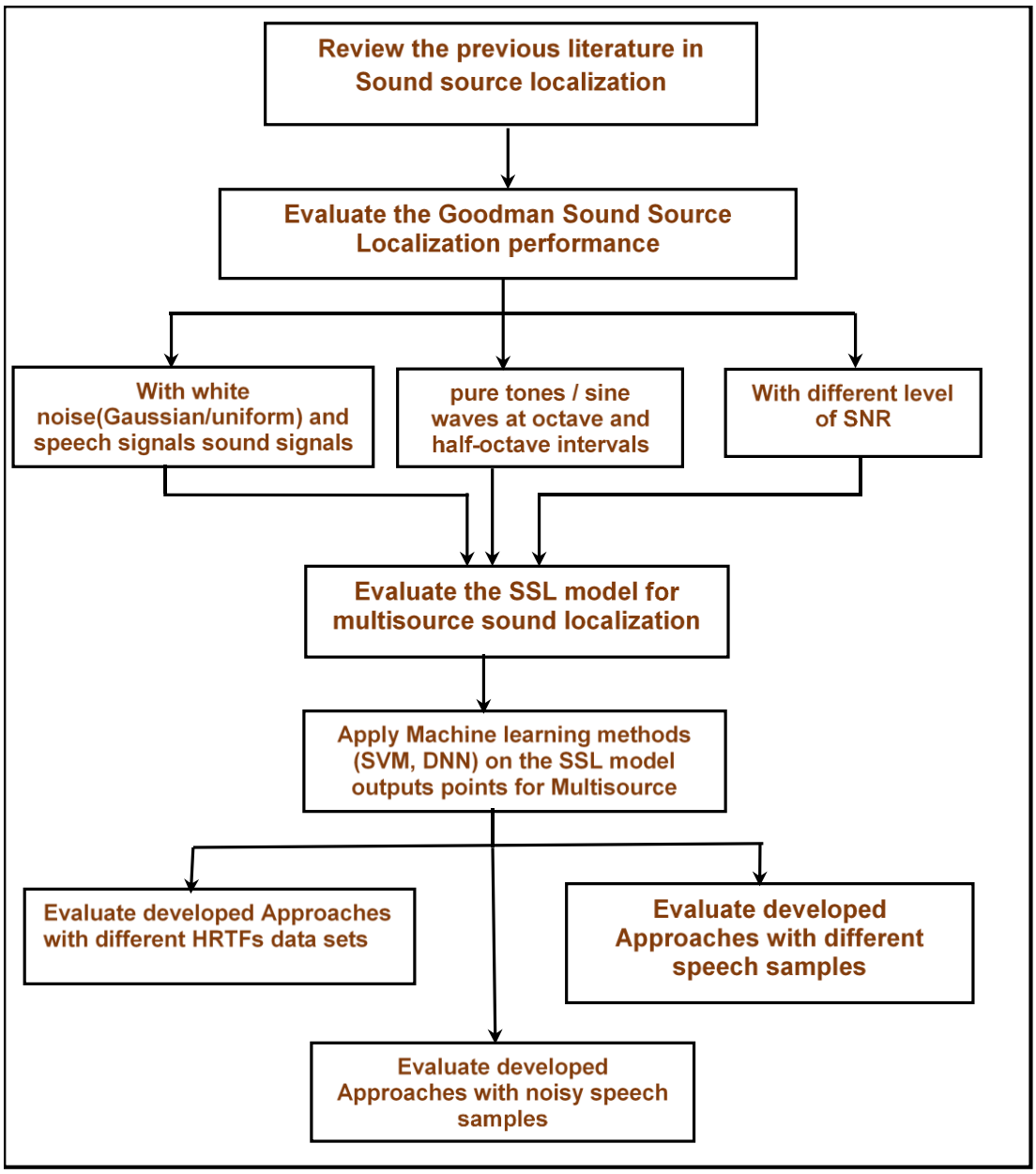


Figure 1.1: Research Methodology.

1.8 Contribution of the Study

The main contribution of this study is a robust approach to multisource localisation from binaural data. Previous methodologies have performed accurate source localisation using binarual data for single sources, the main contibution of this work is the development of a

binarual multiple source localisation method. This has been done by implementing a combination embedding HRTF filters as binaural information filters in a spiking neural networks SNN. The spiking neural network works as an unsupervised algorithm to analyse the spatiotemporal information associated with binaural input signals. Deep neural networks are then taught with supervision to detect patterns in SNN firing rates. The model was investigated using two distinct HRTF data sets (KEMAR HRTFs, IRCAM Listen HRTFs), and in the presence of background noise of varying SNR.

1. Single sound source localization model from earlier work has been replicated and motivated to examine the localization model behaviour with HRTFs of KEMAR dummy head that simulated human head and torso. Then all the outcomes from applying HRTF of KEMAR were compared with the original HRTF data set (IRCAM) of the replicated model which represents a human subject. Each one of these data has special impact on the localization model performance due to the differences in the anatomical parameters (head size, ear shape and torso). Fundamentally, both HRTFs data sets have unique dimensionality characteristic that reinforce the localization model testing by providing a wide range of azimuth and elevation angles.
2. The localization performance with real speech samples has been investigated with a variety of sound durations to investigate the sound signal time duration on localization performance. In contrast, the model is examined with other sound signals forms as likes Uniform white noise, Gaussian white noise and sine wave modulated white noise for evaluation and comparison purposes.
3. Single frequency and octave frequency are investigated accurately to evaluate the localisation reliability in different frequency ranges.
4. The model examines the sound with a various signal to noise ratios of active background noise. The experiments involved generate white noise signal in different SNR level and then add them to the binaural signal to simulate the real-time noise environments.
5. A novel method that combines HRTF data with spiking neural networks and deep neural networks is presented to carry out multisource localization. this novel idea based on applying the spiking neural networks as a pre-processing for features extraction from

binaural data to generate different firing rate outputs. The firing rates were used to train, test and validate the deep neural networks.

1.9 Thesis Outline

The thesis is organised into ten chapters. A brief description of each chapter is given below:

Chapter 2: Literature Review

This chapter starts with a description of the process of human sound source localization. It then presents a review of the literature around binaural sound source localization. This includes a review of different methods in the fields of sound source localization and binaural hearing. A critical evaluation of current techniques and the state of the art of these approaches are presented. Furthermore, approaches that can enhance the performance of localization models to solve multisource sound signal localization are presented. This will include any work on multisource localisation carried out by extracting features from binaural cues (as seen in the next chapter).

Chapter 3: Background and Theory

This chapter describes the fundamental concepts of binaural hearing and Head-related transfer functions. Machine learning algorithms relevant to this research are explained in this chapter, including Spiking neural networks and deep neural networks. Finally, the HRTF and speech databases used in this study are described.

Chapter 4: Single Sound Source Localization model (SSL)

This chapter is a presentation of the single sound source localisation model (SSL). The model structure, the shape of inputs and the pre-processing are explained. Gammatone filter banks and the neural transformations that form the spike train – the spiking neuron model input -output are explained. All experiments and test outcomes of the single sound source model are displayed in this chapter. Various types of input sound signals (Gaussian white noise, Uniform white noise, sine wave modulated white noise and real speech samples) are demonstrated in this

chapter. Single and octave frequency inputs are tested, followed by an experiment to examine the effect of different SNRs on the performance of the model.

Chapter 5: Multisource Localization Model

This chapter describes the required modifications to the model to carry out multisource sound signal localization. The binaural signal mixing process is explained, showing the data from different speech samples used to train and test the multisource localisation model. To decrease the computation complexity and memory cost, the frequency features of impute row data are reduces as explained in this chapter. The results of tests with clean and noisy speech samples are discussed. Furthermore, comparisons between different HRTFs databases are shown. Finally, the results of the new model are compared with the SNN method. The localization performance with background noise conditions and directional noise cases.

Chapter 6: Non-individual HRTFs Localization.

This chapter details the individuality characteristic associated with each HRTF database, followed by a display of the state-of-art in this this area. The proposed model is trained by using data that generated from one HRTFs (IRCAM Listen HRTFs data set) and tested with data that generated from another HRTF (KEMAR dummy head). The model localization behaviour for localizing one and two sound sources with these mismatch HRTFs for training and testing stage are explained. Moreover, the potential approaches towards a generic model that would work with different HRTFs also demonstrate in this chapter. The outcomes are visualized and compared with the outcomes of SNN

Chapter 7: Conclusions and Future Works

This chapter presents the conclusions of this research and provides suggestions for future work. The experiments findings and its conclusion that raising from applying the localization model to solve the multisource problems are explained in this chapter. Number of improvements are suggested to overcome the localization drawbacks and enhanced the localization accuracy.

CHAPTER 2
LITERATURE REVIEW

Chapter Overview

In this chapter, a background survey of the general framework of human hearing system is covered. In addition, a literature related to different approaches that have been applied in the fields of sound localization and binaural hearing. Furthermore, the most important techniques used to enhance the performance of sound signal localization in different environments are also detailed. The chapter starts by illustrating the mechanism of human hearing. Details about binaural hearing are given in section 2.2, and review of the most conventional methods for sound source localization is given in section 2.3. Section 2.4 reviews machine learning and neural networks. The discussion in Section 2.5 is concerned with sound source localization modelling and classification approaches that are applied to different machine learning models. Finally, the state-of-the-art multisource localization methods are reviewed in section 2.6.

2.1 Human Sound Conduction Mechanisms

To clarify the mechanism of human hearing, we should first understand the anatomy of the human auditory system. Generally, hearing can be defined as the process that is performed by the peripheral auditory organs (outer ear, middle ear and inner ear) that converts sound waves into electrical pulses. These electrical pulses are processed by the auditory nervous system (Alberti 2001, Zemlin1968).

The peripheral auditory organs consists of three main components as shown in figure 2.1:

1. The outer ear: the outer ear consists of two main elements: The pinna and auditory canal. The pinna has an ovoid-shaped structure. Humans have two pinnae, each one has an individual structural shape. Pinnae act as a filter that works to collect the sound signals to the ear canal helping with sound source localization. The auditory canal is an auditory tube terminated by the tympanic membrane (eardrum). Its main function is to transmit sound waves to the eardrum and acts to increase the ear's sensitivity, due to its resonance, between 3000 Hz to 4000 Hz. There are many factors which influence the sound intensity in the ear canal. One of these factors is the direction of the sound. Another is shoulder reflection and the acoustical shadow caused by listeners' head and pinna filtering effects. Head, shoulder and pinna influences are increased when their size is close to the sound's wavelength (Alberti 2001).
2. The middle ear: the main part of the middle ear is the tympanic membrane. It has cone shape structure which vibrates in response to the received sound signal. The middle ear changes the pressure changes of sound waves from the auditory canal in to mechanical vibrations. The structure of the middle ear consists of three small bones in the middle ear cavity named; the malleus, incus, and stapes. These three bones compose a ossicular which is a chain connects the tympanic membrane with the oval window. The core function of the middle ear is to transmit these slight changes of the tympanic membrane movements that caused by the auditory pressure on its external side to the inner ear (Maroonroge et al. 2000).
3. The inner ear or cochlear is so called as it has a snail shell shape and consists of two main parts, the scala tympani on the bottom and the scala vestibule on the top, separated by the basilar membrane. The cochlea converts the pressure fluctuations into nerve

impulses that are coded in such a way as to be processed by the brain. The volume of cochlea is around 0.2 ml filled with around 30,000 hearing cells which transform vibrations into neural impulses. Nerve fibres exchange the signals from the hearing cells to and from brain (Alberti 2001). Each auditory nerve fibre responds to a different band of frequencies and sound pressures. Usually, the rate of neuron impulses that are transmitted to the brain are dependent on the presented sound's intensity and frequency (Maroonroge et al. 2000, Kirk and Gosselin 2009).

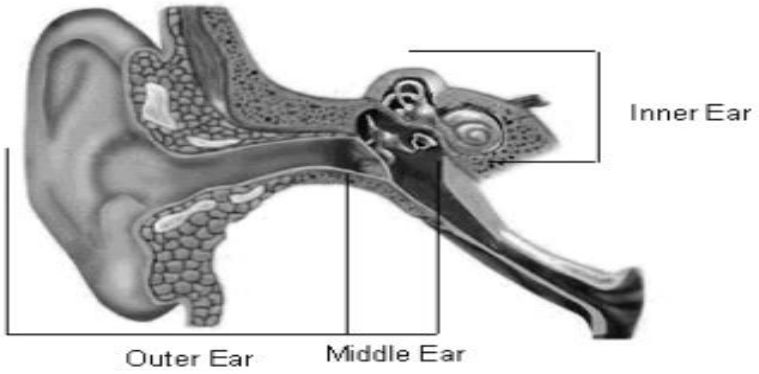


Figure 2.1: Human ear's overall structure explains the outer, middle, and inner ear (Maroonroge et al. 2000).

The main method of sound localization in binaural hearing is coincidence detection. Regardless of whether the input sound signal is plain sinusoidal wave or a more complex sound signal form such as a mixture of voices at cocktail party, the input to the ear are just vibrations at each eardrum. The brain analyses and compares the individual response of each eardrum and then extracts the related localization cues to estimate the location. The hearing cells of the basilar membrane are regularly organized by the frequency of a sound rather than sound spatial location or any other specific characteristics of sound source. For that reason, auditory space representation is done in the central auditory system by converging sound received by the two ears onto a single neuron inside the brain where the physical parameters of sound with its temporal features are analysed deeply and accurately (Grothe et al. 2010).

2.2 Review the Spatial hearing and localization cues

Humans have an extremely complex hearing system. It can identify and locate sounds with remarkable accuracy in azimuth, elevation, and even distance. This function can be performed even using single ear. Psychoacoustic studies demonstrated that the source localisation process

depended on four types of acoustic cues (Blauert 1997). Binaural cues caused by the differentiation of signals from both ears have a significant role in the sound localisation process; these cues include, Interaural Time Difference (ITD), Interaural Level Difference (ILD), spectral cues, and dynamic cues. ITD refers to the first type of localization cue which is brought on by the propagation delay between the ears. ITD is a primary localisation cue of low-frequency signals; below 1.5kHz. ILD is brought on by the head shadowing. ILD is essential for the localising higher frequencies; above 3kHz (Agterberg et al. 2012). Interaural differences (ITD, ILD) are important to localize sound sources in horizontal plane and lateral dimensions (left-right discriminations) (Macpherson and Middlebrooks 2002). Sound scattering, and shadowing are altered by the listener's head dimension. ITD and ILD are often defined by analysis of these changes (Algazi et al. 2001, Kuhn 1977, Ziegelwanger and Majdak 2014). ILD can contribute to localisation separately from ITD, particularly at higher frequencies where the wavelength is small compared to head diameter, producing ambiguous ITD information (May et al. 2011).

The third type are spectral cues, these are brought by reflections and interactions of sound waves from any obstacles including the pinna, the head and torso (Xie 2013). Spectral cues are useful at mid-frequencies (between the low and high-frequency ranges) (Nimityongskul and Kammer 2009). Spectral cues are relevant to the localization in vertical plane and sound source front-back differentiations. The spectral localization cues are primarily described by analysis of the listener's pinnae geometry (Bronkhorst1995, Hebrank and Wright 2005). Dynamic cues are brought on by the relative motion of the ears and the source. Listeners can move their heads, and some animals even move their ears to seek confirmation or better resolution in source localisation (Zhong and Xie 2014).

The literature further discriminates between two classes of hearing cues; binaural cues and monaural cues. Individually, binaural cues are useful only in narrow bandwidths and emphasise sound localisation in the horizontal plane. Monaural cues represent the signal which received by one ear (Ahveninen et al. 2014). Spectral cues are often monaural, the filtering effect of the pinna and head are angle dependant and can be particularly helpful in assessing height (Grothe et al. 2010). Figure 2.2 shows the monaural and binaural hearing cues for sound localization.

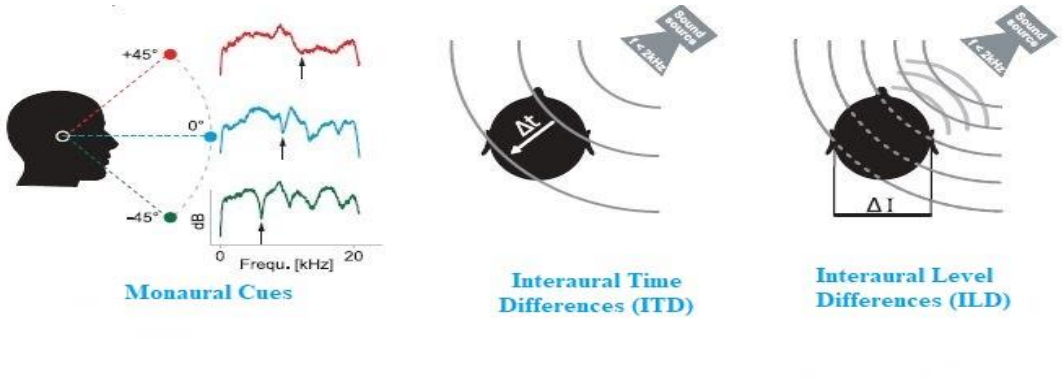


Figure 2.2: Cues for sound localization (Grothe et al. 2010)

Using only two sensors (ears), human and animals succeed in localising diverse types of sound sources. Anatomical parameters such as head and ear shapes play significant role in processing the incoming sound and helping to locate the origin of the sound. The effect of the head on the sound at the ear is captured in the head-related transfer function (HRTF). HRTFs are direction-dependent filters that can characterize the received sound at the outer ear affected by sound scattering and reflections resulting from the head, pinna and torso (Møller et al. 1995, Wightman and Kistler 1989). HRTF contains a filter (impulse response/transfer function) for each angle. Head-related transfer functions (HRTF) pick up transformations of a sound wave propagating from the source to our ears. The transformations contain the diffraction and reflections of the head, pinnae, shoulders and torso. So, the HRTF or HRIR filters capable to create the illusion of spatially located sound (Groethe el at. 2010). HRTFs capture listener-specific cues, including ITD, ILD and spectral cues. Due to the different individuals' dimensions HRTFs are unique to each individual, as are the cues.

2.3 Review of sound source localization methods

Methods that use reduced sensor arrays (two sensors) offer several techniques to achieve high accuracy localisation. These methods use a head and torso simulator where the anatomical parameters, and thus the binaural cues, are known.

These methods have joined azimuth-based models of ITD and ILD. The correlation of ITD and ILD with source azimuth location is a complicated pattern (May et al. 2011). However, because the frequency-based pattern of ITDs and ILDs can change across individuals, an azimuth-based model requires pre-processing or standardisation with the binaural signals and may reduce performance over various binaural setups. Furthermore, strategies differ in the method of

integration of interaural information over time and frequency. Statistical tracking methodologies can be used to integrate localisation estimates over time. However, binaural strategies have concentrated on cases with lower levels of background noise (Mayetal.2011, Roman and Wang2008). When the HRTF is known at each possible location, the localisation process attempts to apply the inverse matching of the observed localisation cues to a source situation (Roman and Wang 2008).

Various methods depend on correlation analysis, beamforming, and signal subspace techniques which apply microphone arrays to source localisation in free space (Knapp and Carter 1976, Ward et al. 1998). The localization accuracy depends on increasing the number of microphones in the sensor array (Salvati 2012). Recently, a method using a 10-element microphone array, consisting of two sub-arrays, was proposed for 3D sound signal localization. The method computes the time delay of arrival (TDOA); each node in the sensors array receives the sound signal and instantly returns an absolute timestamp prior to passing it to the processing unit (Song et al. 2017).

However, the experimental outcomes showed that an accurate estimation of elevation angle value is possible by using the all embedded components of HRTF. When the ITD and ILD localisation cues are integrated with the spectral cues for a better localization resolution. To achieve accurate binaural localisation in predicting both azimuth and elevation values, the HRTF needs to be employed with all embedded binaural cues (Rakerd et al. 1999, Best et al.2005). The effects of sound signal duration and level on the localization accuracy in the vertical plane (elevation) and horizontal plane (azimuth) are discussed by Ruhland Gai et al. (2013). This study offered an experiment on cats to examine the impact of sound signal level and duration on azimuth and elevation localization performance. It showed that any alteration in the sound spectrum can cause significant effect on the elevation localization performance. Therefore, an increasing sound duration caused notable enhancing in localization accuracy in elevation. In contrast, in horizontal plane (azimuth prediction), neither sound duration nor level had an observable influence on localization accuracy, excluding at near-threshold levels (Inoue 2001, Macpherson and Middlebrooks 2000).

One notable limitation of the previously proposed methods is the computational overhead that resulted from using TDOA which increases with the size of microphone array, whereas the

reliability of TDOA estimation requires considerable number of microphones (Nesta and Omologo 2012, Heilmann et al. 2014, Song et al. 2017). A further limitation is source localization accuracy and confusion. For example, a sound source located at any region of the head that results in signals have the same interaural time differences (ITD), known as a "cone of confusion" which basically results from the spherical appearance of human head (Kapralos et al. 2008). The cone of confusion phenomenon could lead to poor localization performance in distinguishing sound sources in the vertical plane when the wavelength of incoming signal is equal to the head diameter (the distance between two ears) in this case both ITD and ILD have zero values in the median plane (Miller 2013).

2.4 The Artificial Intelligence and Machine learning

Artificial intelligence refers to the research that focuses on studying the human brain's capabilities in thinking, learning, problem solving and making decisions to imitate human intelligent behaviour. Artificial intelligence characterizes a machine's ability in mimicking and performing intelligent human capabilities. Artificial intelligence includes different types of approaches that perform different tasks in a variety of sectors, for example, methods based on statistics, artificial neural networks, computational intelligence and probability.

2.4.1. Traditional neural networks

The modern concept of neural networks started in the 1940s with the work of Warren McCulloch and Walter Pitts (McCulloch and Pitts 1943) which represent the first generation of neural networks. They proved that artificial neurones could calculate any arithmetic or logical task (Hagan et al. 1996). Donald Hebb Donald was followed MacCulloch and Pitt, they proposed the earlier technique for learning in artificial neurones (Hebb 1949). After few years of active research, Stephen Grossberg investigated the self-organizing neural networks (Grossberg 1976). Two new key notions related to the artificial neural networks and their application were presented in 1980. The first conception was the utilising of statistical mechanics to clarify the operations in the recurrent networks (Hopfield 1982). The second concept presented when the researchers discovered the backpropagation algorithm for training multilayer perceptron MLP networks. This algorithm was the solution to problem-solving limitations in the earlier neural perceptron algorithms. The most popular back-propagation

network represents the second generation of neural networks and the significant supervised model for many engineering applications.

The significant availability of powerful modern computers and new hardware development enables the processing capabilities to test the latest ideas related to advanced neural networks such as ingenious architectures and training rules. Neural networks, also called parallel distributed processing, have occupied a regular place as very essential and important scientific and engineering tools to be used in appropriate situations. However, there is still la lack of knowledge about biological systems, and all current artificial neural models represent an over simplification in representation the biological neurone models (Kröse et al., 1993). Artificial neural networks researched for many years in the prospect to realise a human-like performance to solving complex functions related to the human perception (Lippmann 1987). There are some attempts to investigate the efficiency of applying the integration between classical neural models and HRTFs to perform localisation tasks in animals (Shimoyama2012). Song et al. (2017) proposed a model to stratify traditional neural networks to extract the ITD and ILD from incoming signal to estimate its sound direction. They used a TDOA method with microphone array with 10 sensors. The findings demonstrated that the suggested method could learn to localise a sound source in the anechoic and reverberant conditions only when the incoming signal was white noise. Youssef et al. (2012) Presented an artificial neural model to predict the azimuth and elevation angle of a sound source. The experiment results showed that the model could estimate azimuth and elevation with limitation for complex signals such as a human speech signals in noisy environments.

2.4.2. Spiking Neural Networks (SNNs)

The term Spiking Neuron Networks (SNNs) simulates the behaviour of natural neurones, highly inspired from computations which are performed naturally in the brain based on recent developments in neurosciences. Spiking neural networks deploy third-generation neurone models and represent a relatively important level of similarity to real neurones in the brain. SNNs are well matched to approximating lower level perceptual functions (Baladhandapani and Nachimuthu 2015, Markowska and Koldowski 2015). SNNs can process and account for time delays in signals; a key feature of third generation models when compared with previous approaches traditional methods (Yu et al. 2016, Diaz et al. 2016). Second-generation methods

of neural networks are represented by threshold and sigmoidal techniques as computational units. The major challenge is to promote effective learning rules that might pick characteristics of the specific features of SNNs while preserving the good aspects of traditionally correlated models. Nevertheless, most of traditional neural networks have many difficulties when dealing with the enormous amounts of data and adjusting to fast changing environments. In addition, there are certain limitations related to refined learning algorithms or artificial neuron models compared to biological processing in natural nervous systems (Paugam-Moisy and Bohte 2012). The major challenge is to develop effective learning rules to overcome the peculiarities of SNNs while preserving the right features of traditionally correlated models. Nevertheless, conventional neural networks have many difficulties when dealing with the large values of data and adjusting to a rapidly changing environment. There are limitations related to refined learning algorithms or neurone model designed artificially compared with biological processing in the natural nervous system (Paugam-Moisy and Bohte 2012). These limitations can be summarized by the availability for a well labelled data to train the machine learning models. a sufficient training data should be available to provide suitable learning patterns for the learner. In addition, the huge data size required strong processing units as like high speed computer with graphical processing unit GPU and big storage capabilities. Networks of spiking neurones provided a more realistic representation of human cellular networks compared with traditional artificial neural networks. Spiking neural networks consist of the neurones that transmit shortened signs (impulses) which are called spikes. The computational units in spiking neurones are composed of three steps: summation of all neurone input stimuli, integration over time, and a spike fire when the membrane potential expands over the threshold which then returns to reset value (Davies 2013).

Inputs and outputs in first-generation neural networks were represented as binary signals [0, 1] and the processing unit inside the neuron represents a fixed threshold value. The computational unit in the 2nd generation artificial neural networks can be summarized as follow: sum all values of synaptic weights, compute the neuron's output signals when the summed amount exceeds the threshold value. The continuous activation function of 2^{nd} generation neural networks makes it convenient in processing analogue input and output stimuli. It accepts any real numbers as an input for this type of neurones, and the output is limited to any number between 0 and 1 (Basheer and Hajmeer 2000). 2^{nd} generation neural networks do not

use individual pulses; the output signals can be representing as normalised firing rates of the neuron through specific period which called rate coding (Gerstner and Kistler 2002, Vreeken 2002). Figure 2.3 explains the essential comparison between SNN and the earlier two types of traditional neural networks.

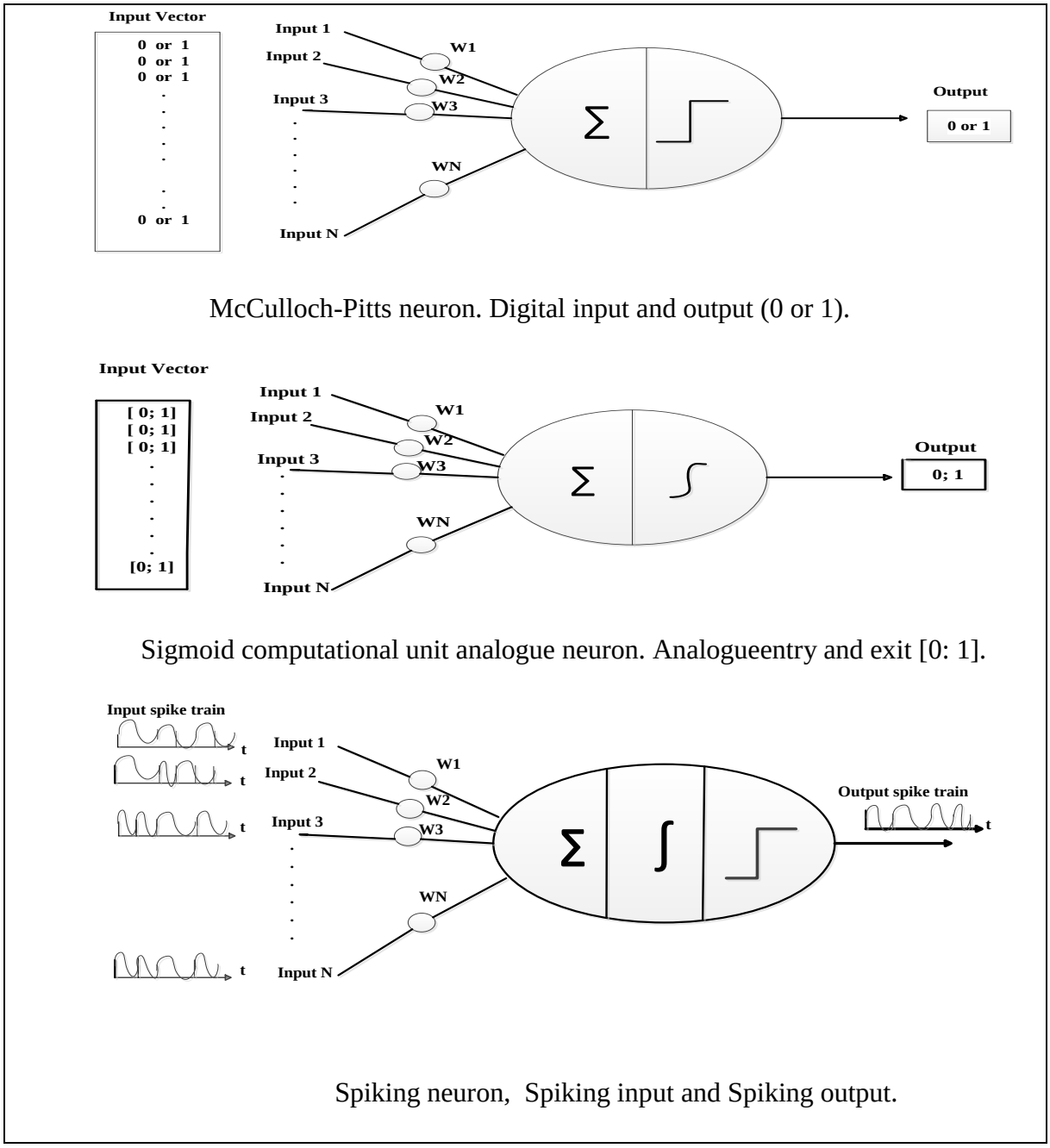


Figure 2.3: Comparison among the three generations of neural networks, type of input, output and the computation types of activation functions for each type.

Spiking neural networks increase the realistic level of artificial neural networks by using individual spikes in processing the temporary input (time delays in signals). This feature in spiking neuron models permits integrating spatial-temporal data in connection and computation, like the action of real neurons (Ferster and Spruston 1995). Spiking neural networks utilise pulse coding rather than rate coding as in traditional neural networks. According to this technique, the neurons in spiking networks receive and send individual pulses which allows for the processing of multiple data at the same time. For example, in the case of sound processing, frequency or amplitude (Vreeken 2002).

Spiking neuron models can be divided into two broad categories based on their level of abstraction. The conductance models and the threshold models, as shown in Figure 2.4. The action potential in the conduction models rises from the continuous dynamics, therefore in simulation the time step has to be small. Whereas, the threshold models use explicit thresholding and resetting to generate action potential, which are algorithmic but efficient in simulation. This represents a key difference between conductance models and the threshold models. Hodgkin–Huxley HH refers to the one of the most important conductance models. It can reproduce all classes of neurons with a good accuracy regard to the shape of spike or complex firing activities compared to threshold models. And, conduction models represent the biologically relevant mathematical neuron models present a more realistic artificial neuron model. Its computation cost represents the mainly drawback for these types of spiking neurons. The major criteria to compare among spiking neuron models are biophysically meaningful and measurable parameters, and whether they can exhibit autonomous chaotic activity. One of the simplest threshold models of a SNN is the leaky integrator and fire (I&F). It can fire tonic spikes with a constant frequency. The resonate-and-fire model produced by (Izhikevich 2001) is a parallel of the I & F neuron and is more efficient. An alternative to the leaky I&F neuron is the quadratic I&F neuron, also known as the theta-neuron (Izhikevich 2004).

The Hodgkin–Huxley model is one of the most important models in computational neuroscience. Researchers denote all conductance-based models as being of the Hodgkin-Huxley-type (HH). Such models are paramount not only because their parameters are biophysically meaningful and measurable, but also because they permit work on research problems linked to synaptic integration, dendritic cable filtering, influences of dendritic

morphology, the interaction between ionic currents, and other matters related to single cell dynamics (Izhikevich 2004). The model is quite computationally expensive. Thus, one can use the Hodgkin–Huxley formalism only to mimic a small number of neurons or when simulation time is not critical.

A modification of the Hodgkin–Huxley model is presented by Izhikevich (Izhikevich 2003). The model gathers advantages of the biological plausibility of Hodgkin–Huxley-type dynamics and the computational qualification of integrate-and-fire neurons. The spike response model (SRM) was generated to decrease the four-dimensional Hodgkin-Huxley model into one equation. It has been proven that the SRM model can predict 90% of the Hodgkin-Huxley spike train correctly (Davies 2013). There is significant different between the spike response model and the leaky integrate-and-fire. A differential equation describes the membrane potential in the case of LIF model and it is voltage dependent. In contrast, response kernels are used to describe the membrane potential for SRM. HH refers to the conductance models whereas Integrator–and-Fire (I&F) with all its related models and SRM refer to the threshold models as shown in figure 2.4.

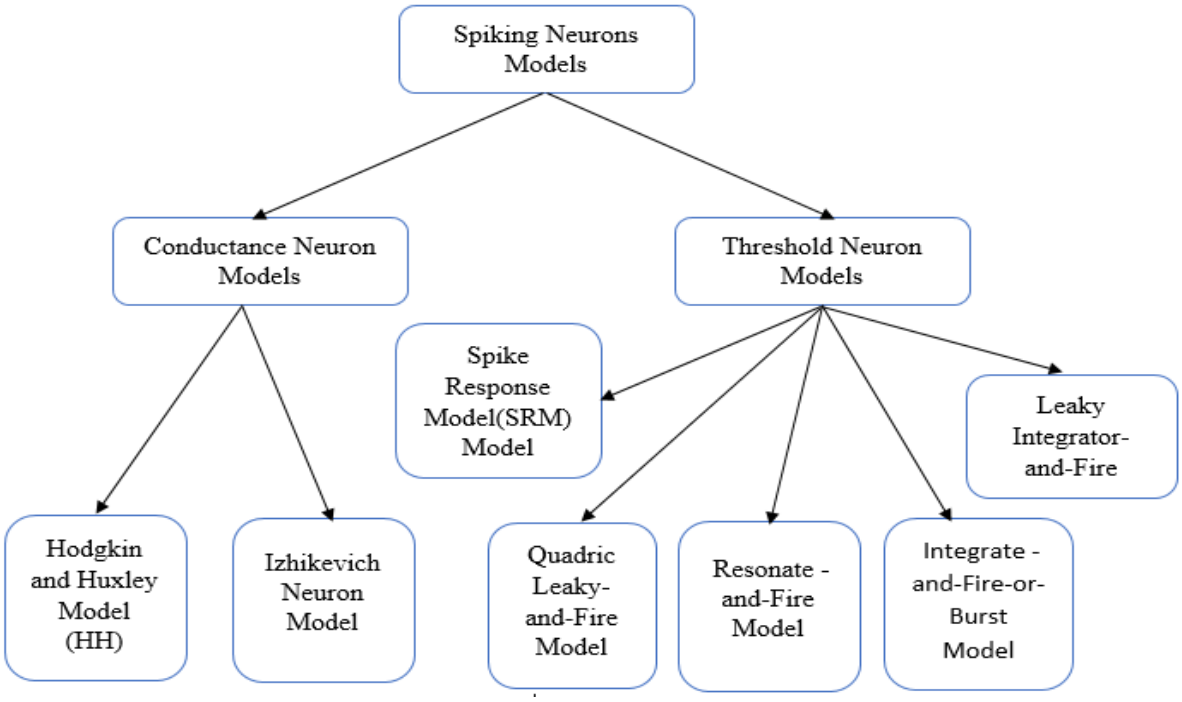


Figure 2.4: Spiking neuron models

2.4.3. Deep Neural Networks

Deep neural networks (DNNs) are an advanced type of artificial neural network (ANN). Deep neural networks have a structure which consists of input and output layers with many hidden layers arranged between them as shown in figure 2.5 (Schmidhuber 2015). The computational models that consist of multiple processing layers allow learning of representations of data by using deep learning. Deep learning applies the backpropagation algorithm on a massive dataset. It enables learning of local parameters and representations in each layer depending on received representations from prior one (LeCun et al. 2015). There are distinct types of deep neural networks architectures: Feedforward deep neural networks, recurrent deep neural networks, deep belief spiking neural networks and convolutional deep neural networks. The studies demonstrate that each type of deep neural structure has significant role to solve problem in specific application domain (Schmidhuber 2015).

Deep learning has a vital role in handling data analysis and learning problems where there is a large amount of input information. Existing artificial networks of spiking neurons still cannot compete with DNNs (Schmidhuber 2015). O'Connor (2012) explored a new model to realize the ability of the brain to build a correlated model of the world around it by executing a model of a recurrent network of spiking neurons with multi-modal integration. The network trains as a deep belief network (DBN) and the learned parameters are mapped onto a spiking neural network. The network trained on three types of stimulus; visual, audio and labels of ten digital groups. New methods rely on the Siegert approximation for integrate-and-fire neurons to apply an offline trained DBN onto an effective event-driven spiking neural network. This technique is proposed by O'Conner et al. (2015) and is appropriate for hardware implementations. The experiment outcomes demonstrate that the system could pick out a valid digit from other unclear inputs. Henderson et al. (2015) presented a new method integrated between SNN and deep learning. Neural network adapts according to the input training data. Neural network training using deep learning methods have been evolved successively to apply to several types of human applications; speech recognition, object detection and image classification (Deng and Platt 2014).

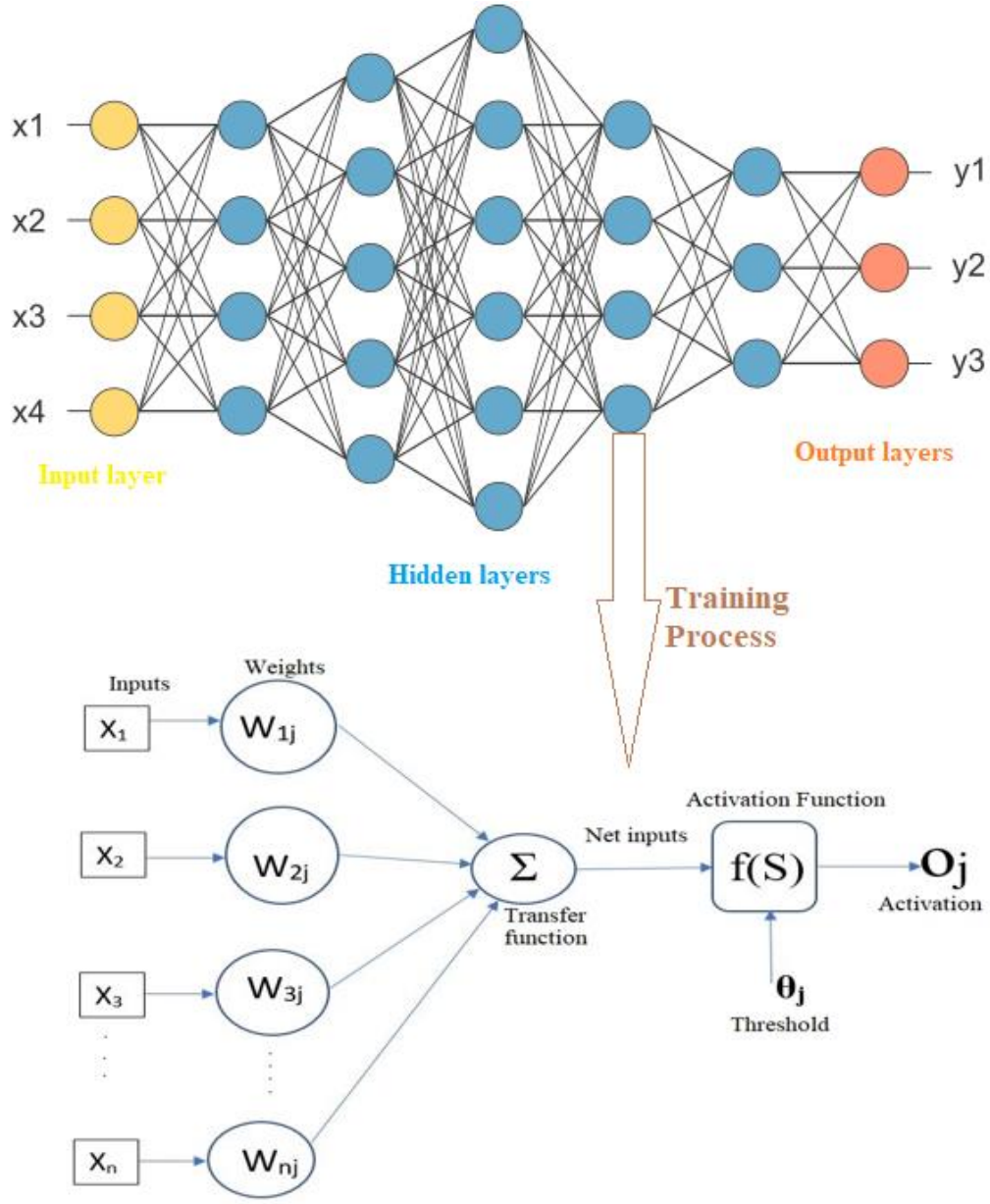


Figure 2.5: Deep neural network structure and network training process

2.4.4. Learning Methods in Neural Networks.

In artificial neural networks, there are several types of learning methods which work to build analytical models automatically without being explicitly programmed. These methods are known as the "learning paradigm" and can be divided into four kinds: supervised, unsupervised, reinforcement and evolutionary. The most common learning mechanisms are supervised and

unsupervised learning (Jürgen 2015). The "learning rule" refers to the algorithm through which the neural network adapts to train the input data (Pietila and Lim 2012).

In supervised learning, each input is labelled with a targeted value, which may be a numeric value (e.g. azimuth angle), or a classification (e.g. speech or music). The training process is carried out by minimizing the error between the ANN output and the desired output (the target). Recently, existing effective applications focused on supervised learning often take the form of pattern recognition competitions (Jürgen 2015, Graves and Jaitly 2014, Graves et al. 2013).

Kasinski and Ponulak (2006) reviewed and evaluated various supervised learning approaches to train spike timing of spiking neural networks. They analysed the major features of each method and its ability in learning spike timing accurately. Their work was based on the suggestion that functional brain computation fundamentally can rely on the precise timing of each spike. Many implementations of SNNs are imitating biological neural networks, as they provide a more accurate representation of realistic networks than traditional artificial neural networks. Also, many types of research have been done in applying SNN to temporal pattern recognition. Some ideas are investigating SNN applications in robotics, most of which are based on evolutionary algorithms to train the network (Bulanova et al. 2012). In 2004, a new supervised learning rule was inferred by Booij (2004) for Spiking Neural Networks (SNNs) involving the gradient descent technique, which can be implemented on networks with a multi-layered design. This algorithm is practically prepared to deal with neurons that fire multiple spikes.

A novel supervised learning algorithm is proposed by (Stromatias 2011) which relies on genetic algorithms. The proposed algorithm is eligible to train both synaptic weights and delay and permit each neuron to emanate many spikes, and so on, taking on the full characteristics of the spatial-temporal coding intensity of the spiking neurons. Also, limited synaptic precision is applied. Furthermore, the readily supervised training algorithms, for instance, SpikeProp and its modifications QuickProp and RProp permit their neurons to spike only once through the simulation time, thus not taking full advantage of the power of SNN. The supervised training algorithm is designed for limited precision feed-forward SNN (SNN/LP) is also proposed as a genetic algorithm and is applied using supervised training. One of the benefits of the GA is that they can adjust the patterns of the spike response model (SRM) (Stromatias and Marsland 2015).

The literature presents many deep neural network approaches that use supervised learning in various applications including speech recognition and sound source localization (Yalta et al. 2017, Takeda and Komatani 2016a). For more details, see section 2.5.1.

In unsupervised learning, the model parameters (weight and bias) are modifying by searching joint characteristics and pattern similarity among system training examples. The model parameters updated depend on internal knowledge that generated throughout the learning process (Baldi 2012). Jürgen (2015) reviewed in his paper most of important efforts that implement unsupervised learning and the various unsupervised methods, for example, auto-encoder hierarchies (Ballard 1987) and restricted Boltzmann machines (RBMs) (Smolensky 1986). Recently, many modern unsupervised deep learning applications have been presented. For example, Kingma and Welling in (2014) offered unsupervised learning framework called variationally autoencoder that was applied to train deep belief networks. Unsupervised learning also played a significant role in learning using deep neural networks.

Karhunen et al. (2015) reviewed most common unsupervised learning methods that have been applied to different machine learning frameworks and deep learning structures including multilayer perceptron networks and deep neural networks. Bengio et al (2014) suggested unsupervised learning for a deep learning structure referred to as a generative stochastic network. The suggested method was based on learning a Markov chain rather than learning the entire probability distribution (Bengio et al. 2014).

In the field of spiking neural networks, Dan and Poo (2004) generated the spike-timing dependent plasticity (STDP) algorithm as an example of using an unsupervised learning paradigm in spiking network training. STDP has a significant role for implementing synaptic plasticity impact in the SNN which broadly mimics the biological brain (Daucé 2014). The reinforcement learning paradigm has also been applied to spiking neural networks using STDP as learning rule with a modulatory signal (Davies 2013).

2.5. Sound Source Localization and Machine Learning Methods

In the last few years, there have been growing numbers of attempts to implement machine learning methods to solve the problem of sound localization (Sun et al. 2018). Berkly (1993) stated that supervised learning methods and backpropagation are the most suitable methods to

solve the source localization problems. Diverse types of machine learning methods with supervised learning have been proposed in the literature. (Chen and Ser 2009) applied least squares support vector machines (LS-SVMs) for acoustic source localization by using arrays of microphones using the time delay of arrival (TDOA) as a feature. The proposed algorithm requires the measuring of the microphone array and prediction of the TDOA. The accurate prediction of the TDOA is determined by the microphone position that related to the sound source. Li and Liu (2013) proposed a sound source localization method by using a Gaussian mixture model GMM. This method was based on analyses of time delay features, and the localization was performed by applying a spatial grid matching (SGM) process. The Gaussian mixture model was structured as the form for each grid based on the feature of acoustic time difference.

Another DOA method using microphone arrays that used the GCC for feature extraction (Sun et al. 2018) was based on applying probabilistic neural network (PNN) as a classifier for DOA estimation. The SSL problem has been addressed as stationary single source localization inside enclosed room. The room was divided into the set of space clusters, each cluster refers to the unique three-dimension coordinate. The classifier is working to determine the cluster that the source belong to. PNN is constructed from four layers. The first layer has a number of neurons equal to the GCC feature dimensions, the second layer is called the pattern layer and has a number of neurons equal to the total number of training samples that are used to train the PNN, the third layer is summation layer and has number of neurons equal to the room space clusters, and the final output layer with only one neuron that is responsible for decision making and selecting the most likely class. Figure 2.6 explains the cross-correlation classification algorithm (GCA) Sound source localization model.

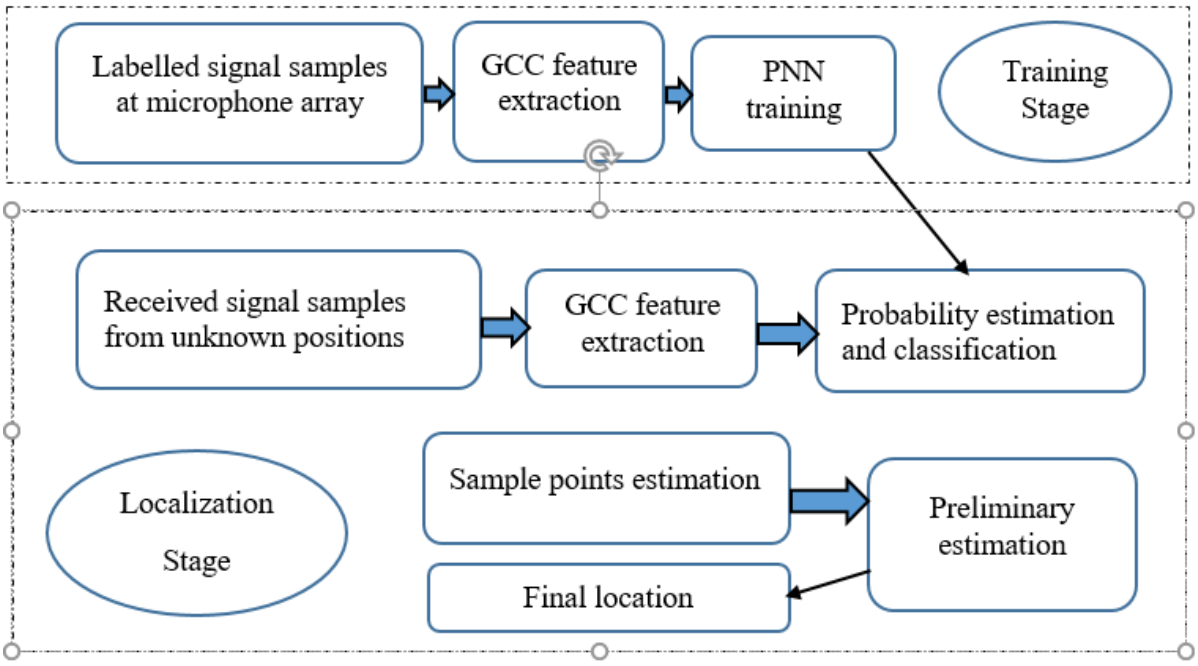


Figure 2.6: Flow chart of the proposed GCA (Sun et al. 2018)

Supervised feedforward neural networks are used for binaural sound source localization by (Datum et al. 1996). Two neural networks work on the same input stimuli to estimate azimuth and elevation angles individually. Each neural network consists of three layers and are trained by applying the Multiple Extended Kalman Algorithm (MEKA). The time and intensity of the received signal are analysed by using a narrow-band filter bank. The input for feedforward neural networks is the intensity differences and time differences of binaural signal.

2.5.1. Sound source localization using multilayer perceptron (MLP)

Xiao et la (2015) proposed a multilayer perceptron approach (MLP) as an advanced step for applying neural networks to solve localization problems in noisy and reverberant environments. This method relies on applying a learning-based method for estimating direction of arrival (DOA) by using microphone arrays. The MLP is trained to perform sound source localization using features extracted from the generalised cross-correlation (GCC). The experimental outcomes from this work appear to enhance the localization performance in noisy and reverberant conditions. It was demonstrated that the model performance was strongly correlated to the size of the training data set. These types of methods which learn by using an enormous size of training data depend on the availability of massive quantities of data.

2.5.2. Sound source localization using deep neural networks

The literature presents many deep neural network methods that implement supervised learning to achieve sound source localization (SSL) in noisy and reverberant conditions (Yalta, Nakadai et al. 2017, Takeda and Komatani 2016a). The most advanced work that has employed deep neural networks was presented by (Takeda and Komatani 2016a, Yalta et al. 2017). The learning-based methods that used larger amount of training data demonstrated solutions to many aspects related to human perception (Xiao et al. 2015). (Takeda and Komatani 2016a) proposed a fully-connected recurrent deep neural network (RDNN) for sound source localization with using discriminative training. The model used the hierarchical integration of directional information at each sub-band of frequency. The Multiple Signal Classification (MUSIC)was used as a feature extractor. This work demonstrated that successful sound source localization method requires frequency domain and time-information. The experimental findings recommended that a well-structured deep neural network can overcome many limitations related to sound source localization such as multisource sound localization and detecting unknown direction (Takeda and Komatani 2016a).

Generally, deep neural networks have led to considerable contributions in signal processing fields. Deep learning has enabled considerable progress in computer vision (Krizhevsky et al. 2012) and speech recognition (Deng and Platt 2014). Yalta et al. (2017) adopted a novel deep convolution neural networks (DCNN) for sound source localization in noisy and reverberation environments. He et al. (2016) argued that applying deep learning rather than MUSIC for localizing sound source showed increased performance. The model was based on microphone arrays with CDNNs trained by applying residual learning. The model tests outcomes appeared to be effective when localizing a single sound source in an anechoic, low noise environment. Performance became degraded at higher reverberation times (e.g. >500 ms). Furthermore, the model is limited to localizing single source audio and not multisource. The authors mention many suggestions for improving the model localization abilities, such as determining hyperparameters for the suggested model and further investigation into the residual learning capabilities for SSL in challenging conditions.

2.5.3. Binaural hearing and Spiking neural networks

Spiking neural networks are useful in the processing of binaural signals as they process signal in the time domain, highlighting relative time variations between signals. Due to the ITD, a time delay can occur between sound signals at both ears. In contrast, the SNN captures timing information as rate of spike input trains into its networks.

Many important research papers have contributed to developing a variety of spiking neural models inspired by the biological processes that happen inside the brain. Glackin et al. (2010) introduced a spiking neural network (SNN) based on the model of the medial superior olive (MSO). The model was tested on an HRTF data set taken from an adult domesticated cat, measured over a limited azimuth range (from −180 angle degree to 170 angle degree). The researchers investigated the impact of adjusting the ITD on the algorithm. ITD is important localisation cues at low frequencies in the range of (270 -1500 kHz) where the wavelength of the arriving signal at each ear is greater than the head diameter. The researchers used Jeffers's model (1948), to process interaural time differences inside the brain (figure 2.7). Two key features characterise the Jeffress model concept which is the axonal delay lines produce internal delays and coincidence detector neurones fire at the maximum rate if excited simultaneously from both sides (Calmes, 2009).

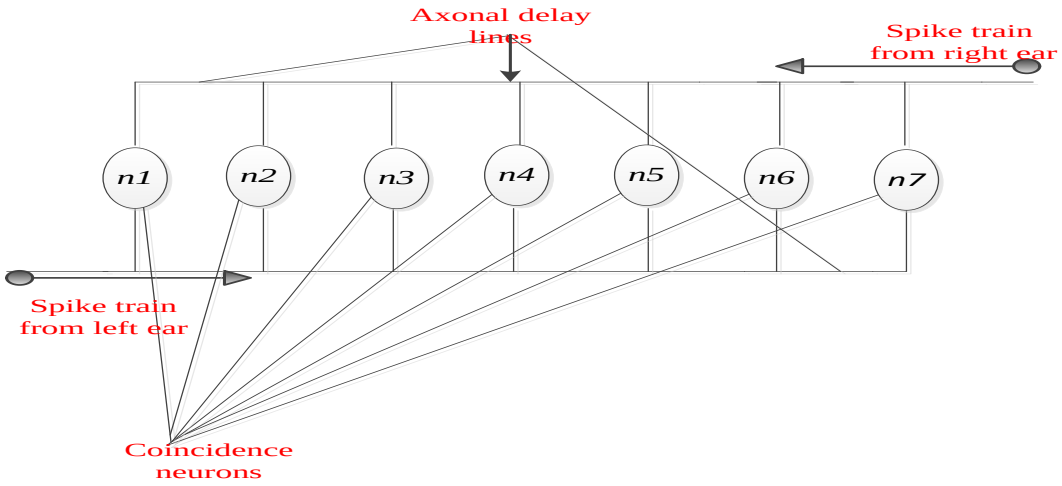


Figure 2.7: Coincidence neurons of Jeffress Model

Kriener and Pfeil (2014) explored the impact of different synaptic parameters on localisation accuracy. The results suggested that if the input frequencies and the number of neurones are selected in a suitable way, it leads to successful localisation performance using

systems analogous to the Jeffress model. A huge stride forward in natural sound localisation modelling has been established by Wall et al. (2012) when they produced a biologically inspired SNN-based algorithm modelled on mammalian auditory pathways. Experimentally determined HRTF datasets for left and right ears were utilised as part of the training data of the model. Moreover, the ITD was extracted and used to detect the sound source position by using the azimuthal angle. This model used a supervised learning method to assess the capabilities of the SNN model and achieved a high accuracy for the localisation process. Sound localisation with cochlear implants (CIs) at different signal-to-noise ratios have been explored and compared to the actual hearing system (Kerber and Seeber2012). The outcomes confirm that CIs appeared less effective in noisy circumstances.

The function of cochlear implants and superior olivary complex (SOC) were investigated by Jindong et al. (2008). A spiking neural network has been applied to compute the two-dimension ITD and ILD spike maps across frequency. Then, these maps have been scaled considering the progress of ITD in low frequency and ILD in high and middle frequencies. Then, ITD and ILD maps had integrated to perform sound source estimation. Pourmohammad and Ahadi (2013) Provided details regarding the of time delay estimation (TDE). TDE is usually applied in N-dimensional wideband sound source localisation in free field environments using, at minimum, N + 1 microphones. The main target of this research was to decrease the number of microphones used. Moreover, the researchers proposed and actualized TDE-ILD-HRTF-based 3D whole space sound source localisation by applying three microphones.

Goodman and Brette (2011) presented the location estimation process depends on spike timing that transfers information about auditory stimuli precisely in the auditory system. They suggested two different ways to process the input binaural signals. The first method, which called the ideal model, is depended on representing the complete set of HRTFs (i.e., all possible locations in the selected HRTF data set). While the second method, which called the approximate model, is relied on representing only gains and delays that extracted from the input binaural signal. The location estimation process depends on spatial-temporal filtering and spiking nonlinearity. The auditory pathway (cochlea) are simulated using set of Gamma-tone filter banks applied on the resulting signals, followed by neural filtering. They employed the key feature of spiking neural networks in processing the temporal signal to solve the sound

source localization problem by using only two sensors. Figure 2.8 explains the sound source localisation (SSL) model that applied two methods (ideal model or approximate model) for handling the binaural transformations to localize sound sources.

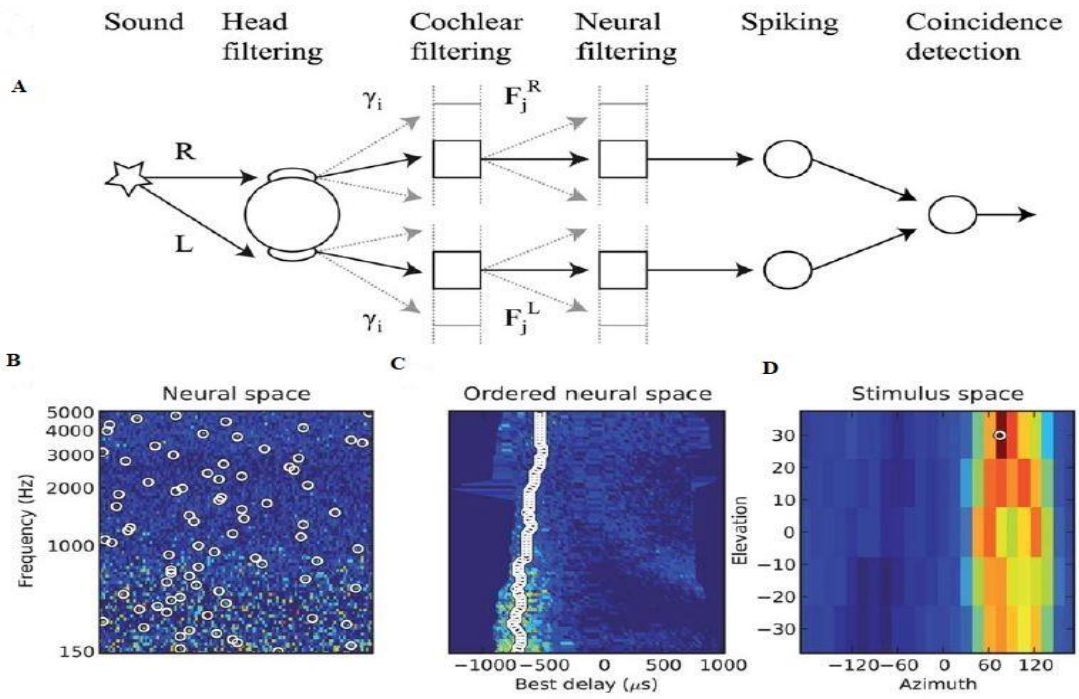


Figure 2.8: Short heading of above images,

(A) Description of model structure. (B) Model response to a sound emitted from at a certain location. Colours indicate to the firing rate of post synaptic neurons, vertically ordered by preferred frequency (the horizontal axis represents a dimension orthogonal to the tonotopically axis). The white circles refer to the neural assembly that encodes the given location. (C) like (B), but neurons are sorted by preferred interaural delay. (D) complete response of all neural assemblies to the same sound submitting, as a function of their appointed location which represent by most activated neurons assembly.

2.6. State-of-Art Multisource Localization

There is lack of research that emphasises multisource localization (Takeda and Komatani 2016b). In signal processing, multi-sound source localisation is a generic issue often described as the 'cocktail party effect' problems. The MUSIC algorithm is the earliest approach that uses DOA prediction for multisource localization. It works by searching the for spectral peak in the

spatial spectrum that results from applying orthogonality of the signals and then using it to localise sources by detecting DOAs (Schmidt 1986). Later, Bechler and Kroschel (2003) suggested an improvement of the Generalized Cross-Correlation (GCC) algorithm to consider the second peak as an index of the second source in multisource localization. The study suggested that both incoming sound sources have equal power leading to two peaks in comparable order of magnitude in the GCC function. The outcomes showed that the estimation accuracy of the second source by using the 2^{nd} peak criteria was 47.83%. The computational cost of localization methods that use TDOA is increased as the size of microphone array increases. However, the reliability of TDOA estimation depends on large numbers of microphones (Nesta and Omologo 2012). Many studies have focused (Ishi et al. 2009, Shiiki and Suyama 2015) on enhancing the MUSIC algorithm, but still there are many limitations related to multisource localization tasks. The first limitation is the computational overhead, while the second is the requirement of prior awareness about the number of original sources (Ishi et al. 2009).

Source separation methods that are based on the statistical independence of the individual sources, such as independent component analysis (ICA) (Comon and Jutten 2010), have been broadly used for multisource localization (Loesch et al. 2009, Lombard et al. 2011). Nesta and Omologo (2012) assumed that the number of dominant sources exactly match the number of microphones in each time-frequency domain. Likewise, the localization methods that apply sparse component analysis (SCA) (Swartling et al. 2011, Pavlidi et al. 2012) are performed under the supposition that in each time-frequency region there is always one source that has energy which is much higher than the other sources.

Pavlidi et al. (2013) proposed a method for multiple sound source localization that relied on detecting time-frequency regions where there is only one source is active. The proposed method attempted to cope with the ambiguity introduced by the linear array by applying a circular array. This concept is improved by Jia et al. (2017). A soundfield microphone was used. The suggested method, shown in figure 2.9, explains how a relaxed sparsity constraint of the speech signal was applied to search the presence of “single-source” region among the sound field microphone's recorded signals. The method was based on detecting the single source

region and then predicting the DOAs of active sources using a peak searching method on the predicted TOA's normalized histogram.

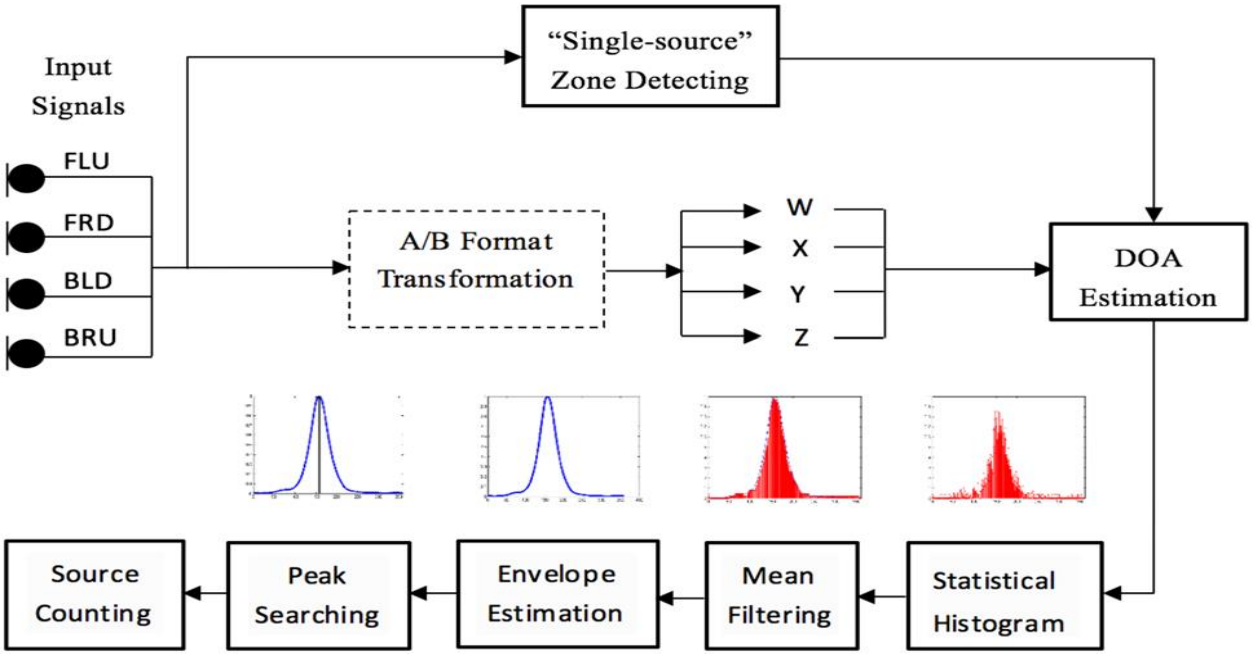


Figure 2.9: The multisource localization model presented by (Jia et al. 2017)

2.7. Chapter Summary

This chapter has reviewed previous work on sound source localization particularly in noisy and reverberation environments. Firstly, a review of the structure and function of the human hearing system was presented. Then, spatial hearing and the basic features of localization cues were presented with comparison between two distinct types of localization; binaural and monaural localisation. The chapter covered the type of features and the most commonly used techniques like, correlation analysis, beamforming, and signal subspace methods. This chapter has also reviewed various machine learning approaches including deep neural networks and spiking neural network approaches to sound source localization, and the diverse types of learning algorithms that were applied with these techniques. Finally, the chapter covered some of the literature that deals with multisource localization and the features that should adopted to improve the localization accuracy in this field.

From this chapter the following points can be summarised as follows:

• Most source localisation methods are designed for a single one source and use microphone arrays. These methods are based on estimating the TDOA of received signal at each microphone. There are many limitations related to using microphone arrays, for instance, the high localization performance requires increased the number of microphones. The computational cost of localization methods that use TDOA is increased as the size of microphone array increases. However, the reliability of TDOA estimation depends on large numbers of microphones.

• Many single source localisation models have been improved to work in noisy and reverberant environments. However, more studies are needed to investigate the localization performance with long reverberation times (>500ms) and lower signal-to-noise ratios.

• Machine learning approaches for SSL were reviewed. Most of these efforts used supervised learning for sound source localization. The researchers argued, the better localization performance required supervised learning. This is inspired from the human abilities in audio-vision integration to determine the right sound signal direction.

• To decrease the computational complexity, various approaches have used only individual localization cue (ITD or ILD) to get the binaural information for sound source localization. While, the literature studies demonstrate the accurate sound source localization required all localization cues ITD, ILD and spectral cues integrated together for better binaural sound representation and localization performance.

• DNN and SNN are advanced machine learning networks that show promise for solving sound source localization from binaural signals. Many research papers have contributed to promote a variety of SSL models that used spiking neural models inspired by the biological processes that happen inside the brain. The methods that used Jeffers's model concept to translate the binaural time delays and neuronal firing rate coincidence detectors are the most successful method for binaural localization.

•For multiple sound source localization, in the last few years, methods were developed that detecting time-frequency zones where there is only one source is active. One of the significant limitations that captured in the reviewed methods for multisource localization is, all these

methods are based on one concept which is 'there is only one sound signal has energy over the other sources in each time instant'.

CHAPTER 3

BACKGROUND AND DATA SETS

Chapter Overview

The previouschapter presented are view of previous work that focuses on general methods *of sound source localization and machine learning algorithms employed in the field of sound source localization. This chapter covers the background, methods and materials that have been used in this research. In addition, speech and HRTFs databases that are used in this research will be described in this chapter. Section 3.1 explains binaural source localisation and head-related transfer functions. Section 3.2 demonstrates the basic components and principles of spiking neural networks (SNNs). Deep neural networks (DNNs) and their main functions and constructors are described in section 3.3. Section 3.4 explains the mathematical concepts of backpropagation learning algorithms as a general-purpose learning method. Support vector machines (SVM) for multi-class classification is introduced in section 3.5. Finally, a description of the two HRTF databases and the speech database adopted in this thesis are detailed in section 3.6.*

3.1 Binaural Source localisation

3.1.1 Binaural hearing and sound source localisation

Binaural hearing refers to a feature of the human auditory system which utilises several cues, extracted from the signals from both ears, to provide spatial information about sound sources. Binaural cues, caused by differentiation of the signals between the ears, have a significant role in the localisation process.

The left and right HRTFs (H_L and H_R) for both ears are defined by the following equations

$$H_L(r,\theta,\emptyset,f,a) = \frac{P_L(r,\theta,\emptyset,f,a)}{P_0(r,f)}, H_R(r,\theta,\emptyset,f,a) = \frac{P_R(r,\theta,\emptyset,f,a)}{P_0(r,f)} \quad 3.1$$

Sound source location is described using a spherical coordinate system (r, θ, ϕ), P_L and P_R refer to sound pressures for left and right ears in the frequency domain; P_0 is the free-field sound pressure in the absence of a head, r is the distance between the sound position and head centre, θ is the azimuth (0° to 360°), and ϕ denotes elevation r (-90 to 90). Depending on r, the HRTF is classified as far-field HRTF when r has a value greater than $1.2m$ and near-field for values below this (Duda and Martens 1998, Sheaffer 2013). The parameter 'a' in the above equation indicates the set of factors determining the dimensions of the pertinent anatomic shape of each human. For far field source localisation, the three-dimensional location of the sound source in a free field space can be defined by two angles; a horizontal angle (azimuth) θ and a vertical angle (elevation) ϕ. This representation is illustrated in Figure 3.1.

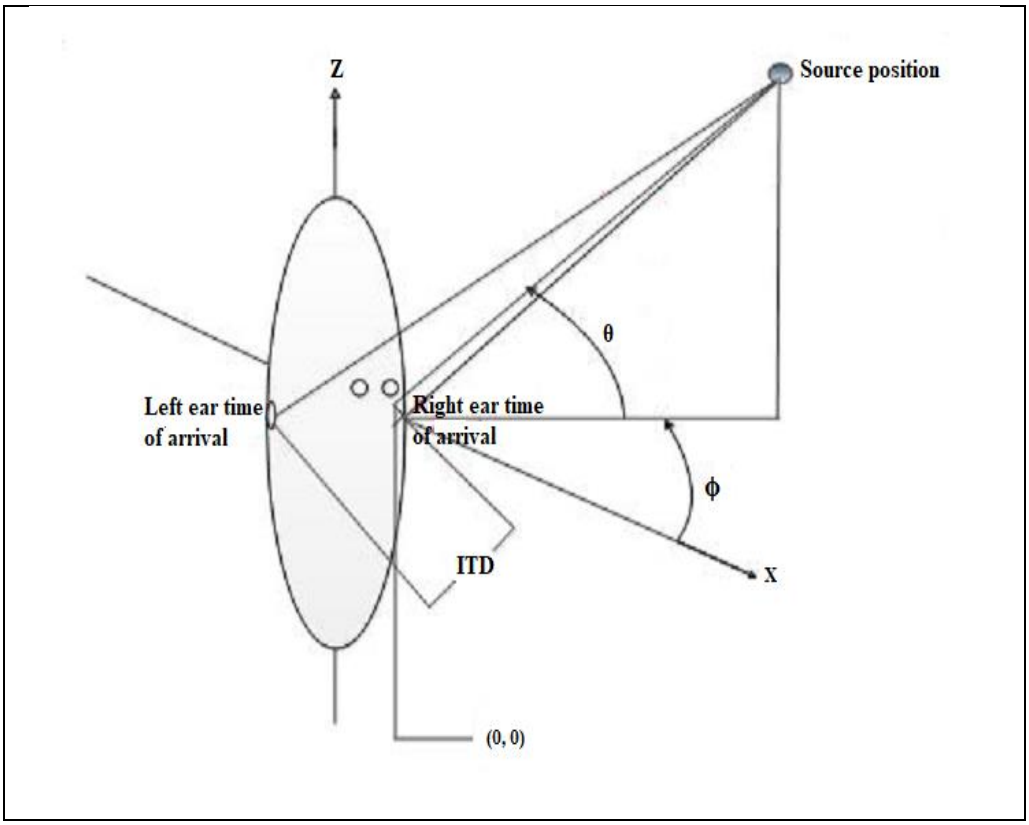


Figure 3.1: Interaural time differences for the arrival of the signal at both ears.

HRTFs have many physical characteristics related to frequency and time domain properties while various localisation cues can be evaluated from measured HRTFs (Zotkin et al. 2003). The impact of these localization cues on the proposed localisation model will be assessed separately by changing the scope of frequencies for the recieved sound signals. HRTFs or HRIRs consist of the localization cues ITD and ILD which are described as:

$$ITD_P(\theta, \phi, f) = \frac{\Delta\psi}{2\pi f} = -\frac{\psi_L - \psi_R}{2\pi f} \qquad 3.2$$

Where:

ψ_L refers to distributed phase of the left ear.ψ_R refers to the distributed phase of right ear.

$$ILD(r, \theta, \emptyset, f) = 20\log_{10}\frac{H_R(r,\theta,\phi,f)}{H_L(r,\theta,\phi,f)}\ (dB) \qquad 3.3$$

While the HRTF are defined in spherical coordinates, perception relies on only ITD and ILD cues (Xie 2013, Ahveninen et al 2014). However, sound sources can be located at various positions that produce the same ITDs; these sources lie within what is known as the "cone of confusion" (Kapralos et al 2008). The cone of confusion phenomenon can lead to poor localization performance in distinguishing sound sources in vertical planes especially when the wavelength of a continuous signal approaches the head diameter in this case both ITD and ILD have zero values in the median plane (Wallach1940, Miller 2013).

HRTFs are never directly accessible to the auditory system because they are always embedded with the source signal. To model the HRTF, a finite impulse response filter (FIR) or infinite impulse response filter (IIR) is captured that captures the source-to-receiver transfer functions for a range of positions. IIRs tend to be implemented in the time domain and FIRs can be either time or frequency domain(Hao et al. 2007). High quality HRTF datasets are vital to modelling binaural hearing model accurately (Zhang et al. 2014). Generally, HRTF measurements process can be time consuming with a high degree of complexity due to the requirement for a high degree of thoroughness. To minimise these difficulties while achieving good control of the measurement environment, dummy head HRTF data setis suggested for most research purposes (Carty 2010).

A comprehensive HRTF dataset is selected to test our model for sound source localisation in a free field environment. Gardner and Martin presented a vast collection of head related transfer function measurements (Gardner and Martin 1995). More details about these data are discussed in section 3.6.

3.1.2 Head related transfer function and inverse problems

Some numerical techniques have been improved on the using the sophisticated geometry of a dummy or human head. Boundary elements method (BEM) is one tool used to solve the scattering problem of the acoustic wave. This method works in two steps; firstly, the acoustic wave is mutated into a boundary surface integration. Secondly, The boundary surface is descritised into a mesh of elements, leading to a set of linear algebra equations (Xie 2013). There is a body of research using this method, and these attempts investigate the relationship between the computational complexity of BEM as a function of frequency and showing that there is a directly proportional relationship between them (Ziegelwanger et al. 2015).

A transfer function captures the gain and phase transformation of a linear-time-invariant system. HRTFs can be captured in anechoic environments as the Fourier transform (FT) of the head related impulse response (HRIR) that constitutes the binaural impulse response from a given source position in the time domain. The HRTF and HRIR are linked by Fourier transform as explained in the following expressions (Xie 2013):

$$h_L(r,\theta,\emptyset,t,a) = \int_{-\infty}^{+\infty} H_L(r,\theta,\emptyset,f,a)e^{j2\pi ft}df,$$

$$h_R(r,\theta,\emptyset,t,a) = \int_{-\infty}^{+\infty} H_R(r,\theta,\emptyset,f,a)e^{j2\pi ft}df; \qquad 3.4$$

$$H_L(r,\theta,\emptyset,f,a) = \int_{-\infty}^{+\infty} h_L(r,\theta,\emptyset,t,a)e^{-j2\pi ft}dt,$$

$$H_R(r,\theta,\emptyset,f,a) = \int_{-\infty}^{+\infty} h_R(r,\theta,\emptyset,t,a)e^{-j2\pi ft}dt.$$

To simulate the binaural signal that would be present at the ears for a source at a particular location, a waveform (*S)* is convolved with the HRIR filters (F_L and F_R) that have the closest azimuth and elevation to that of the source:

$$F_L * S, F_R * S$$

The process (*)represents a convolution process between HRIR and incoming sound signal. In the frequency domain (HRTF) the convolution becomes a multiplication process.

There are three ways to obtain HRTFs; measurements, computation and customization. the most accurate and common method is using measurements, particularly for human individuals. This method is performed in an anechoic chamber. The measuring signal generated by a computer is passed through a digital/analogue (D/A) converter and a power amplifier and then delivered to a loudspeaker. A pair of microphones simulate the human ears are used to record the resultant signals and then delivered to the computer after pass through amplifier and analogue/digital (A/D) converter. After do some necessary signal processing steps, the final HRIRs or HRTFs

are acquired. The second method used the mathematical and physical concepts to obtain HRTFs computationally. Some simplified human anatomical geometry can solve the analytical solution of HRTFs. One of the simplest models for HRTF calculation called spherical-head model where the head is simply indicated as a static circle shape with radius, and the ears are indicated as two opposites points on the circle. The main advantage to use this method is to overcome the scattering issue that results from human geometry (head and torso). The third method used the customization to approximately obtain the individualized HRTFs. This depends on the fact of existing a strong correlation between the individual HRTF and its individual anatomical parameters because the HRTFs characterize the interaction between received sound signals and human anatomical shapes (Zhong and Xie 2014).

It is not possible to invert the binaural signals effectively (at least in real-time). This is because the HRTF is non-minimumphase system and unlike a minimum phase system, the inverse is non-causal(Nam et al., 2008). Minimum-phase refers to a specific feature of a system where both the system and its inverse are causative and stable(Callister and Rethwisch2007).Zeros and poles refer to the roots of the numerators and denominators, respectively, of the polynomial transfer function of a system. Commonly, the poles and zeros of the transfer function are complex, and their positions can be plotted on the S-plane in order to graphically represent the system. The S-plane is also called a zero-pole plot. The Z-plane is a discrete time approximation of the S-plane. The Z-plane is a bilinear transformed representation of the s-plane which, using complex conjugation, expresses the periodicity of a frequency response once it has been discretised, normalised against 2π*F_s (Oppenheim and Schafer 2014). Figure 3.1 shows the pole-zero plot of a transfer function and the Z-plane representation. The unit circle is the equivalent of the y axis on the s-plane folded round and reflected between 0*F_s and π*F_s, betweenπ*F_s and 2π*F_s there are conjugates. This represents the ‘aliased’ response outside the temporally representable portion of the system response. The position of the poles and zeros on the S-plane supply specific view into the response features of a transfer function.

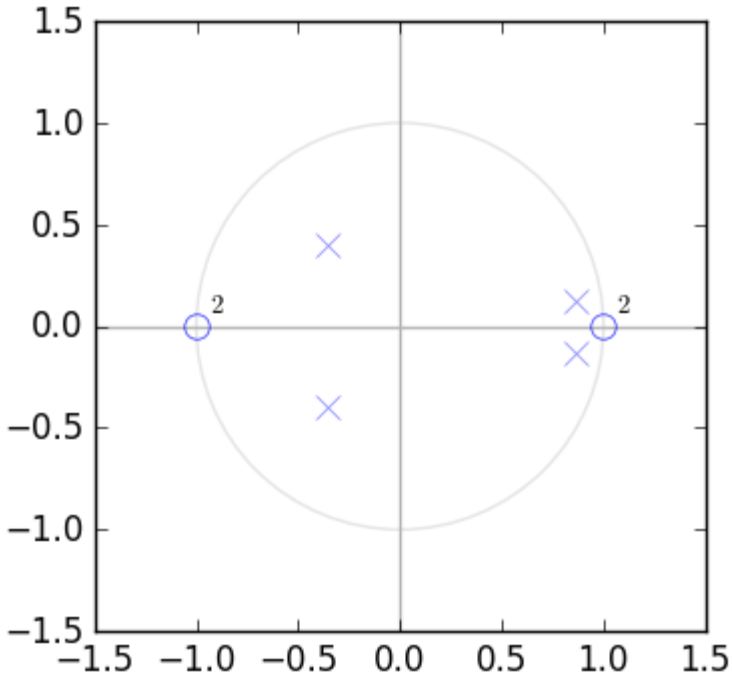


Figure 3.2: Pole and zero plot of transfer function and the z-plane representation.

As time increases in a stable system all its elements of the homogeneous response have to reduce to zero. The system is unstable, if any pole and its complex conjugate lies outside the unit circle. while the zeroes will not influence on the system stability if they lie outside the unit circle, however, this points out to the system that is not invertible (Oppenheim and Schafer 2014). In case of an unstable system, the pole, lying in the outside the unit circle of the *Z*-plane, produces an element in the system homogeneous response that rises without limit from any restricted initial conditions. HRTFs are non-minimum phrase systems and as a result are non-invertible functions. Ziegelwanger et al. (2015) presented work that has been carried out on the numerical calculation of HRTFs. Conversely, the approximation of minimum phase HRTFs examined by (Kulkarni et al. 1995) and the outcomes suggest a description of HRTF phase as a position-dependent ITD that is frequency independent. Approximation of minimum phase HRTFs may computed mathematically, but logically, it is inapplicable because it will lead to the unstable and non-causal system.

3.2 Spiking Neural Networks (SNNs)

Networks of spiking neurons provide a more realistic representation of biological networks compared with traditional artificial neural networks (Bulanova et al.2012). The key feature of spiking neural networks is a temporal coding principle (Figure 3.3) where individual pulses (spikes) are emitted at particular moments in time. Spiking neurons can process multiple information sources into a single flow of signals, such as the amplitude and frequency of sound signals in the aural system (Gerstner et al. 1998).

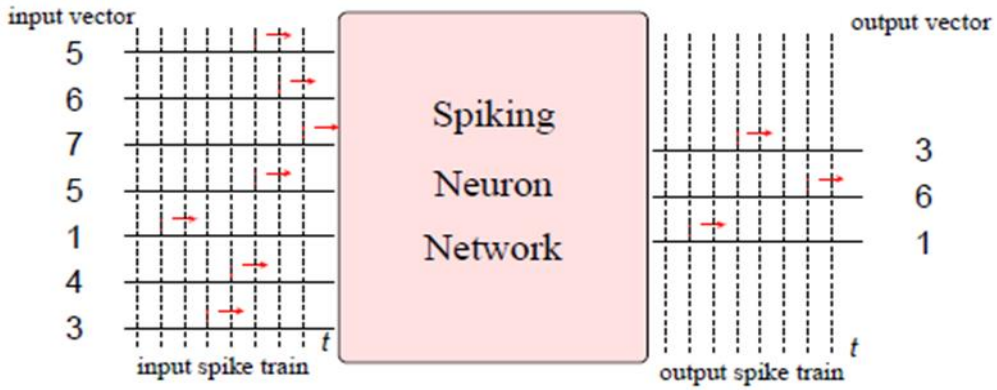


Figure 3.3 The temporal coding principle for encoding and decoding real vectors in spike trains (Paugam-Moisy and Bohte 2012).

3.2.1 Neurons in spiking neural networks

Spiking neural networks consist of the neurons that connect through shortened signs known as spikes. The computational units in spiking neurons include three steps: summation of all neuron input stimuli, integration over time, and finally a spike fires when the membrane potential expands over the threshold and returns to a reset value (Davies, 2013).

A spiking neuron model considers the impact of firing action potentials (spikes) on the internal state of targeted neuron as well as the relation between this neural internal state and the spikes the neuron fires (Paugam-Moisy and Bohte 2012). The membrane potential of each spiking neuron has positive charge it. The inner surface of membrane is filled by negative charge and the outer surface occupied by a positive charge. Those charges generate the membrane potential. In a resting state, the membrane potential does not receive any input cues and are at resting potential. An action potential (spike)occurs when the membrane potential reaches the threshold value and fires the signals. Absolute refractory time refers to the minimum period

between two fired spikes. Hyperpolarization is when the membrane potential is more negative at a certain point on the neuron's membrane, while depolarization is when the membrane potential turns out to be more less negative (more positive). Figure 3.4 explains spiking neural network firing (spikes) and excitatory and inhibitory postsynaptic potentials over time processes. It describes the biological process that will be modelled (Gerstner and Kistler, 2002). The membrane potential represents the internal state of active neuron. Each neuron model describes a different membrane potential which causes firing action potentials when neuron received an enough energy.

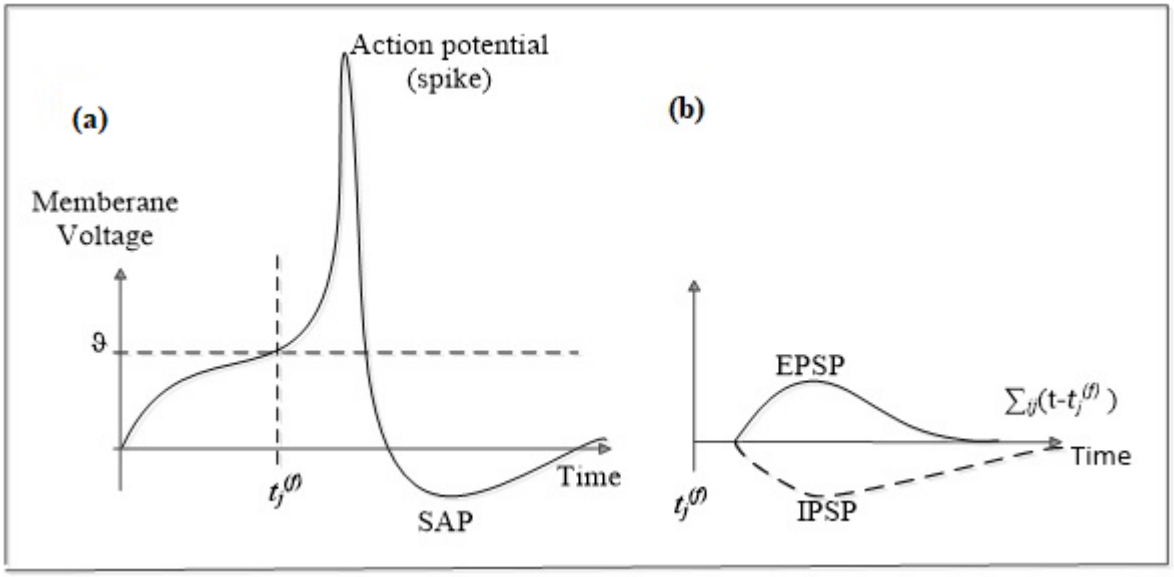


Figure 3.4: Firing process

The part (a)refers to the spike neural network firing process (spikes): The part (b) refers to the excitatory and inhibitory postsynaptic potentials over time general shape (Gerstner and Kistler2002).

The firing process results from the movement of negative and positive ions across voltage-gated ways. The spikes have the same form and are not affected by signal movement between presynaptic and postsynaptic neurons. Spiking neurons communicate through the spikes, and the synapses are responsible fortransferring an electrical or chemical signal between neurons. Figure 3.4 explains the firing process within neurons: each neuron emits spike-trains, which change significantly in frequency across a small period of time. Neurons employ spatial and

temporal information of input spike samples for encoding their data and send it to other neurons (Davies, 2013).

3.2.2 Leaky integrated and fire neuron model

The Leaky integrate and fire spiking neural model (LIF) represents a very simple spiking neuron model. Its analysis and simulation process are relatively easy, and it is widely used. The neuron in a LIF model, in its simplest form, is modelled as a "leaky integrator" of its input current I(t). Figure 3.5 illustrates how the pulse is transferred through the integrate-and-fire neuron. It explains the firing process within neurons; each neuron emits spike-trains, which change significantly in frequency across a small period of time. Neurons must employ spatial and temporal information of input spike samples for encoding their data and send it to other neurons (Davies 2013). The spike-train can be described using the form:

$$F_i = \{t_i^1, t_i^2, t_i^3, t_i^4 \ldots\ldots\ldots, t_i^f\} \qquad 3.5$$

The i indicates the neuron and f refers to the number of the spike, t_i^f refers to the firing time.

$$F_i = \{t \mid V_i(t) = \vartheta \wedge V_i'(t) > 0\} \qquad 3.6$$

The variable *Vi* refers to the membrane potential which explains the internal state of neuron.

A spike comes through the input (the axon) and, using a low-pass filter, converts the spikes from short pulse into an elongated pulse which takes the form $I(t - t_j^{(f)})$. Where *j* refers to the neuron and *(f)* refers to the number of the spike, $t_j^{(f)}$ refers to the firing time. This is used as input to charge the integrate-and-fire circuit which increases a value representing a postsynaptic potential $\varepsilon(t - t_j^{(f)})$.A spike (*1ms*) is generated when the membrane potential of neuron rises over threshold value ϑ.

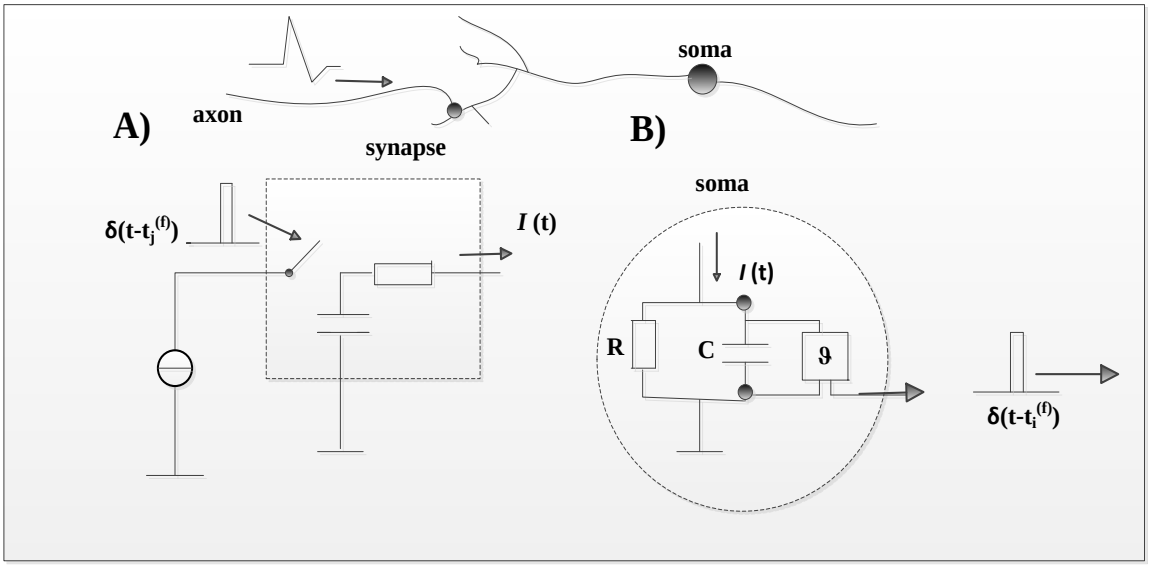


Figure 3.5: The integrate-and-fire neuron schematic design

The integrate-and-fire neuron schematic design. In part 'A' the spike transforms a current pulse I(t) using a low-pass filter and then charges the capacitor. On the right, the schematic shape of the soma, a spike generates when voltage V along the capacitor rises above the threshold (Gerstner and Kistler 2002).

Integrate-and-fire neurons are rely on electronics concepts. Once the voltage rises over the capacitor threshold value ϑ, the neuron fires a pulse itself. The integrate and firing neural model can be described mathematically as:

$$\tau_m \frac{\partial u}{\partial m} = -V(t) + RI(t) \qquad 3.7$$

τ_m refers to the membrane time constant where voltage leaks away so that this model is occasionally called the leaky integrate and firing model. *R* represents the membrane resistance, and equation 3.8 characterises a straightforward resistor-capacitor (RC)circuit where the leakage term results from the resistor and the integration of *I(t)*due to the capacitor being placed in parallel to the resistor (Gerstner and Kistler2002). The spike generates a short pulse(δ) when the neuron fires as soon as *V* passes threshold ϑ. In a refractory period, which is the state that happens after neuron firing directly *V* set to a baseline.

The spiking events in the LIF model start when the membrane potential *V(t)* reaches a certain spiking threshold, it is immediately reset to a reset potential and the leaky integration operation characterised by Equation 3.8starts again with the reset potential initial value (Vreeken 2002).

3.3 The mathematical description of Deep Neural Networks DNNs

A DNN is a sequential neural network that has many successive nonlinear hidden layers. An input feature vectorx_t is transformed among these hidden layers by applying a nonlinear mapping. A DNN can be describe by the following expressions.

$$z^0 = x_t \quad 3.8$$

$$y_i^{(l+1)} = \sum_{j=1}^{N^{(l)}} w_{ij}^{(l)} z_j^{(l)} + b_i^{(l)} \quad 3.9$$

$$z_i^{(l+1)} = \sigma(y_i^{(l+1)}) \quad 3.10$$

Where $N^{(l)}$ refers to the number of units in the lth layer, $W^{(l)}$ is a weight matrix and $b^{(l)}$ denotes to the bias vector in this detected layer. $\sigma(x)$ refers to the activation function which is nonlinear. There are many types of activation functions for different tasks, the most common ones are the sigmoid activation function (equation 3.11), hyperbolic tangent function (equation 3.12), rectified linear unit (ReLU) activation function (equation 3.13) and soft-plus activation function which is an analytic function defined as a smooth approximation of rectification (equation 3. 14) (Goodfellow et al. 2016).

$$\sigma(x) = \frac{1}{1+exp^{(-x)}} \quad 3.11$$

$$\sigma(x) = \frac{1-exp^{(-2x)}}{1+exp^{(-2x)}} \quad 3.12$$

$$\sigma(x) = \max(0, x) \quad 3.13$$

$$\sigma(x) = \log(1 + e^x) \quad 3.14$$

Deep learning is appropriate when enormous amounts of training data are available. It demonstrates state-of-the-art performance for solving problem in various fields involving text,

sound, or image. Furthermore, deep learning has been used in many advances in computer vision and speech recognition. One of the most significant characteristics of deep learning is working with feature representation or abstract representation of training data (Schmidhuber 2015).

For multiclass classification deep neural networks, the goal of training is to determine the boundaries between the different classes in the features-space. Recently, Softmax activation functions have played significant role in solving multiclass classification problems. Softmax is used in the output layer to represent the probability of a particular classes from input vector as explained in the following (Chung et al. 2016):

$$p(s/x_t) = softmax(x_t) = \frac{\exp(w_s y^L)}{\sum_{n=1}^{N^L} \exp(w_n y^L)} \quad 3.15$$

Deep neural networks are trained by updating the weight matrix and bias vector applying a gradient descent algorithm which minimises a cost function across a dataset. The process of learning the weights and biases, is described in equation 3.16 And 3.17, where ϵrefers to the learning rate and $\alpha$is the momentum, respectively. The cross-entropy *C* is computed between the output of Softmax *p(x)* and the target probability *p(x)* as shown in the following equation (Bengio 2012):

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) - \epsilon \frac{\partial C}{\partial w_{ij}(t)} \quad 3.16$$

$$\Delta b_i(t) = \alpha \Delta b_i(t-1) - \epsilon \frac{\partial C}{\partial b_i(t)} \quad 3.17$$

$$C = -\sum_x p(x) \log p(x) \quad 3.18$$

3.4 Learning paradigm

Currently, there are two forms of multi-label classification methods: batch learning and online learning, batch learning was applied. Classes are defined as binary vectors, where each index corresponds to a pair of angles. In the training phase, the categorical cross-entropy was used to determine the error between the neural network outputs and the desired class for each batch. Backpropagation was used to update the learning parameters using the value of the cost function (weights and biases). The partial derivatives of the cost function with respect to the weights and biases (and $\partial C/\partial b$) can be computed by using equation 3.19 (Nielsen 2017):

$$\frac{\partial C}{\partial w_{ik}^{l}} = a_k^{l-1}\delta_i^l \text{ and } \frac{\partial C}{\partial b_i^l} = \delta_i^l \qquad 3.19$$

Where δ^l_i, denotes to the error in the i^{th} neuron in the l^{th} layer and w^l_{ik} refers to the weight value that connect between k^{th} neuron in the layer to the i^{th} neuron in the next layer. b^l_i is the bias of the i^{th} neuron in the l^{th} layer. a^l_i refer to the activation function of the i^{th} neuron in the l^{th} layer.

Figure 5.8 explains the training and validation process steps that were applied to train and validate the DNN for multisource localization. The DNN has been trained using the full size of training data that was generated using various speech samples belonging to 17 speakers to predict 4032 angle pairs (classes) in case of using IRCAM and 4800 angle pairs for KEMAR.

To validate the multisource localization model performance, a new data set that was generated from completely new speech samples belonging to 3 speakers (two males, one female) was used in the model validation stage. Generally, all the results that shown in this chapter resulted from applying all suggested models with validation speakers. So that completely fresh samples have been introduced to the previously trained localization model to investigate its ability in predicting the sources from unknown sound signals.

3.5 Backpropagation learning Algorithm

The original backpropagation algorithm consisted of two complementary steps. The first is the forward step, where the network outputs are computed from its inputs and initial weights. In the second step, a cost-function (cross-entropy) is computed by comparing the true classes with the predicted classes of the training data. This is then propagated back to update the networks parameters (weight and bias) which gives the algorithm its name ‘backpropagation’ (Nikoskinen 2015). The backpropagation algorithm is the fastest algorithm used for computing the gradient of the cost function (Buscema 1998). The Gradient-based concept is the principle of most optimization algorithms that are used to optimize the loss function with respect to the neural network parameters (Bengio2012).

In the case of multi-class classification, the backpropagation algorithm with cross entropy for multiple hidden layer networks can be described by the following steps:

$$F(x) = \begin{pmatrix} P(Y = \frac{1}{x} \\ \vdots \\ . \\ P(Y = \frac{K}{X} \end{pmatrix} \quad 3.20$$

Where *K* refers to the class classification problem and *F(x)* is the output layers when the output activation function is considered as Softmax (see equation 3.15)

$$softmax\ (x_1, \dots\dots, x_K) = \frac{1}{\sum_{i=1}^{K} e^{x_i}} (e^{x_1}, \dots\dots, e^{x_K}) \quad 3.21$$

to compute the gradient, the following computations are described:

$$\frac{\partial softmax(x)_i}{\partial x_j} \begin{cases} = softmax(x)_i\ (1 - softmax(x)_i\ )\ \ if\ \ i = j \\ = -softmax(x)_i\ softmax(x)_j \qquad if\ \ i \neq j \end{cases} \quad 3.22$$

Then equation 3.23 is introduced, where $f(x)_k$ refers to the kth components of $f(x)$: $(f(x))_k = P(Y = \frac{k}{x})$.

$$(f(x)_y) = \sum_{k=1}^{K} 1_{y=k} (f(x)_k) \quad 3.23$$

Then

$$-log(f(x)_y) = \sum_{k=1}^{K} 1_{y=k} \log(f(x)_k) = \delta(f(x), y) \quad 3.24$$

The symbol δrefers to the loss function correlates to cross-entropy (Buscema 1998, Sadowski 2016)

3.6 Support Vector Machine SVM

The support vector machine (SVM) is one of the discriminative binary classifiers proposed by Vapnik as a set of related supervised learning techniques (Vapnik 1995). Initially, SVMs were developed to perform classification functions (Burges 1998) and then expanded for regression tasks (Smola and Schölkopf 2004). In binary SVM classifiers, each data point belonging to only one of two classes is represented by an n-dimensional vector. A linear classifier is trained to separate these two classes of data using a hyperplane. SVMs can use a kernel function to learn non-linear boundary regions between training samples by mapping the input samples into higher dimensional space (Pillay and Govender 2017). A classification score for a data point is

acquired by evaluating the distance of the predicted sample to the hyperplane (Evgeniou and Pontil 1999). The mathematical expression for the training data set can be described as:

$$D = \{(x_1, y_1) \quad (x_2, y_2) \dots \dots \quad (x_m, y_m)\} \qquad 3.25$$

D refers to the input data point that used as training set for SVM classification function.x_i refers to m-dimensional vector. $y_i$indicates the class of the sample x_i , taking either 1 or -1. The SVM classifier F(x) can be described by using equation 3.26.

$$F(x) = w.x - b \qquad 3.26$$

w and *b* indicate to the weight and bias vectors, which are updated throughout the training process by the SVM. For multiclass classification implementations, binary classifiers can have combined by using pairwise coupling (Hastie and Tibshirani 1998). Other methods used to implement SVM for Multi-class classification include, one-against-one (OAO) and one-against-all (OAA) classifiers. In the OAA, M binary SVM classifiers are constructed for M-class problems. In the training phase, the samples in the one class are labelled as positive samples while all the rest samples are labelled as negative. In the prediction stage, the classification is acquired from all M-SVM classifiers. The test sample is labelled by using the maximum output among the M classifiers (Vapnik 1998, Hsu and Lin 2002). In One-Against-All method, the expectation of the likelihood of involving errors on the test examples are computed by applying the following equation:

$$E[P(error)] = \frac{E[\,number\ of\ training\ points\ that\ are\ support\ vectors]}{(number\ of\ training\ vectors)-1} \qquad 3.27$$

This equation calculates the proportion of the expectation of the number of support vectors which are training points to the number of training set examples (Vapnik 1995). Generally, the studies demonstrate that the OAA using a polynomial kernel performs better in solving multi-class classification problems compared with other types of multi-class SVM approaches (Chamasemani and Singh 2011).

3.7 Research Databases

In this section, the HRTF and the speech databases used in this research will be briefly explained. The KEMAR HRTF data set and IRCAM HRTF data set are used to investigate sound source localization models in different experimental conditions. Two HRTF data sets were used to train and validate the model for single and multisource localization. These datasets

have different anatomical parameters (head size, ear shape and torso), this improves the model generalisation over previous work which was limited to a single torso simulator. Furthermore, anechoic speech samples were convolved with binaural responses were used to test and validate different sound localization models in this research.

3.7.1 KEMAR Dummy HRTF Dataset

The KEMAR Dummy head HRTF dataset was captured from sound sources in a free field environment. Gardner and Martin presented an enormous collection of head related transfer function measurements which was published as 'HRTF KEMAR' dummy head data sets (Gardner and Martin 1995). The measurements were carried out in an anechoic room. The KEMAR mannequin was raised vertically on a portable turntable that allowed an accurate controlled rotation in any azimuth. The speaker was raised on a microphone stand which provided a precise elevation about the mannequin. The measurements are presented one elevation at a time. Elevations in the range of -40 to 90 degrees with 10° regular step increments. While the azimuth within the range 0 to 360 degrees with asymmetrical increment degrees to cover a variety of spherical coordinates around KEMAR head (Gardner and Martin, 1995). Table 3-1 explains the dimensionality of KEMAR dummy head HRTF database including the number azimuth measurements for each elevation. It has 710 locations along vertical and horizontal planes and the sample frequency is 44.1 KHz.

Table 3-1: KEMAR dummy HRTF number of measurements and azimuth increment at each elevation.

Elevation	Points Per elevation	Azimuth Increment Per elevation (degree)
-40	56	6.43
-30	60	6.00
-20	72	5.00
-10	72	5.00
0	72	5.00
10	72	5.00
20	72	5.00
30	60	6.00
40	56	6.43
50	45	8.00
60	36	10.00
70	24	15.00
80	12	30.00
90	1	x.xx

The following figures show samples of impulse responses for the left and right ears from KEMAR dummy dataset in particular directions.

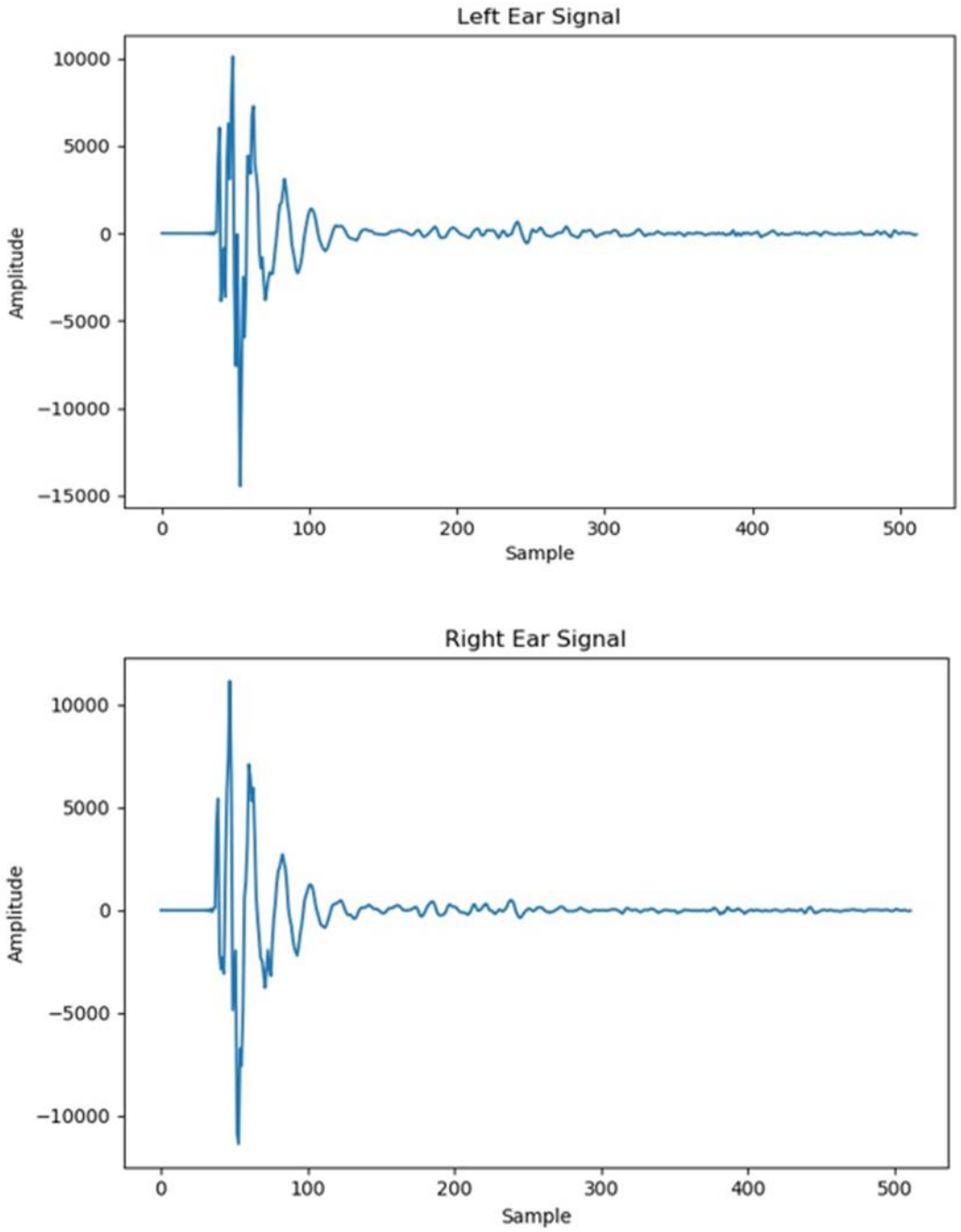


Figure 3.6: Impulse responses for the left and right of KEMAR dummy ears in the time domain with azimuth=0˚ and elevation=0˚.

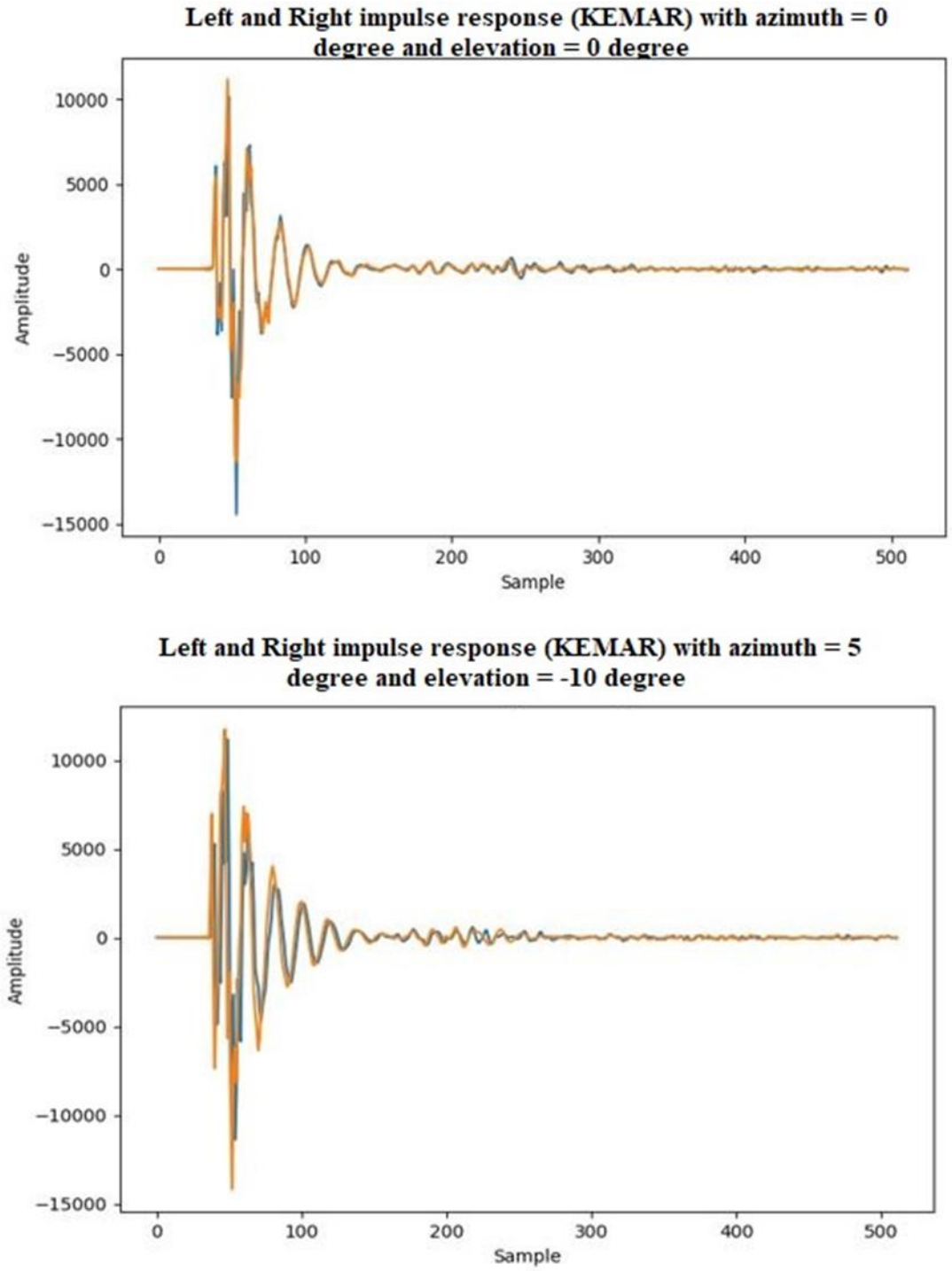


Figure 3.7: Head-related transfer function for the left and right KEMAR dummy ears in the time domain.

Frequency domain response, viafast Fourier transform (FFT) on a windowed frame, using a Hanning window with size 512, of impulse responses data samples are shown in the following figures:

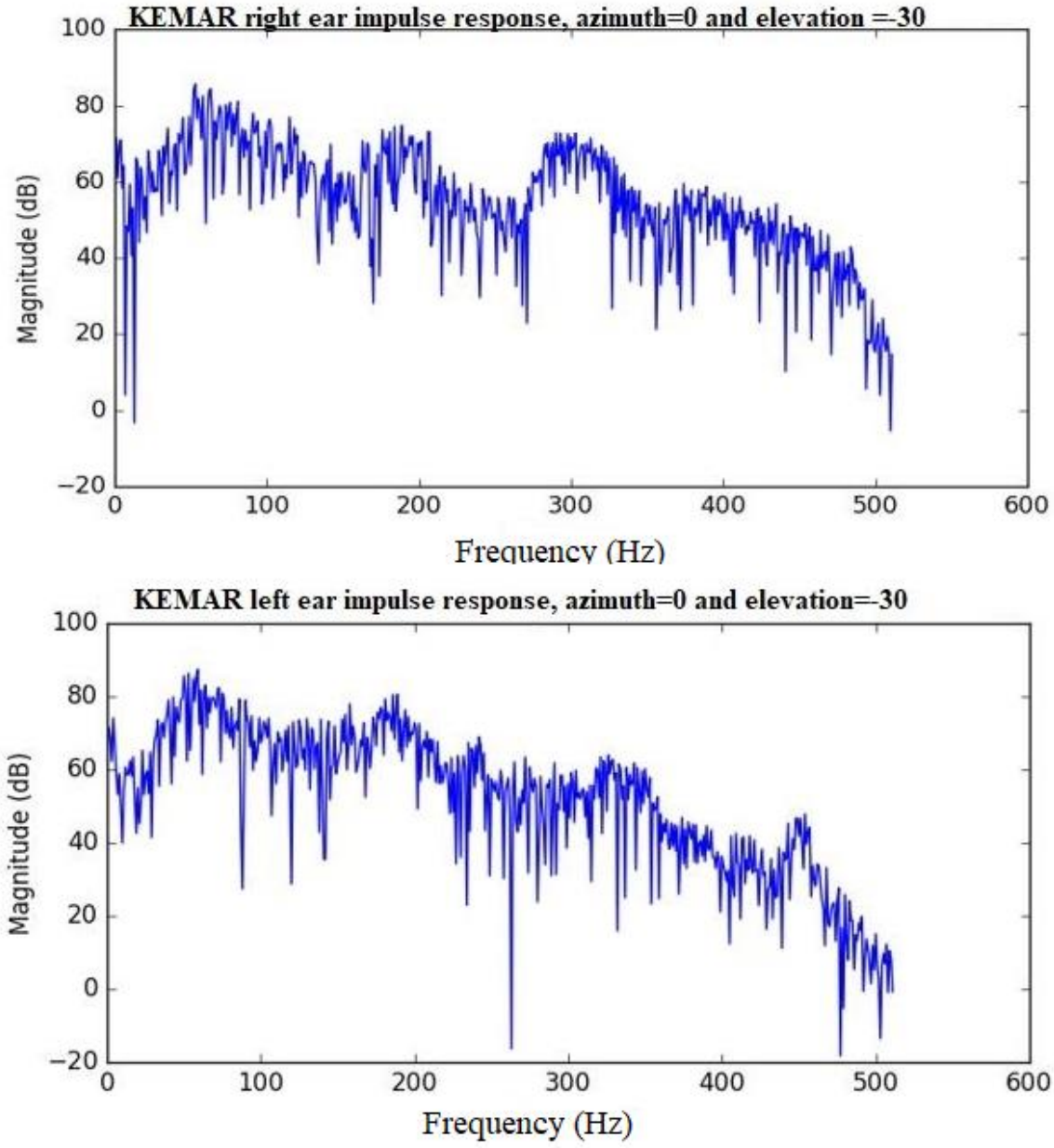


Figure 3.8: KEMAR normalised impulse responses in the frequency domain.

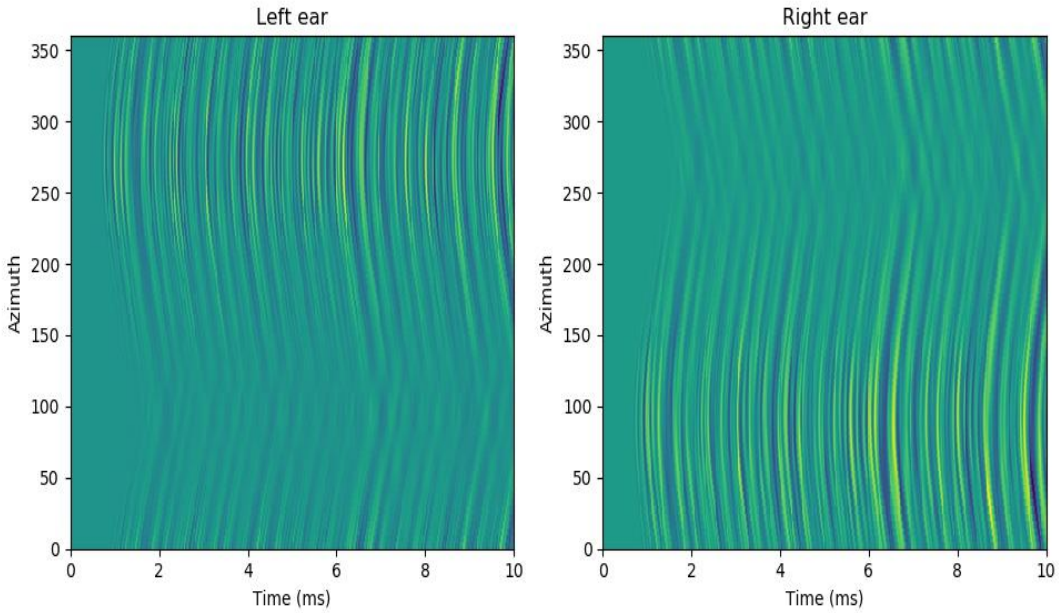


Figure 3.9: Impulse response of KEMAR database in the horizontal plane when elevation = 0 degree. The colours refer to different response along different azimuth angles(locations)

3.7.2 IRCAM LISTEN HRTFs Dataset

The second HRTF data set is known as the IRCAM-Listen HRTF database (Goodman and Brette 2011). This data set consists of a general purpose HRIRs measurements for 51 different subjects. Subject 1002, a male human subject, is selected for all experiments in this research. This data has 187 locations which are referred to different HRTF containing different elevation and azimuth measurements. Loudspeaker is moved by a U-shaped structure called crane which is made from metal that completely enveloped with melamine panels. The crane has been elevated by a couple of step-by-step motors controlled by the computer. A measurement software was used to choose the elevation angle and an angular sensor is used to send a feedback. The elevation values in the range of -45 to 90 degrees and azimuth within the range 0 to 360with 15° regular step increments. Table 3-2 shows the dimensionality of IRCAM- Listen HRTF data set involving the number azimuth measurements for each elevation and the sample frequency is 44.1 KHz (Blauert 1997).

Table 3-2: IRCAM LISTEN HRTF database number of measurements and azimuth increment at each elevation.

Elevation	Points Per elevation	Azimuth Increment Per elevation (degree)
-45	24	15.00
-30	24	15.00
-15	24	15.00
0	24	15.00
15	24	15.00
30	24	15.00
45	24	15.00
60	12	30.00
75	6	60.00
90	1	360.00

The following figures show samples of impulse responses from the left and right from IRCAM Listen database in particular directions which are presented for data visualization purposes.

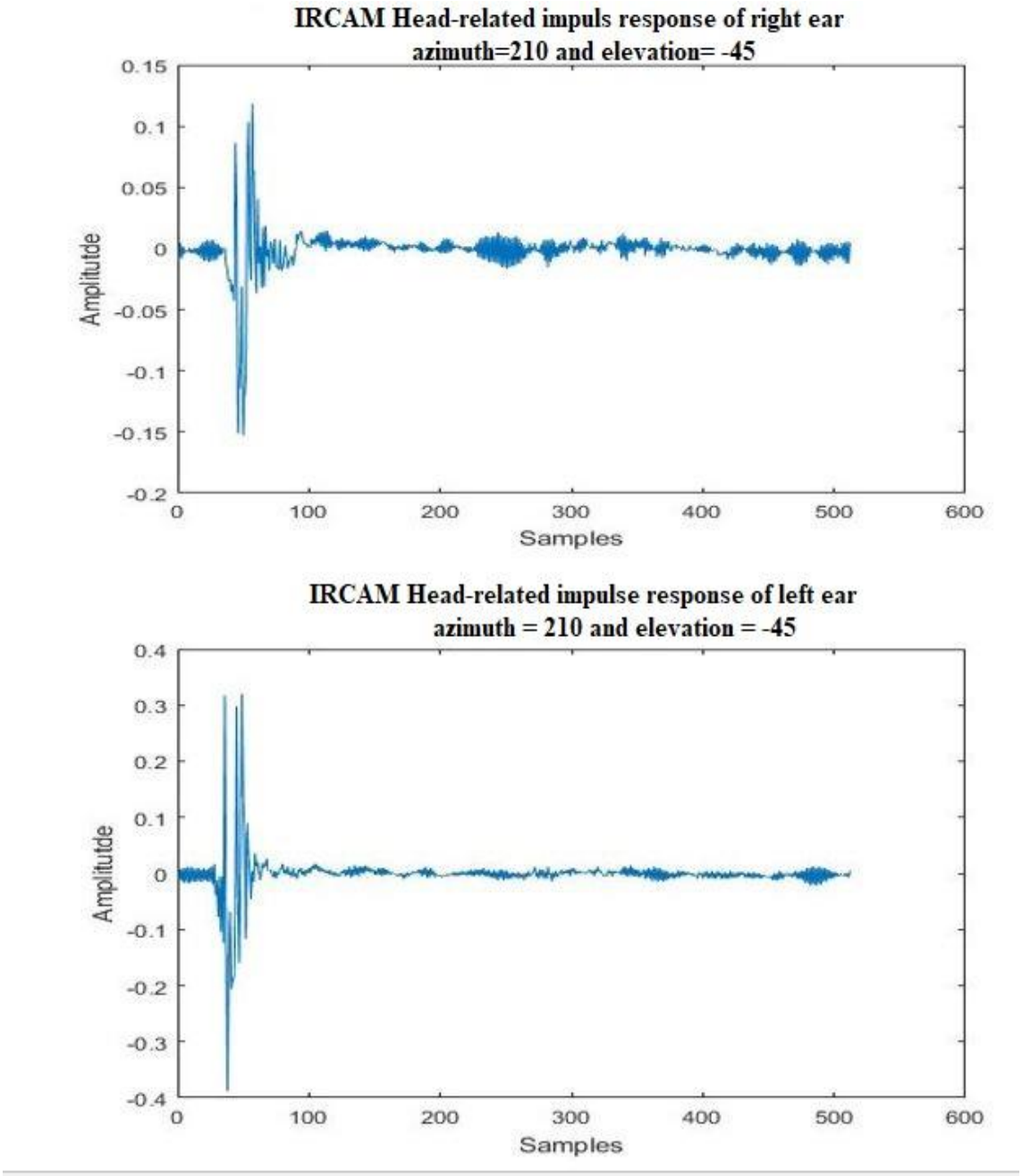


Figure 3.10: Head related impulse responses for IRCAM subject (left and right ears) in the time domain.

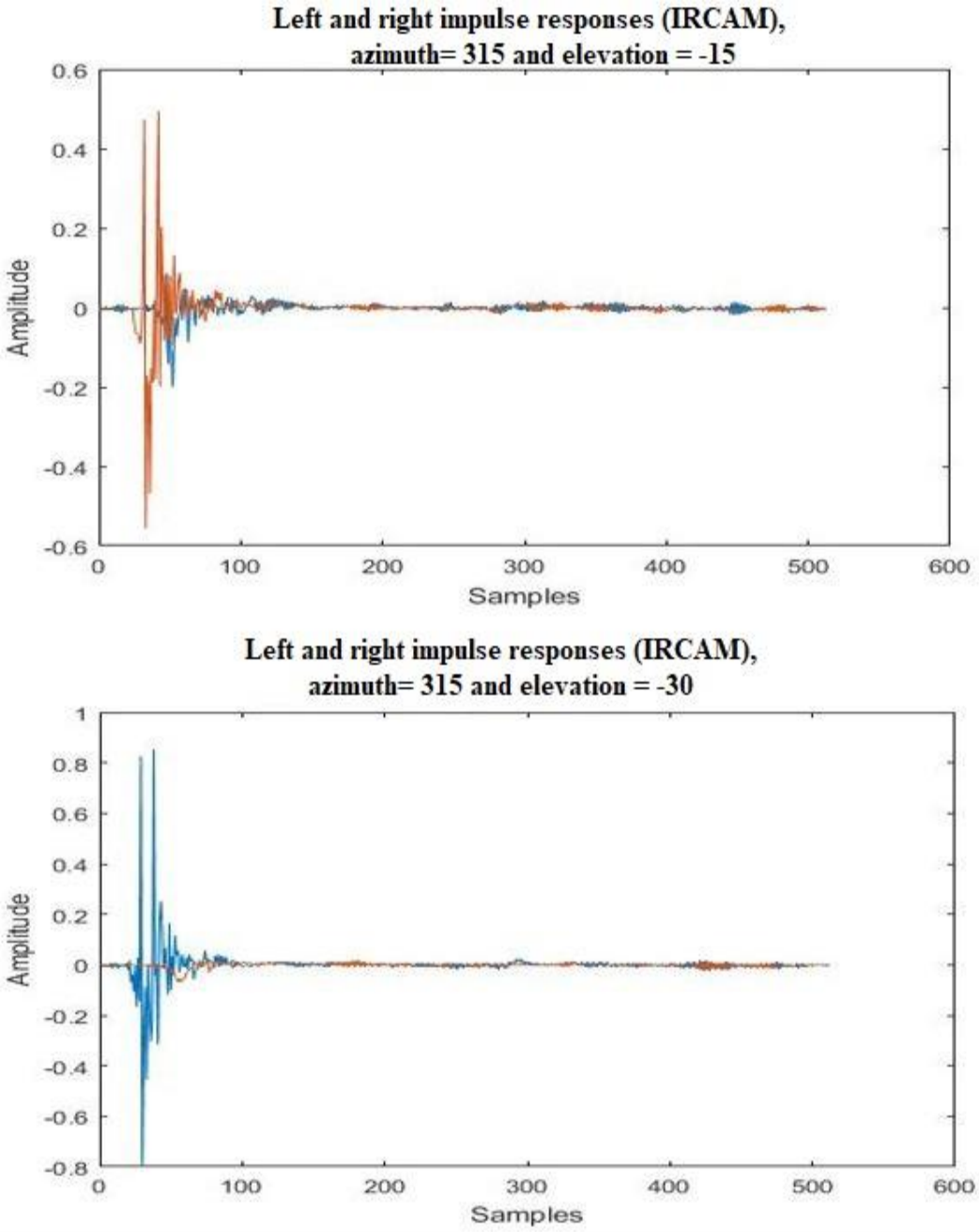


Figure 3.11: Plots of the pair of impulse responses from particular directions of IRCAM selected subject.

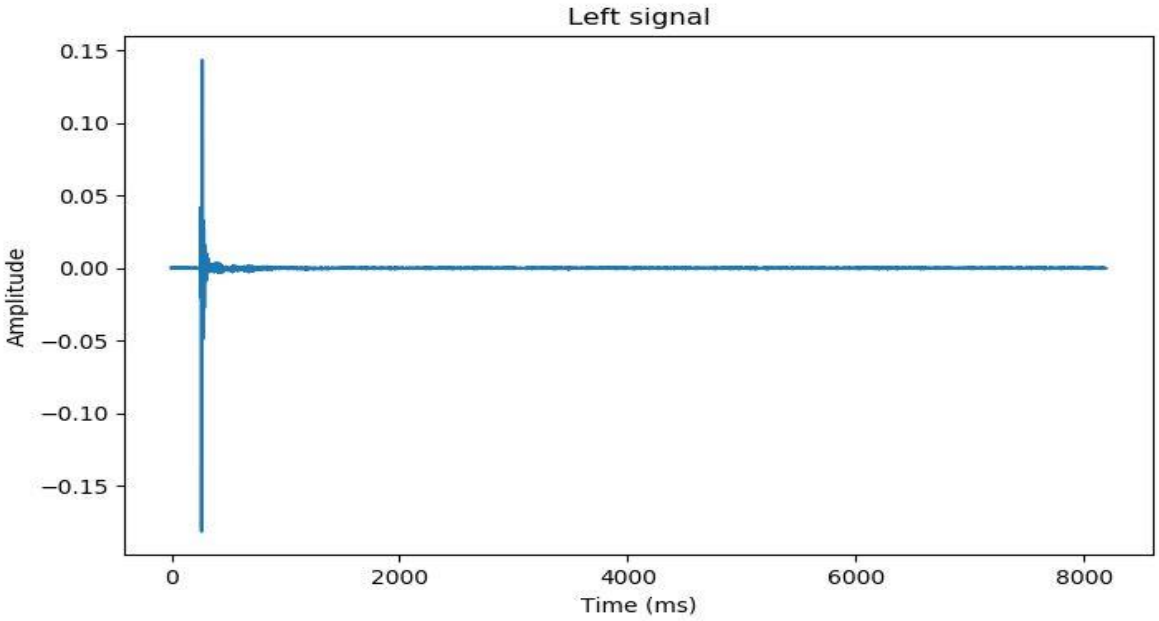


Figure 3.12: Impulse responses in the time domains from left ear of IRCAM, azimuth = 0 degree and elevation = 60 degree.

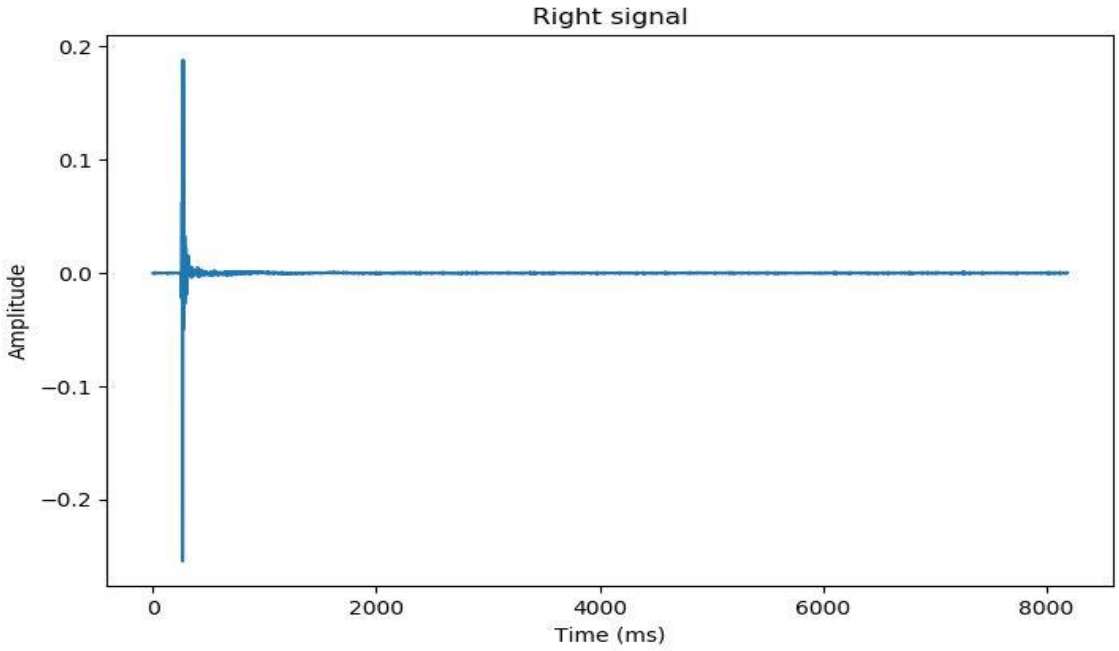


Figure 3.13: Impulse responses in the time domains from right ear of IRCAM, azimuth = 0 degree and elevation = 60 degree.

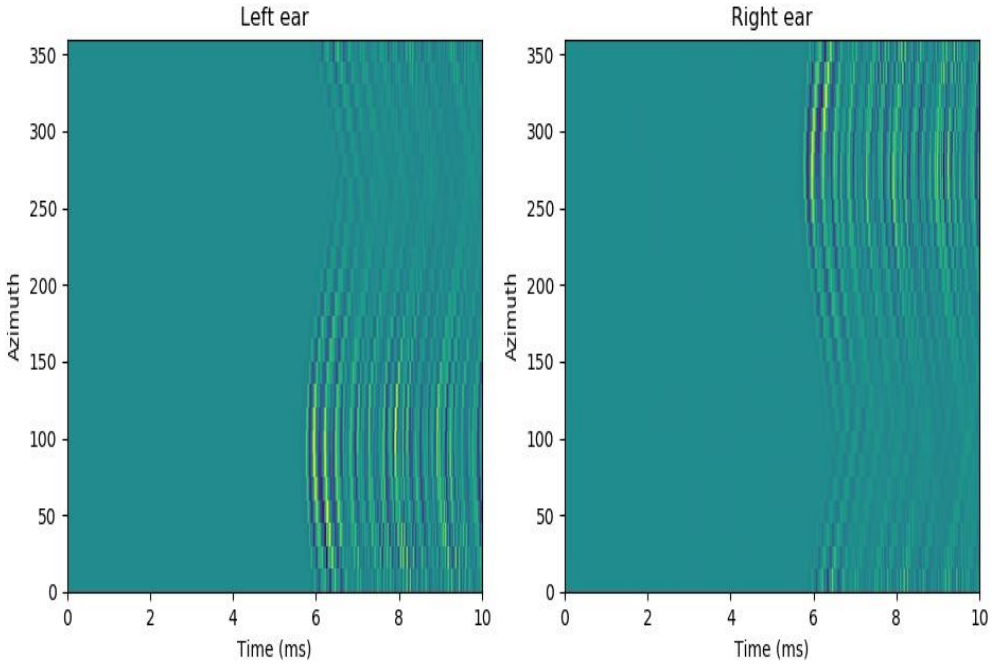


Figure 3.14: Image illustrates the impulse response of IRCAM database in the horizontal plane when elevation = 0 degree.

3.7.3 Speech Databases

The speech database SALU-AC (Al-Noori 2017, Al-Noori et al. 2015) recorded at the University of Salford was used. The data includes variety of speech samples of English spoken by native and non-native speakers in addition to speech samples in different languages. These speech samples were recorded in an anechoic enviroment. The objective in using this data is the efficient testing of sound source localization models due to the variety of speech samples belongs different languages. The main features of this database can have summarized as following; It contains very clean speech samples because it was recorded in an anechoic chamber, it contains English speech samples were recorded from 55 males and 55 females, it included a variety of speech sample recorded from English native and non-native speakers, the time duration for each speech sample is 5 seconds with sample rate 16 kHz. This provides flexible sound samples and durations for different sound source localization experiments. The SALU-AC has 110 different speakers, 55 of males and 55 of females recorded at 16 kHz sample rate.

3.7 Chapter Summary

1. This chapter has focused mainly on giving the background to binaural hearing and sound source localization. The mathematical description of HRTFs as acoustical filters and their basic components have been demonstrated.
2. Also focused on explaining the principles of SNN and demonstrated the principle of encoding and decoding spike trains.
3. Furthermore, this chapter has given a description of the machine learning methods that are used in this research for single and multisource sound localization. These methods include DNN and SVM.
4. The chapter covered briefly some computational concepts behind these two machine learning approaches as well as the most commonly used learning algorithm (backpropagation).
5. Finally, this chapter has given a description of the speech database and the HRTF databases that contributed to testing different sound source localization methods in this research. In addition to the basic characteristics of KEMAR dummy head HRTFs, IRCAM Listen HRTFs and speech data set that is used in this research.

CHAPTER 4
SINGLE-SOUND SOURCE LOCALIZATION PROPOSED MODEL

Chapter Overview

This chapter focuses mainly on explaining the structure of the sound source localization model that is suggested by Goodman and Brette (2011). Single sound source localization modelled using a spiking neural network is replicated by using KEMAR dummy head-related transfer functions. The model component description is shown in section 4.1. The experiments and the outcomes of investigating the performance of the localization model using various kinds of input sound signals including two types of white noise signals and various speech samples are examined in section 4.2. This work is different from Goodman and Brette by investigate the model performance to localize the sound sources under different conditions; researching the localization performance at single and octave frequency. Also, section 4.3 shows the impact of varying levels of environmental noise with different signal-to-noise ratios (SNRs) on the robustness of the single sound source localization model. Comparison between spiking neural networks (SNN) and other machine learning methods (support vector machine (SVM)) for single source localization is shown in section 4.4. Multisource localization by using SNN based localization model is presented in section 4.5. A motivated localization model based on applying the spiking neural networks as pre-processing method integrated with different machine learning methods is presented in section 4.6.

4.1 Spiking Neural Networks localization model

SNNs can process and account for time delays in signals; a key feature of third generation models when compared with previous approaches (Yu et al. 2016, Diaz et al. 2016). This feature is essential in spatial signal processing, as much of the information is encoded in the interaural time difference and interaural phase shifts of different frequency components. Therefore, the current work attempts to explore the suitability of the spiking neural network model as a signal feature extraction tool to provide information on sound source localisation from binaural signals. Previous methods performed well when one source was present, but performance was poor when multiple simultaneous sources were present as will explaining in the current chapter. Therefore, like the way in which our brains have learnt to interpret the neuron firings from the auditory nerve, a supervised learning algorithm is trained to process the firing rate from the SNN and learn to perform multi-source localisation as a novel idea to solve multisource separation and detection problems. This will clearly show in chapter 5.

4.1.1 Single-sound source localization model (SSL)

Work conducted by Goodman on single sound source localization was replicated and tested to investigate its ability to localize single and multi-sound sources (Goodman and Brette 2011). A simple spiking neural model was designed to predict the location (azimuth and elevation) of a single sound source in spherical coordinates. The location estimation process relies on spatial-temporal filtering and spiking nonlinearity. Figure 4.1 explains the sequential steps of Goodman model that was applied on an input signal to analyse the binaural information. The measured HRTFs simulate the acoustical filtering of the source signal that received by the two ears. The first application of the HRTF is generating a simulation of a signal capture by a dummy head. Then there is another application of the HRTF, this time for all possible angles. The localization model in the simulation stage simulates all possible HRTF pairs from data sets. The left and right channels of HRTF are reversed to decrease ITD and ILD impact on the received signal and make a left and right signal close to identical at (0, 0) which is the receiving end of the incoming sound signal. The left and right channels' reversal supports the training stage which does not take account of the gain and delay in teaching the spiking model to localise the sound source.

A set of gammatone filters, that simulate the auditory pathway (cochlea), are applied on the resulting signals, followed by neural filtering. The two monaural signals were transferred to the cochlear for analysing into multiple frequency bands. The filtered signals are transferred into spike trains by the monaural neurons which form the spiking neuron model (in this work a leaky integrate and firing model is used). These spike trains are the input to the neurons in the second layer (binaural neurons) which are coincidence detector neurons. The binaural neurons fire when receiving coincident inputs. When two neurons are firing concurrently they are linked together, and then the weight of their connection will influence the action potential to cross the neuron threshold value and firing spike. A sound source location is detected by analysing the output of the coincident neurons for each location. The coincidence detection neuron outputs refer to the synchrony fields of their inputs that contain that location. Location-specific synchrony samples are thus matched to the activation of neural assemblies which equivalent the sound directions (azimuth, elevation).

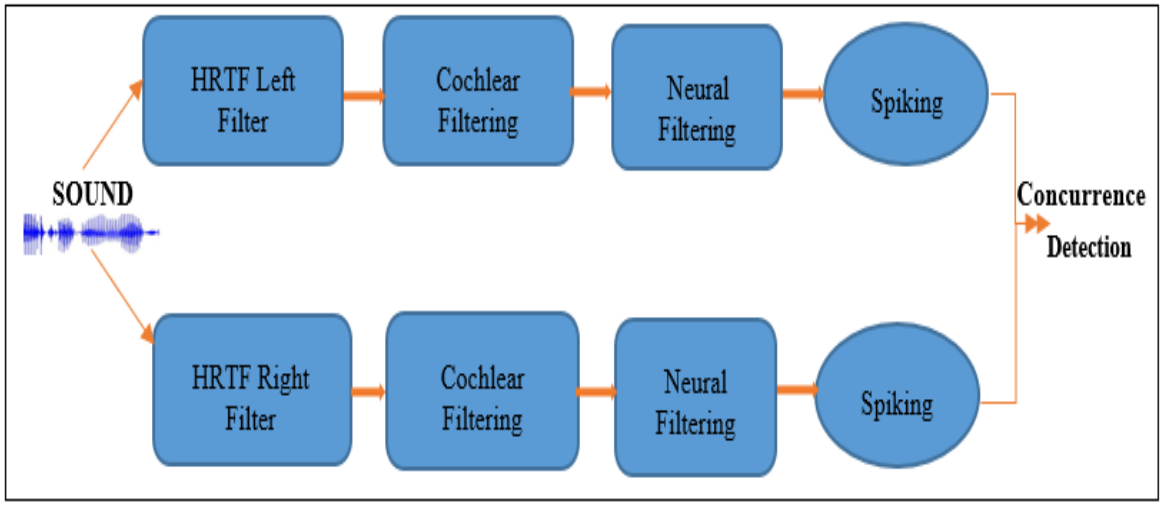


Figure 4.1: Sound source localisation model (Goodman & Brette model)

The sound signal S(t) emanates from sound source at a given azimuth (θ) and elevation (∅). The signals present (S_L(t) and right S_R(t)) at each ear can be simulated by the embedding of S(t) with two linear filters:

$$\boldsymbol{S_L = HRIR_L(\theta, \emptyset) * S, S_R = HRIR_R(\theta, \emptyset) * S} \qquad (\mathbf{4.1})$$

The Goodman model embeds the HRTF with the left and right filters swapped for each angle. This has the effect of removing the ITD for the filter pair representing the true source angle.

$$\boldsymbol{S_{N,L}(\theta_j, \emptyset_k) = HRIR_R((\theta, \emptyset)) * S_L,\ S_{N,R}((\theta, \emptyset)) = HRIR_L((\theta, \emptyset)) * S_R} \qquad (\mathbf{4.2})$$

The $S_{N,L}((\boldsymbol{\theta}, \emptyset))$ and $S_{N,R}((\boldsymbol{\theta}, \emptyset))$are then used as inputs to individual spiking neurons. The spectro-temporal receptive field of the neuron (STRF) define as a filter works to save the incoming sound signal frequency and time representations in to the matrix and then filter them into spike trains. Equation 4.3 describes the left and the right sound signals that are filtered through the neuron's spectro-temporal receptive field (STRF) to transform into the signals into spike trains. STRF can reasonably estimate the responses of auditory nerve neurons from new a stimulus (Zhao and Zhaoping 2011). Spectro-temporal receptive fields (STRFs) can be defined as linear approximations of the signal transform from sound waves to neural responses along the auditory pathway. STRFs depend on the ensemble of incoming stimuli and this has been investigated mechanically and computationally as a potential composite nonlinear process (Kim and Young 1994).

$$\boldsymbol{S_{N,L}(\theta, \phi) = N_A * HRIR_L(\theta, \phi) * S, S_{N,R}(\theta, \phi) = N_B * HRIR_R(\theta, \phi) S} \qquad (\mathbf{4.3})$$

N is the spectro-temporal receptive field for a given input signal. Generally, the received signals at the two ears are filtered transformations of the source signal. In the acoustic environment, the filters are specified by the head and source relative positions. When the sound is filtered through the spectro-temporal receptive field of neuron N, it is converted in to a spike train. The spike trains are generated by the leaky-integrate-and-fire(LIF) algorithm.

In the integrate-and-fire model structure, each presynaptic spike produces a postsynaptic current pulse. More accurately, if *j* represents the presynaptic neuron released a spike at time $t_j^{(f)}$ and a postsynaptic neuron *i* received a current with time cycle (t - $t_j^{(f)}$). The input current at neuron *i* is computed by the summation the total current pulses as explained by equation 4.5(Paugam-Moisy and Bohte 2012):

$$\boldsymbol{I_i(t) = w_{ij}\left(t - t_j^{(f)}\right)} \qquad (\mathbf{4.5})$$

The operator w_{ij} represents a measurement of the synaptic effectiveness between neuron *j* and neuron *i*. The above formula considers a synaptic interaction model.

The spiking neural network is constructed from two fully connected layers. The monaural input neurons are connected to a second layer of neurons referred to as coincident detection neurons. Each coincident detection neuron has two inputs from two of the input neurons. There are as many coincident neurons as there are angles in the embedded HTRF. The model is based on the principle that neurons synchronize when their inputs are similar. When the firing of the two inputs are similar, the firing rate of the coincident neuron will be high. This indicates a strong correlation between the signals at both ears and thus a high likelihood that the sound originated at the angle represented by that input pair.

A bank of Gammatone filters (GFs) is implemented in our work to simulate the frequency resolution of human hearing cochlea. Gamma-tone filters take the form of cascades of four 2nd-order IIR filters corresponding to a 4th-order gamma-tone filter (Slaney 1993). A gamma-tone filter (GF) can be statistically described by equation 4.6 as a shape of impulses responses in the time range

$$\boldsymbol{g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \theta)} \qquad \textbf{(4.6)}$$

When t > 0, the symbol '*a'* indicates a constant which is responsible for regulating the gain, with θ representing phase which is normally set to 0. The filter bank is patterned as group of equivalent bandpass filters, each band limited to an independent frequency (Slaney 1993). The fundamental parameters of the gamma-tone filters are *b* and *n*. The *b* depends on the value specified for the duration of the impulse response; n refers to the order of the filter. This is generally accepted as analogous to the magnitude response of the human auditory filter. The human data summarized on the equivalent rectangular bandwidth (ERB) of the auditory filter by applied the following function:

$$\boldsymbol{ERB = 24.7 + 0.108 * fc} \qquad (\mathbf{4.7})$$

f_c represents the centre frequencies of the bands that make up the filter bank. Gamma tone filter banks can be used to acquire signal features at different frequency levels. (Ma, et al, 2015). However, the literature suggests that the equivalent rectangular bandwidth (ERB) provides more accurate estimation of the auditory filter bandwidth (Singh et al., 2012).

The number of neurons is increased, such that for each frequency and angle there is an input neuron pair and a coincident neuron. The maximum firing rate for the active neurons assembly are computed to indicate the source location. Goodman used a winner-takes-all approach, where the azimuth and elevation of the neuron with the maximum firing rate is taken as the optimal prediction.

4.2 Experiments and results

As mentioned in the previous section, a single source localization model by Goodman and Brette (2011) is replicated to investigate the model performance in different conditions. Two HRTF data sets were used to test the model for single source localization in different conditions. These datasets have different anatomical parameters (head size, ear shape and torso), this improves the model generalisation over previous work, which was limited to a single torso simulator. In the following, the performance of the spiking neural model as a single sound source localization model with IRCAM and KEMAR HRTF datasets was examined. The single sound source localisation framework trained by using 710different azimuth and elevation combinations existing in the KEMAR HRTF data set. In addition, it was trained using 187 various locations in IRCAM HRTF data set. The model localization performance was investigated under various condition; different types of sound signals, different frequency levels, a variety of input signal durations (100ms - 500ms), and noisy signals with different SNRs. In all experiments, the signed angle error is computed. The angle error between the actual and predicted angles is computed by finding the difference between the true angle and the predicted angle for azimuth and elevation. The localization accuracy for azimuth angles is also calculated by finding the average of angles that were predicted correctly with 0˚or 15˚angle error over the total number of points that participated in the model validation stage according the equations 4.8 and 4.10. while, the elevation angle localization accuracy is calculated by finding the average of angles that were predicted correctly with 0˚ or 10˚angle error over the total number of points that participated in the model validation stage according the equations 4.9and 4.11.

$$\boldsymbol{Azim\ L.Acc_{IRCAM}} = \frac{\boldsymbol{number\ of\ azimuth\ angles\ that\ predicted\ correctly\ \pm 15^{\circ}}}{\mathbf{187}}. \qquad \mathbf{(4.8)}$$

$$\boldsymbol{Elev\ L.Acc_{IRCAM}} = \frac{\boldsymbol{number\ of\ elevation\ angles\ that\ predicted\ correctly\ \pm 15^{\circ}}}{\mathbf{187}}. \qquad \mathbf{(4.9)}$$

$$\boldsymbol{Azim\ L.Acc_{KEMAR}} = \frac{\boldsymbol{number\ of\ angles\ that\ predicted\ correctly\ \pm 15^{\circ}}}{\boldsymbol{710}}. \qquad \textbf{(4.10)}$$

$$\boldsymbol{Elev\ L.Acc_{KEMAR}} = \frac{\boldsymbol{number\ of\ elevation\ angles\ that\ predicyed\ correctly \pm 10^{\circ}}}{\boldsymbol{710}}. \qquad \textbf{(4.11)}$$

in above equations, the 15˚ refers to the minimum increments step in the azimuth measurements of IRCAM dataset while the minimum increment step for elevation measurements is 10˚. For KEMAR database, the azimuth increment steps are irregular with minimum step 5˚. So that, the$\boldsymbol{Azim\ L.Acc_{KEMAR}}$ consists of all angle errors in range 5˚ to 15˚. While the elevation measurement has regular increments step with 10˚. In the following, various experiments are carried out to investigate localization performance under different conditions.

4.2.1 Testing distinct types of input sound signals

The single sound source localization performance is investigated with different types of sound signals. This experiment included examining a variety of generated sound signals which are embedded in various locations from KEMAR and IRCAM HRTFs data sets. Firstly, the model is tested with diverse types of white noise input signals including gaussian white noise (GWN) and uniform white noise (UWN). GWN returns samples from the "standard normal distribution" while UWN represents to the random samples from a uniform distribution. Figure 4.2. shows samples of Gaussian and uniform white noise input signal embedded with IRCAM HRTFs. The model is tested with different samples of Gaussian and uniform white noise with 300 ms samples duration.

Secondly, sinewave modulated white noise (SMW) sound signals are used to test the localization method. The process of modulation denotes regularly employing the information signal to modify some parameter of the carrier signal. The carrier signal is usually only a simple, single-frequency sinusoid (modified in time such a sinewaves). This type of modulation method is called amplitude modulation (AM) that is used in electronic communication to convey information through a radio carrier wave. The signal strength of the carrier wave is varied in proportion to that of the message signal being transmitted (Peterson, Smith et al. 1996). To simplify the modulation process, 100% modulation has been used in this experiment as shown in figure 3.4. It refers to the maximum possible amount of modulation when the level of

modulation can be increased to a level where the envelope falls to zero and then rises to twice the un-modulated level.

The importance of this experiments can be summarized by the advantaged of using modulation in the communication system. One of the advantages of modulation is the multiplexing that refer to the ability to transmit two or more signals over the same communication channel at the same time. Hence, the modulation helps to Avoids mixing of signals. If the baseband sound signals are transmitted without using the modulation by more than one transmitter, then all the signals will be in the same frequency range i.e. 0 to 20 kHz. Therefore, all the signals get mixed together and a receiver cannot separate them from each other. Hence, if each baseband sound signal is used to modulate a different carrier then they will use different channels. Thus, modulation avoids mixing of signals.

In this experiment, sinewave is used as an envelope signal at 5Hz and 0.3s duration and the white noise of 0.3s duration is the carrier signal to examine its influence on localization performance. The 5Hz refers to the lowest modulation frequency of baseband signal. The low frequency signals unable to travel long distance when they are transmitted. They get heavily attenuated. The attenuation reduces with increase in frequency of the transmitted signal, and they travel longer distance. The modulation process increases the frequency of the signal to be transmitted. Therefore, it increases the range of communication. The modulation enhances the communication signals by overcome on all transmission limitations. This experiment helps to investigate the localization performance with amplitude modulation signal that required limited bandwidth and low frequency carrier.

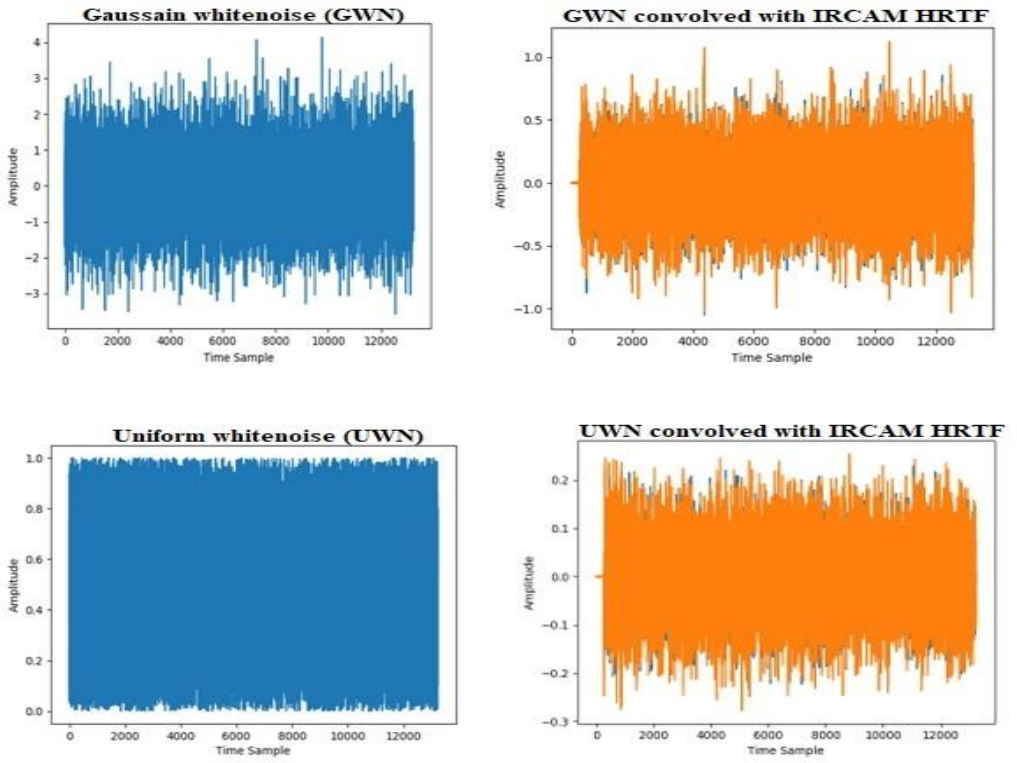


Figure 4.2: Gaussian and Uniform white noise input signal convolved with IRCAM HRTFs.

Figures 4.3 and 4.4 show the form of amplitude modulated signal convolved with KEMAR and IRCAM HRTFs that used to test the single source localization model.

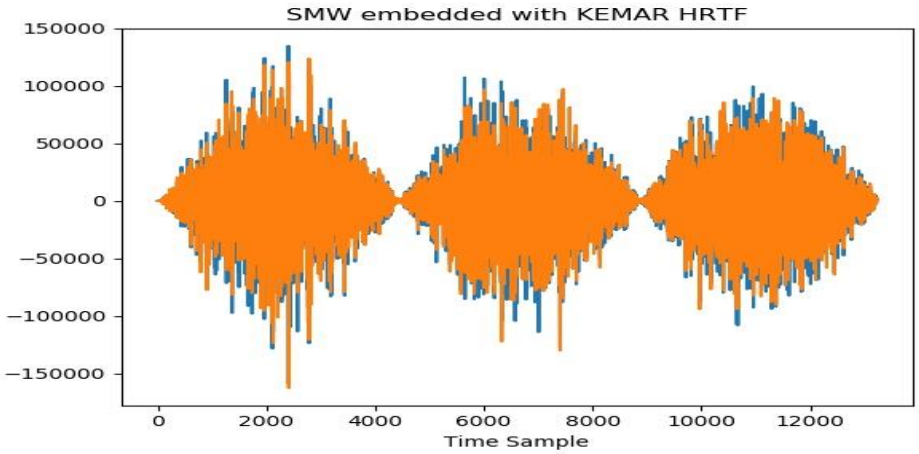


Figure 4.3: Sinewave modulated white noise signal input signal convolved with KEMAR HRTFs.

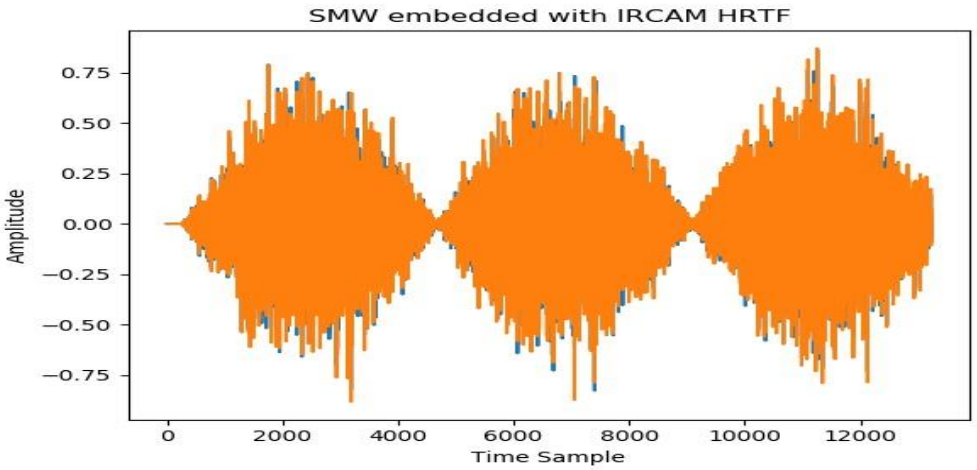


Figure 4.4: Sinewave modulated white noise input signal convolved with IRCAM HRTFs.

Thirdly, clean speech samples were convolved with binaural responses were used to test the Spiking neural network localization model for sound source localization. In this experiment. Different speech samples in different languages were used to investigate the localization model performance with realistic sound samples and prepare it for real-time application. The localization model was investigated with different speech samples with 0.3s duration including different utterances of 10 speakers (5 male and 5 female).

The experimental outcomes of applying these diverse types of artificial sound signal and real speech samples are explained in figure 4.5. It's noticeable that these different sound inputs have different level of impacts on the localization model performance. Quantitatively, for the localization model of 40 Gamma-tone frequency channels, the average estimation error for speech signal was the less compared with the other types of input signal. Table 4-1 explains that the azimuth angle estimation accuracy ±15˚ for IRCAM and KEMAR. Whilst, the elevation angle estimation accuracy is ±15˚ for IRCAM and it is ±10˚ for KEMAR. The results show that with the IRCAM model for speech 93% of predictions are within 15˚(azimuth) and 91% for elevation. And, it shows that with the KEMAR model for speech 89% of predictions are within 15˚ (azimuth) and 87% for elevation. UWN has higher localization Accuracy compared to GWN and SMW, this is because uniform white noise (UWN) represents random samples from a uniform distribution while gaussian white noise (GWN) represents samples from the standard normal distribution. In case of SMW, amplitude modulation has low efficiency in term of its use of power and spectrum. whereas, the average estimation error is slightly increased for

azimuth and elevation angles from KEMAR HRTFs. One of the assumptions it is because it was trained with a database that offered a finer resolution. And, the other assumption, it is the dimensionality of KEMAR data set that contains large variety of measurements for azimuth and elevation that represents different locations along the horizontal and vertical planes.

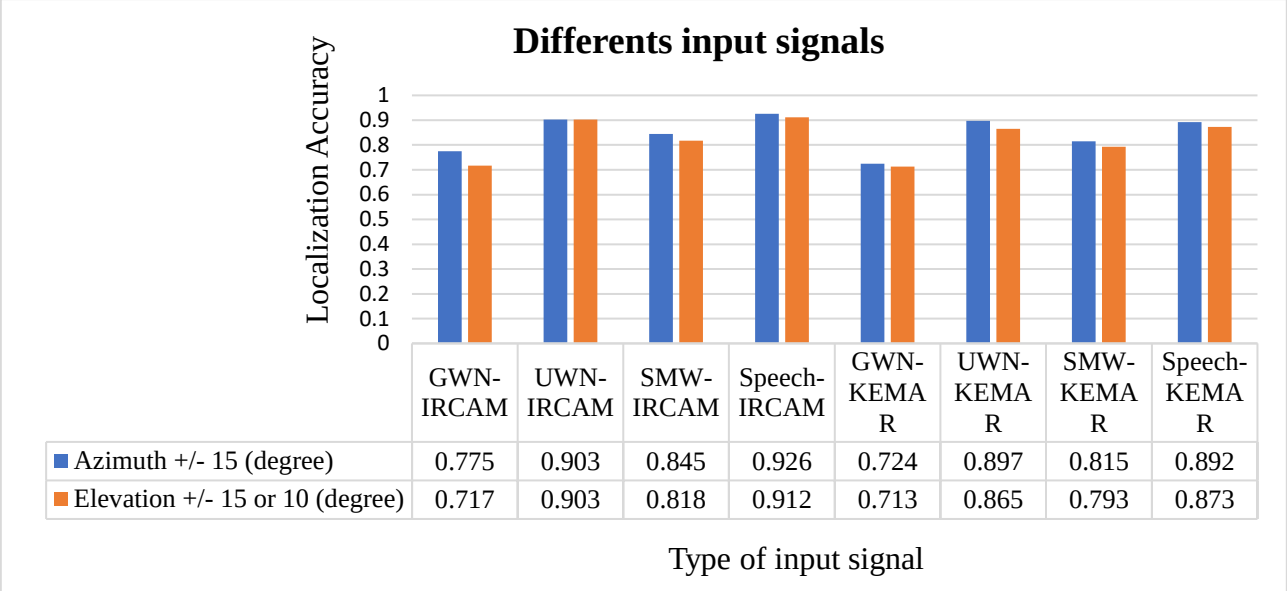


Figure 4.5: Comparison between GWN, UWN, SMN and speech types of input signal effectiveness on the Azimuth and elevation estimation accuracy.

Table 4-1: The experimental results from applying SNN localization model for different types of inputs signal with both KEMAR and IRCAM HRTF databases.

Type of input signal	Azimuth Angle Estimation Accuracy (+/- 15˚) IRCAM	Elevation Angle Estimation Accuracy (+/- 15˚) IRCAM	Azimuth Angle Estimation Accuracy (+/- 15˚) KEMAR	Elevation Angle Estimation Accuracy (+/-10˚) KEMAR
Gaussian white-noise	0.775	0.717	0.724	0.713
Uniform white-noise	0.903	0.898	0.897	0.865
Sine wave modulated white-noise	0.845	0.818	0.815	0.793
Speech	0.986	0.982	0.956	0.948

4.2.2 Testing different frequency ranges

The sound source localization model was investigated over different frequency ranges by applying it with different single frequencies and octave frequency. The single frequency takes the frequency in the range (63 Hz, 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz). An octave refers to the interval between one frequency and its double or its half. For example, there is one octave band between frequencies 1 000 Hz and 2 000 Hz. There is another one octave band between 1 000 Hz and 500 Hz. An octave frequency is applied to increase the resolution of the received signal at two ears as shown in figure 4.7. The localization model with40 Gamma-tone frequency channels is tested with tone signals of 0.3s duration over these various frequency ranges of single and octave frequency. Figures 4.6 and 4.7 visualize the shape of the tone input signal with single frequency 63 Hz and octave frequency that embedded with the binaural signal from KEMAR and IRCAM HRTF databases.

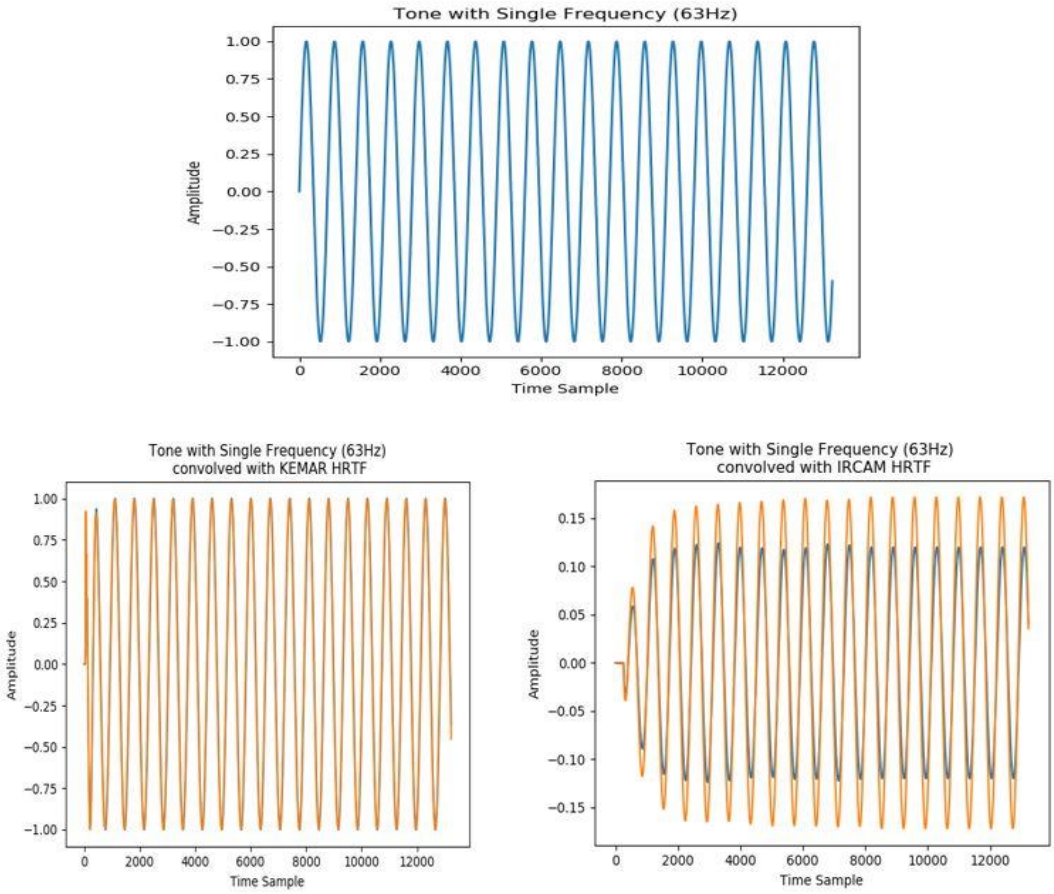


Figure 4.6: Sinewave of single 63 Hz embedded with KEMAR and IRCAM HRTF data sets.

The sampling frequency is 44.1 KHz, the sound signals with the duration of 0.3 seconds will have 13230 samples. The sample rate refers to the number of samples per second in a sound signal.

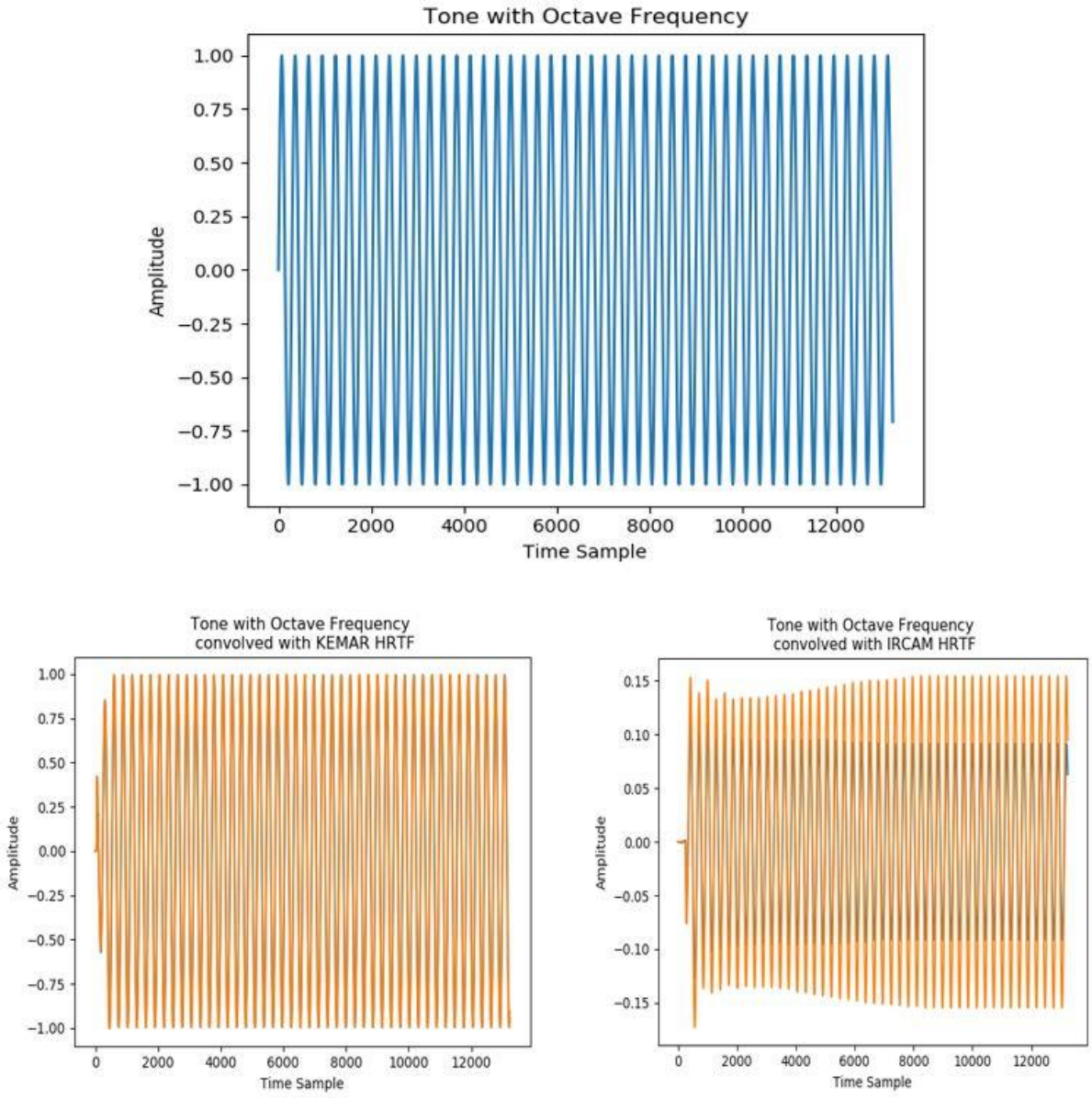


Figure 4.7: Sinewave of octave frequency embedded with KEMAR and IRCAM HRTF data sets.

The localization model was tested with these various levels of single frequencies and octave frequency to investigate the localization cues impact on localization performance. The signed angle error between the actual and estimated angles for both azimuth and elevation is

computed. The estimation accuracy is figured from finding the ratio of correctly predicted angles (the angles that have 0˚ and 15˚ angle error) to the total numbers of locations in the HRTF data set. Figures 4.8 and 4.9 show the estimation accuracy of azimuth and elevation angles for both HRTF data set with pure tones of single frequency.

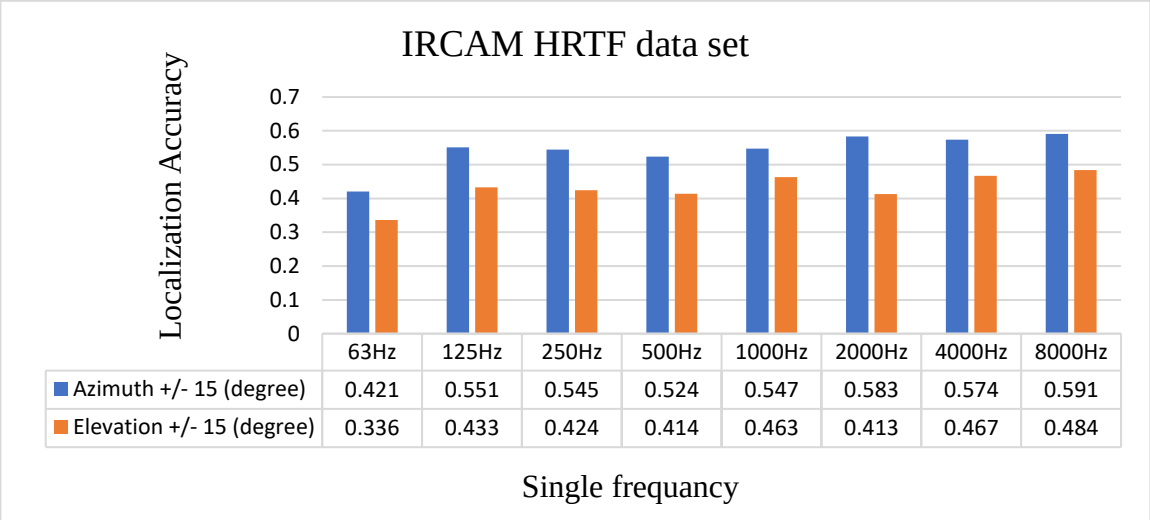


Figure 4.8: Azimuth and elevation angles estimation Accuracy with pure tones with single frequency for IRCAM HRTF data sets explains the model performance in different range of frequency.

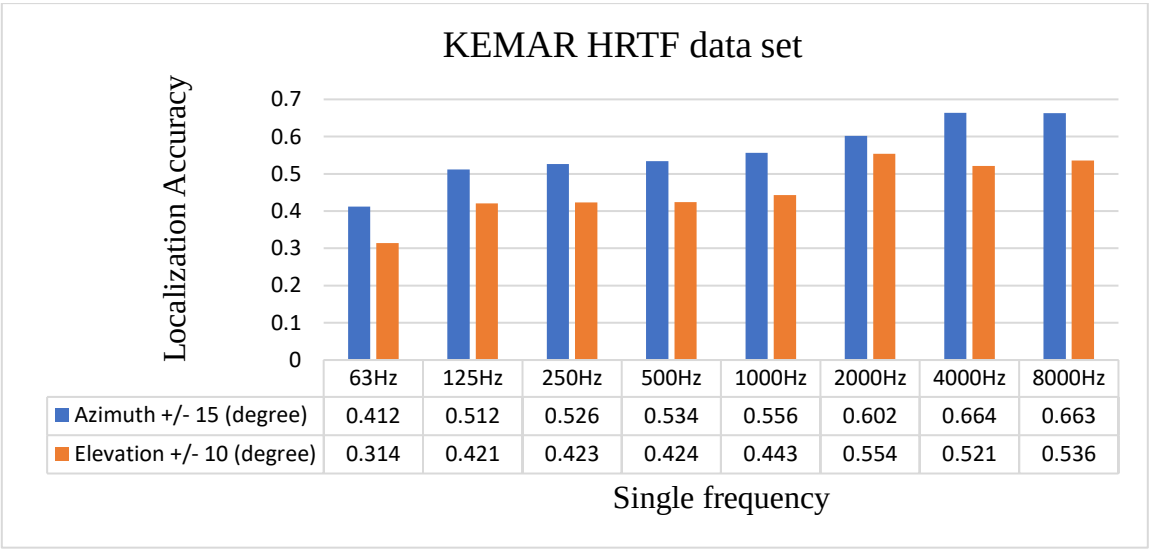


Figure 4.9: Azimuth and elevation angles estimation Accuracy with pure tones with single frequency for KEMAR HRTF data sets explains the model performance in different range of frequency.

Figure 4.10 and 4.11 demonstrates the estimation accuracy of azimuth and elevation angles for both HRTF data set with pure tones with octave frequency. There is a clear improvement in the localization performance compared with of single frequency.

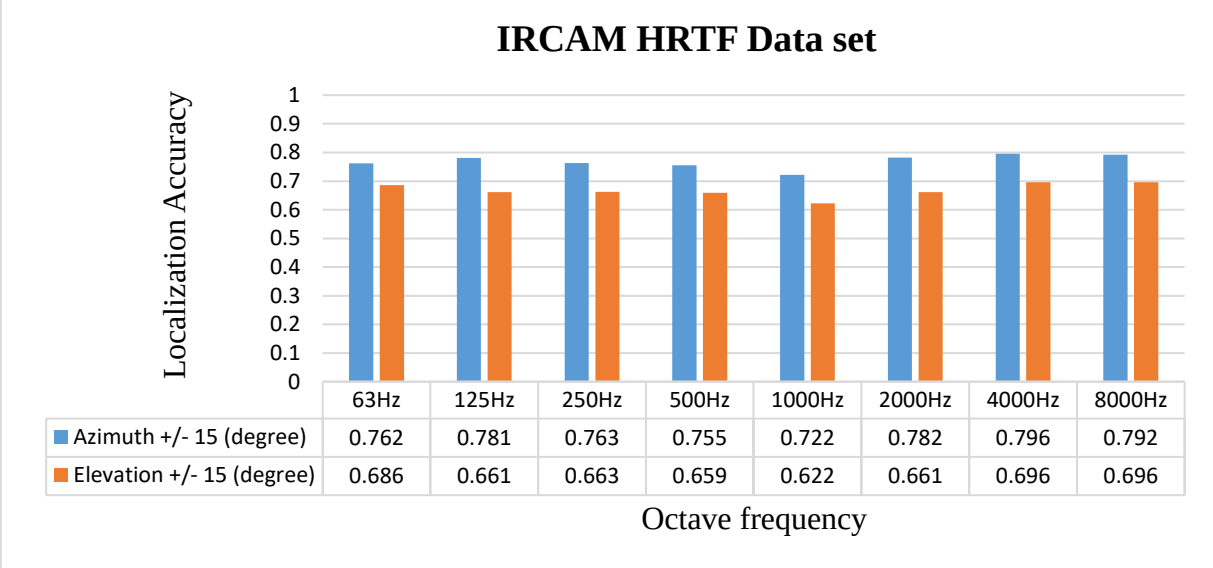


Figure 4.10: Azimuth and elevation angles estimation Accuracy with pure tones of octave frequency for IRCAM HRTF data sets.

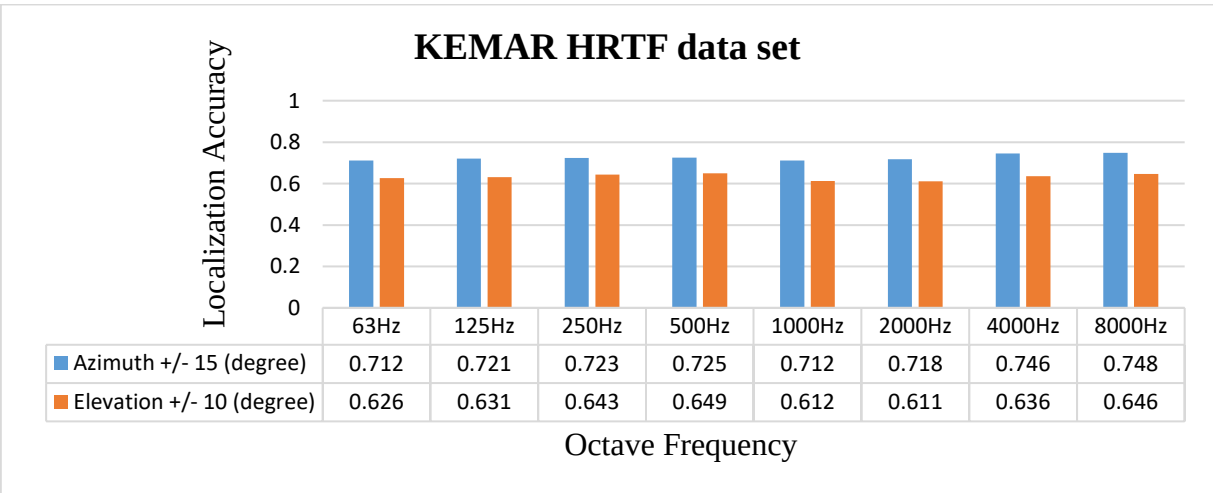


Figure 4.11: Azimuth and elevation angles estimation Accuracy with pure tones of octave frequency for KEMAR HRTF data sets.

As known, each localisation cues type plays a significant role to localise sound in certain frequency range. And to investigate the impact of each individual localization cues type on the

localization model performance, different hearing frequency were tested in this experiment. And as shown in the previous figures, the average estimation error was higher for pure tones, especially for elevation angles. That occurred because the ITD cues are ambiguous due to periodicity in the high frequency domain, and ILD cues are powerless in the low frequency domain, presenting just one dimension in the binaural cues. This experiment was as an evidence of the importance of using the full HRTF cues for sound signal localization rather than use only ITD or ILD to localize the sound signals.

4.2.3 Testing the signal duration and number of Gamma-tone frequency bands

From the experimental results above, two training parameters appear to highly influence of the localization model performance; the signal duration and gamma-tone filter bank number of channels. Firstly, the effectiveness of incoming sound signal duration on the model performance is examined by testing different sound lengths varied from (100*ms* to 500*ms).* Table 4-2 explains the impact of the incoming signal duration on the azimuth and elevation angles estimation accuracy. The speech sound samples were used in this experiment to consider the required signal duration for a better localization performance. The selected sound signal duration will then use as a fixed value in the upcoming tests. The experimental findings demonstrate that the localization model needs no less 500ms signal duration for a better performance over different conditions as shown in figure 4.12.

Table 4-2: Azimuth and elevation angles estimation accuracy under different lengths of input signals.

Signal duration in second	Azimuth Angle Estimation Accuracy (+/- 15˚) IRCAM	Elevation Angle Estimation Accuracy (+/- 15˚) IRCAM	Azimuth Angle Estimation Accuracy (+/- 15˚) KEMAR	Elevation Angle Estimation Accuracy (+/- 10˚) KEMAR
0.1	0.807	0.791	0.815	0.778
0.2	0.871	0.85	0.871	0.832
0.3	0.954	0.949	0.924	0.911
0.5	0.996	0.992	0.965	0.954

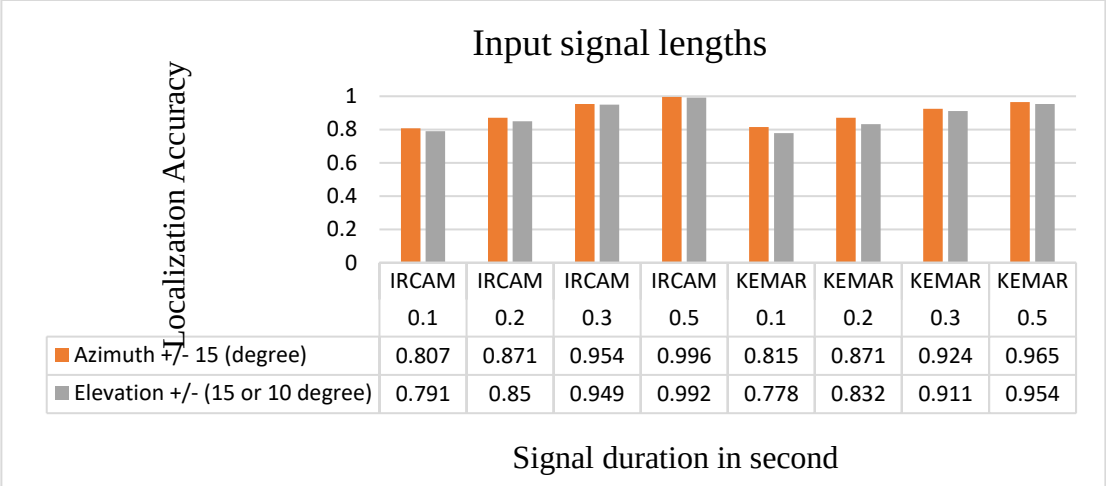


Figure 4.12: The impact of input signal duration on localization model performance.

To study the gamma-tone futures selection, the number of channels of gamma-tone filter bank was also investigated to test its impact on the localization performance. Table 4-3 demonstrates a notable improvement of the model localization performance with 80 gamma-tone frequency bands. However, increasing the gamma-tone frequency channels cause a clear impact on the localization model execution time, particularly, for KEMAR binaural signal. Figure 4.13 shows the azimuth and elevation estimation accuracy ±15˚ for IRCAM and azimuth±15˚ and elevation ±10˚ estimation accuracy for KEMAR binaural signals.

Table 4-3: Azimuth and elevation angles estimation accuracy under different Gamma-tone filter bank frequency channels.

Number of Frequency channel	Azimuth Angle Estimation Accuracy (+/- 15˚) IRCAM	Elevation Angle Estimation Accuracy (+/- 15˚) IRCAM	Azimuth Angle Estimation Accuracy (+/- 15˚) KEMAR	Elevation Angle Estimation Accuracy (+/- 10˚) KEMAR
40	0.958	0.952	0.942	0.936
80	0.996	0.992	0.991	0.988

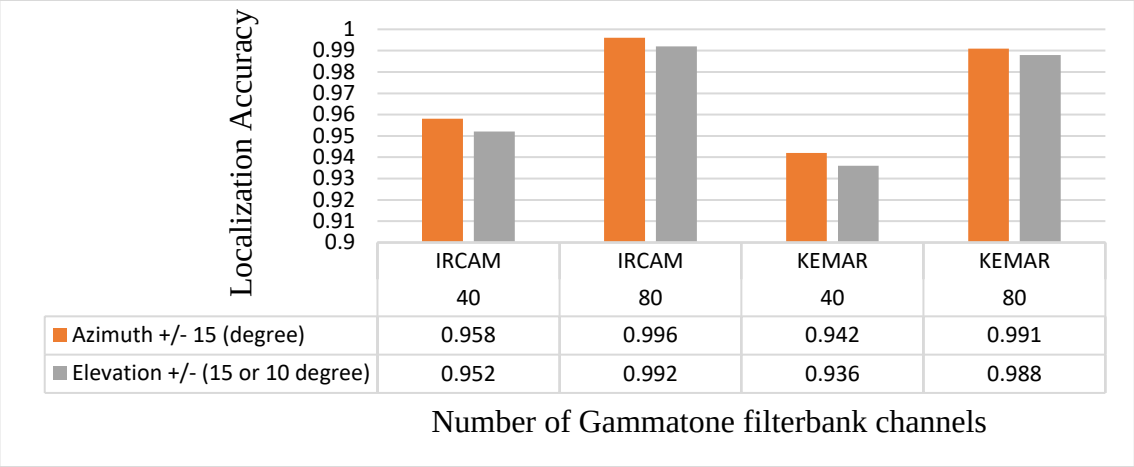


Figure 4.13: The impact of number of Gamma-tone filter bank frequency bands on localization model performance.

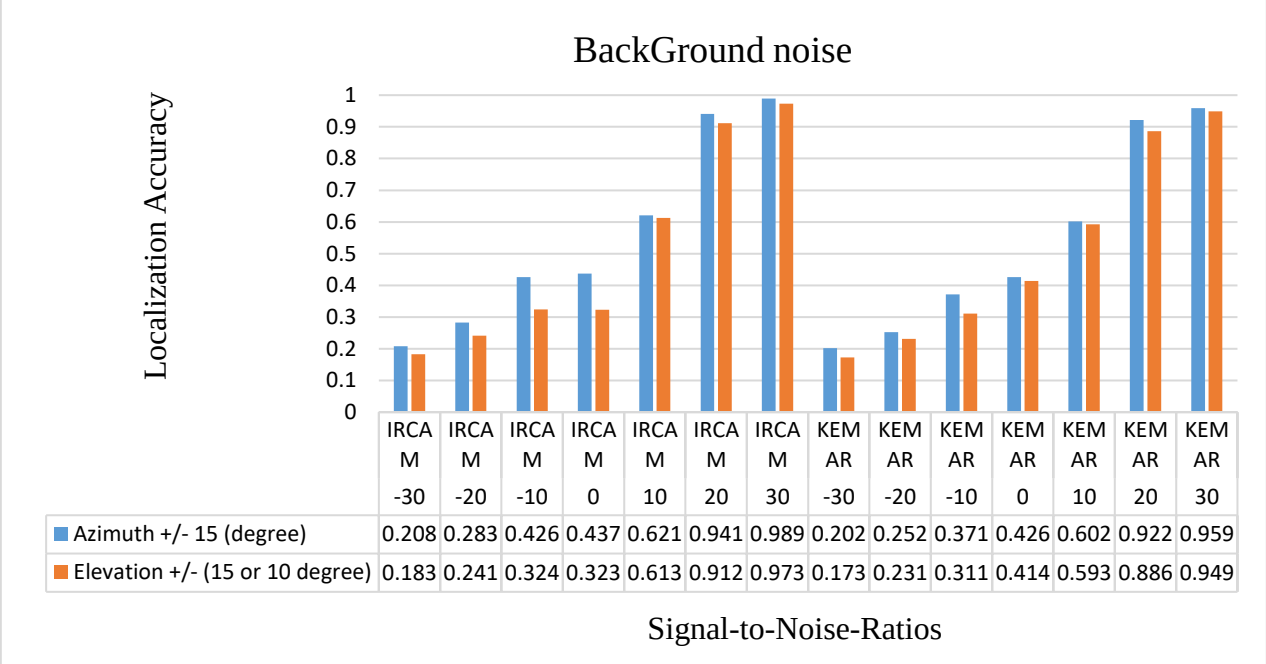
4.3 The impact of environmental noises on the performance of SSL

In the last section, a detailed description of the single sound source localization model based SNN was given as well as clarification of the model testing experiments in different conditions. In this section, the single source localization model performance is investigated in noisy environment when the 500ms of white noise as a background noise was added to the 500ms of speech signal to mimic the noisy signal. These noisy speech samples are generated by adding various levels of white noise to the incoming binaural signals. The impact of the background noise with various signal-to-noise-ratios (SNRs) on the sound source localization performance is investigated. SNR is the power ratio between the signal and noise. SNR is normally measured in decibels (dB), for example, SNR= 0dB when the ratio of the speech signal is equal to the ratio of additive noise. The logarithmic decibel scale is used to measure SNR for any noisy signal as showing in following equation:

$$\boldsymbol{SNR_{db} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)} \qquad (4.9)$$

Where P_{signal} represents the average power of speech signal and P_{noise} refers to the average power of additive noise. The aim of this test is to demonstrate the effectiveness of noisy speech samples contaminated with different SNRs on the localization model accuracy. The experiments were conducted on the speech samples from 100 speakers (50 Male and 50 Female) from the SALU-AC speech database. This experiment includes testing different speech samples

contaminated with different signal to noise ratios. Figure 4.14 illustrates the degradation accuracy based on computing the absolute angle error for azimuth and elevation. The x-axis refers to the SNRs value (in dB) between clean signals (which is usually greater than 30 dB) and 0 dB (where the level of speech and noise are equal) and each bar in the figure refers to a different level of noisy speech. While the y-axis represents the localization accuracy for Azimuth and elevation angles that computed from equations 4.8 and 4.9. The experimental outcomes show background noise (white noise) of different SNRs degrade the performance of estimation azimuth and elevation angles with both HRTF databases (KEMAR AND ICRAM).



	IRCAM -30	IRCAM -20	IRCAM -10	IRCAM 0	IRCAM 10	IRCAM 20	IRCAM 30	KEMAR -30	KEMAR -20	KEMAR -10	KEMAR 0	KEMAR 10	KEMAR 20	KEMAR 30
Azimuth +/- 15 (degree)	0.208	0.283	0.426	0.437	0.621	0.941	0.989	0.202	0.252	0.371	0.426	0.602	0.922	0.959
Elevation +/- (15 or 10 degree)	0.183	0.241	0.324	0.323	0.613	0.912	0.973	0.173	0.231	0.311	0.414	0.593	0.886	0.949

Figure 4.14: Sound Source Localization Performance for different SNRs values.

It can be noticed that the localization model has a stable performance in higher SNR, for example, the azimuth estimation accuracy is (0.98) with 30dB and 0.94 with 20dB) for IRCAM and (0.95) with 30dB and 0.92 with 20dB) for KEMAR HRTFs. While the effect of background noise starts increasing gradually at (10 dB, 0dB) then becomes higher at -10dB, -20dB and -30dB) (i.e. when SNR are reduced). The localization has a better performance in the moderate level of background noise. Whereas, there is a clear difference on localization performance over various levels of SNRs. The experimental findings demonstrate that the localisation model will

do well in low levels of background noise at 40dB of SNR or above. This experiment is important to prepare the localization model to employ in the real time circumstances as like the room environment.

In the previous experiments, a sound source localization model based on spiking neural network has been presented. The model has been tested under various conditions. The model appeared to have reliable performance in localizing different signal types including real recorded speech sounds. It demonstrated ability to localize different speech samples convolved with different binaural signals that measured under different conditions. The localization performance demonstrates that using HRTFs and spiking neural networks are convenient for solving binaural localization problems.

4.4 Applying a support vector machine for binaural localization

One of the most unsolved challenges in the spiking neural networks fields is, there is no clear comparison between spiking neural networks as advanced machine learning method with other low-level machine learning algorithms. In a current section, a support vector machine (SVM) as an alternative machine learning method to the SNN is applied for single sound source localization. The main reason for this is to investigate the SVM strength in processing the binaural responses and to compare with the performance of the SNNs.

A support vector machine (SVM) is applied with a linear kernel approach as a multiclass classifier to predict all possible locations in a certain HRTF data set. Linear SVM has more plasticity in selecting penalties and loss functions, and it is better when handling lots of samples. In this test, a support vector classifier as a supervised machine learning method is applied on the filtered binaural signals by 40 frequency bands of a gamma-tone filter bank. The output of the cochlear filter is reshaped to construct the input features of linear-SVM. The input features array was shaped as:

$$\textit{Input array= sample rate*number of indices, 2 * 40} \qquad \textbf{4.10}$$

In this case, the sample rate is 44.1kHz and number of locations is 187 for IRCAM and 710 for KEMAR. In supervised learning, the support machine classifier used the labelled training data to find an optimal hyperplane as the output to the learning phase which categorizes new

examples.

Figure 4.15 explains the estimation accuracy of azimuth and elevation angles from applying the localization model based on SVM and compared the results with SNN based localization model. Table 4-4 demonstrates the azimuth and elevation estimation accuracy of these two algorithms that applied with KEMAR and IRCAM HRTF data sets. The experimental outcomes show a weak performance of the SVM in handling the binaural information to predict the single sound location compared with SNN. The key feature of SNN is ability to process the spectro-temporal characteristic of the complex data. Traditional machine learning, including SVM, struggled to deal with the complexities of Spatio-and spectro – temporal data (SSTD). SSTD is a term that relates to processing the data depending on finding the correlation between time and place (Scott 2015). Also, a lot of the strength of SVM comes from the non-linear kernel, and because the dimensionality of the data is so high it is completely unsurprising that is didn't work very well. Furthermore, it is a multiclass problem with a big number of classes need to classify and yet SVM is not useful to solve multi- class classification problem despite of its effectiveness as binary classifier.

However, to apply SVM successfully in binaural localization field, it is required more pre-processing for the binaural sound information to analyse the correlation between the two ear signals. For example, applying cross-correlation (CC) to estimate the time-delay between the two ears signals.

Table 4-4: The localization accuracy for SNN model and SVM model for single sound source localization.

Machine learning model	Azimuth Angle Estimation Accuracy (+/- 15˚) IRCAM	Elevation Angle Estimation Accuracy (+/- 15˚) IRCAM	Azimuth Angle Estimation Accuracy (+/- 15˚) KEMAR	Elevation Angle Estimation Accuracy (+/- 10˚) KEMAR
SNN	0.986	0.982	0.957	0.946
SVM	0.642	0.532	0.527	0.419

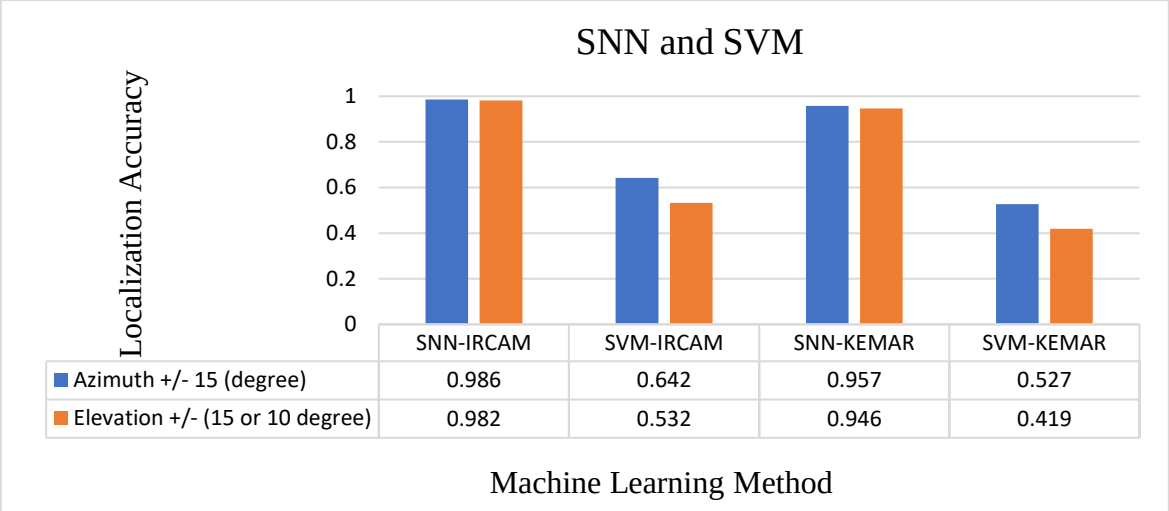


Figure 4.15: Comparison between SNN and SVM for binaural sound source localization.

4.5 Multisource sound localization based on SNN

The multisource localization is known to be one of the greatest challenges in hearing perception fields since it's significantly compromised the system reliability due to ambiguity in HRTFs channels. In the previous sections, a sound source localization model based on SNN is presented and examined with two HRTF data sets. It investigated the effect of different types of input signals and different SNRs on the robustness of a single sound source localization model. In this section, the SNN based localization model was investigated for multisource localization. To simulate the signal at the ears when two different sounds are emitted from two separated locations, a mixing process is carried out as explained in figure 4.16.

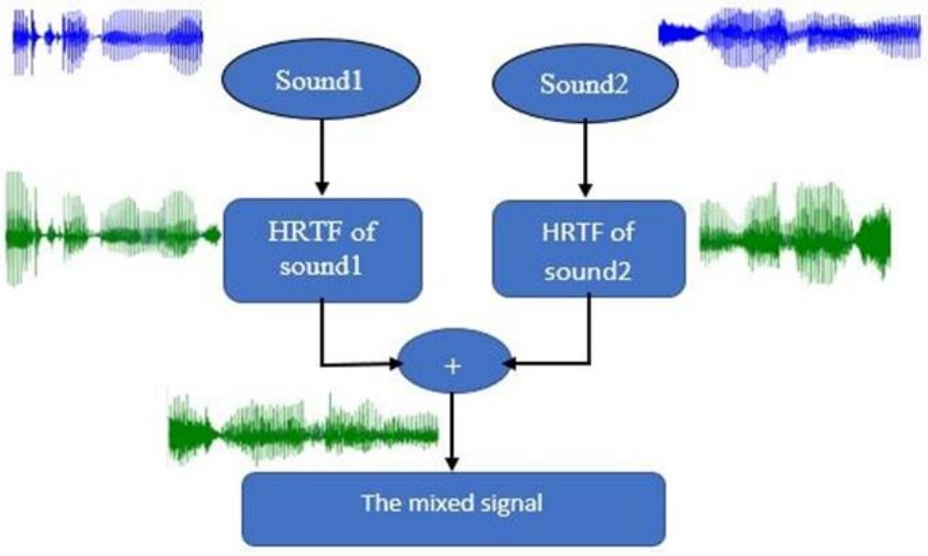


Figure 4.16: The mixing process for two different speech signals from two locations

The multi multisource localization model was implemented based the simple heuristic of choosing on the two most active coincident neurons. This is followed by the same rule for localizing a single sound source, a 'two-winners-takes-all' concept was implemented to detect the two locations. In this case, the method expects there are two winners that representing two sound locations and the experimental results, and two pairs of coincident neurons with the highest and 2nd highest firing rates are identified. Figures 4.17 to 4.31 illustrate the performance of this methodology.

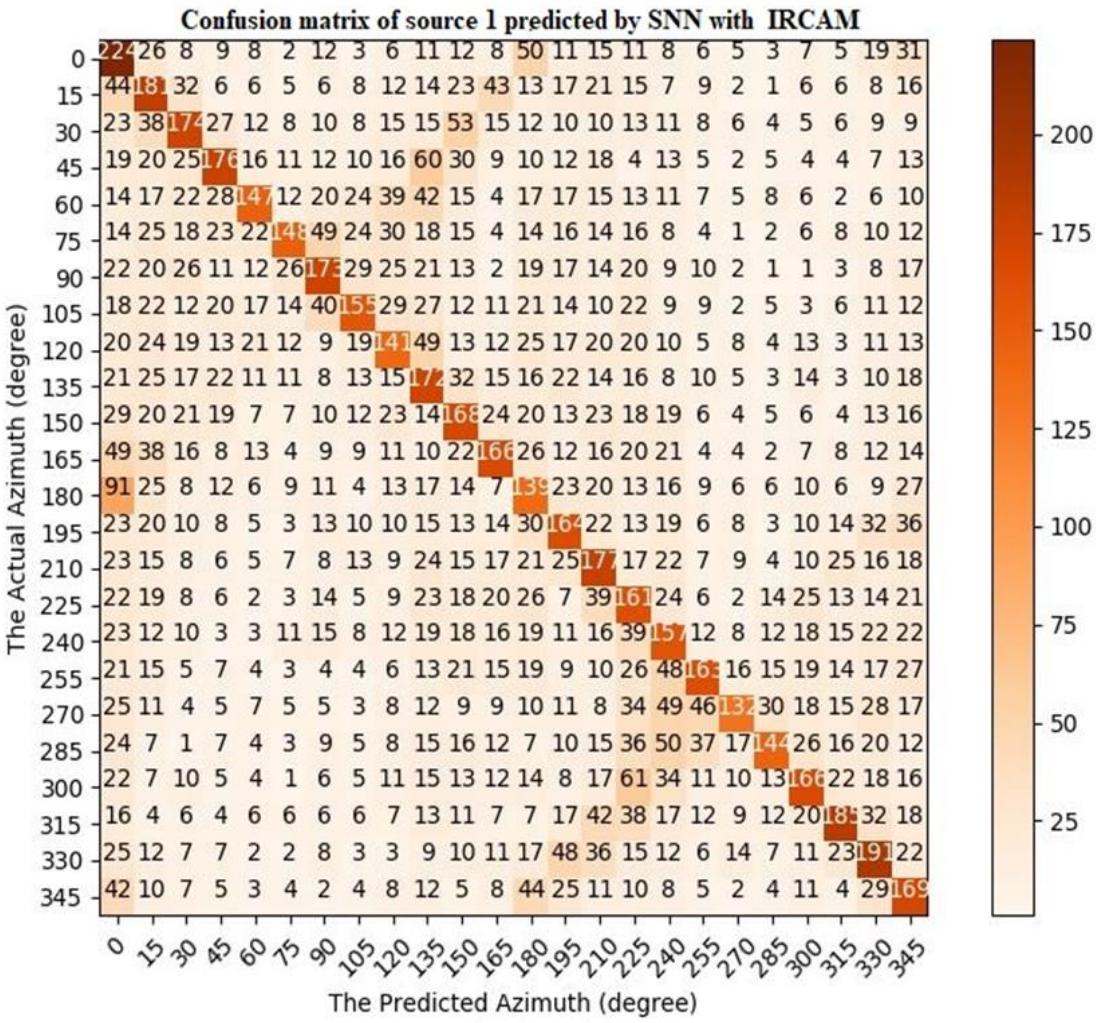


Figure 4.17: The confusion matrix plot for the source one azimuth angles predicted by multisource localization based SNN model with IRCAM and validation speakers.

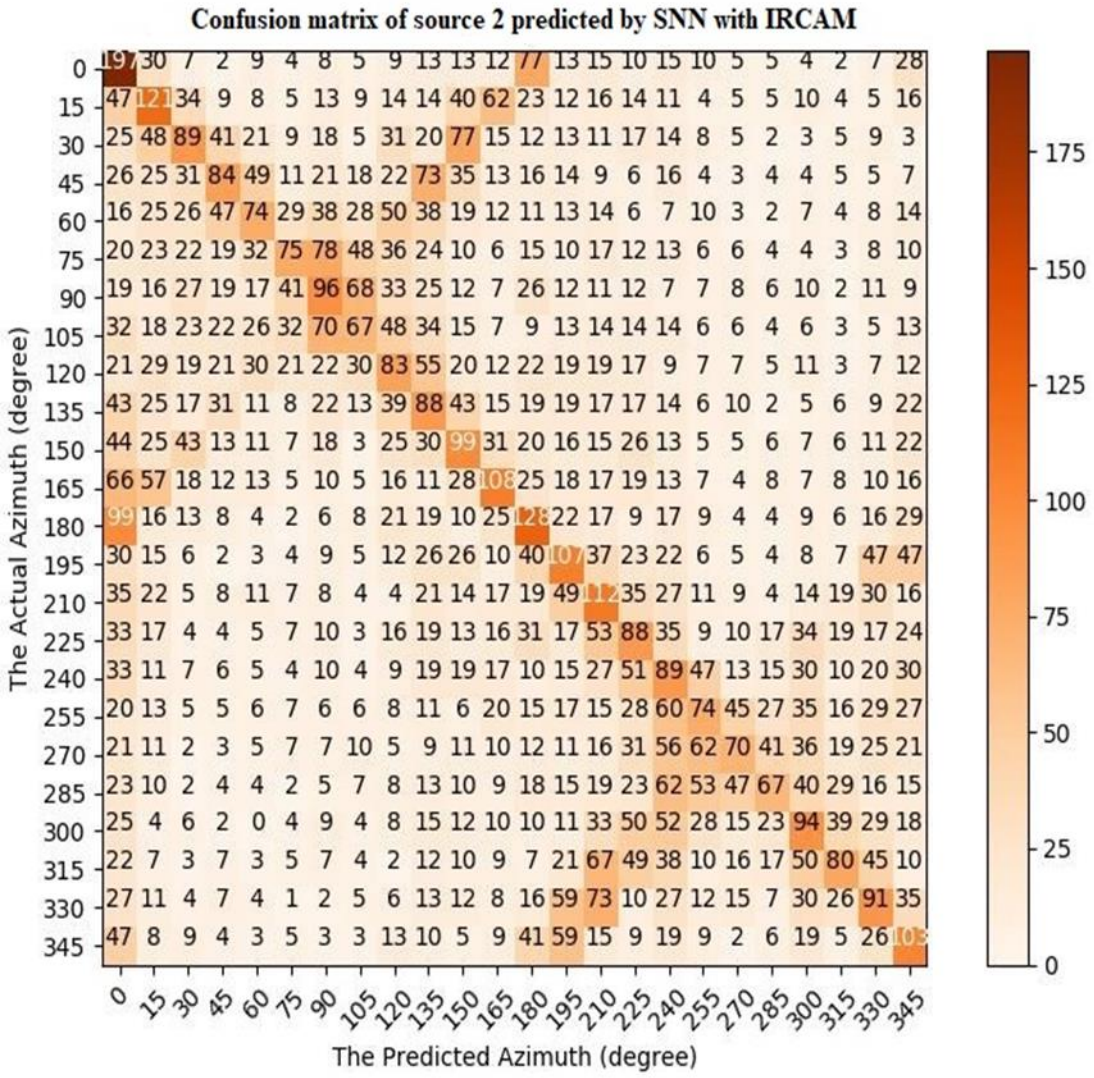


Figure 4.18: The confusion matrix plot for the source two azimuth angles predicted by multisource localization based SNN model with IRCAM and validation speakers.

Figures 4.19 and 5.20 show the absolute angle error between the original and predicted locations for location one and location two that predicted by SNN with IRCAM HRTFs.

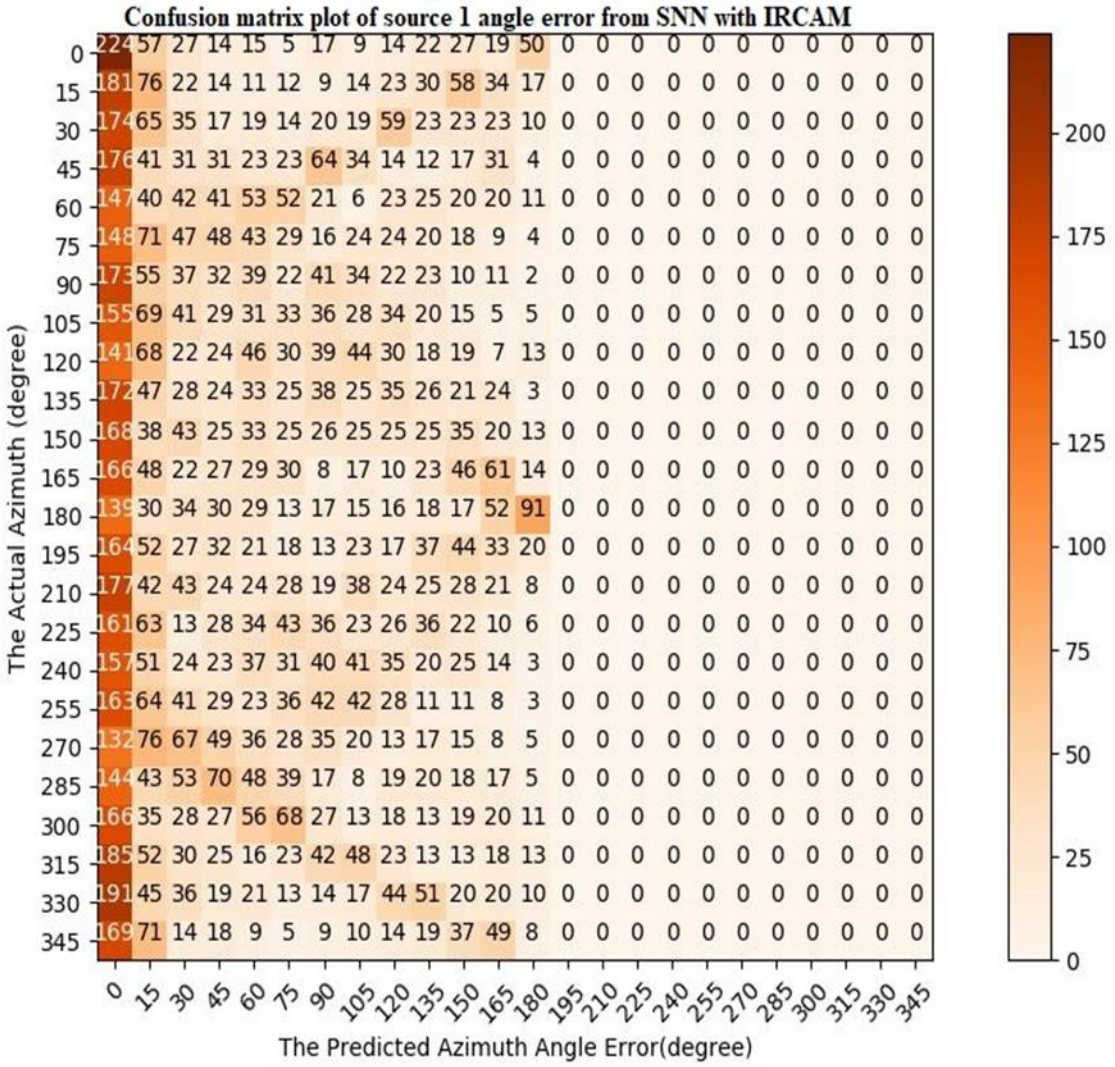


Figure 4.19: The source one azimuth angle errors from applying multisource localization based SNN on IRCAM with validation speakers.

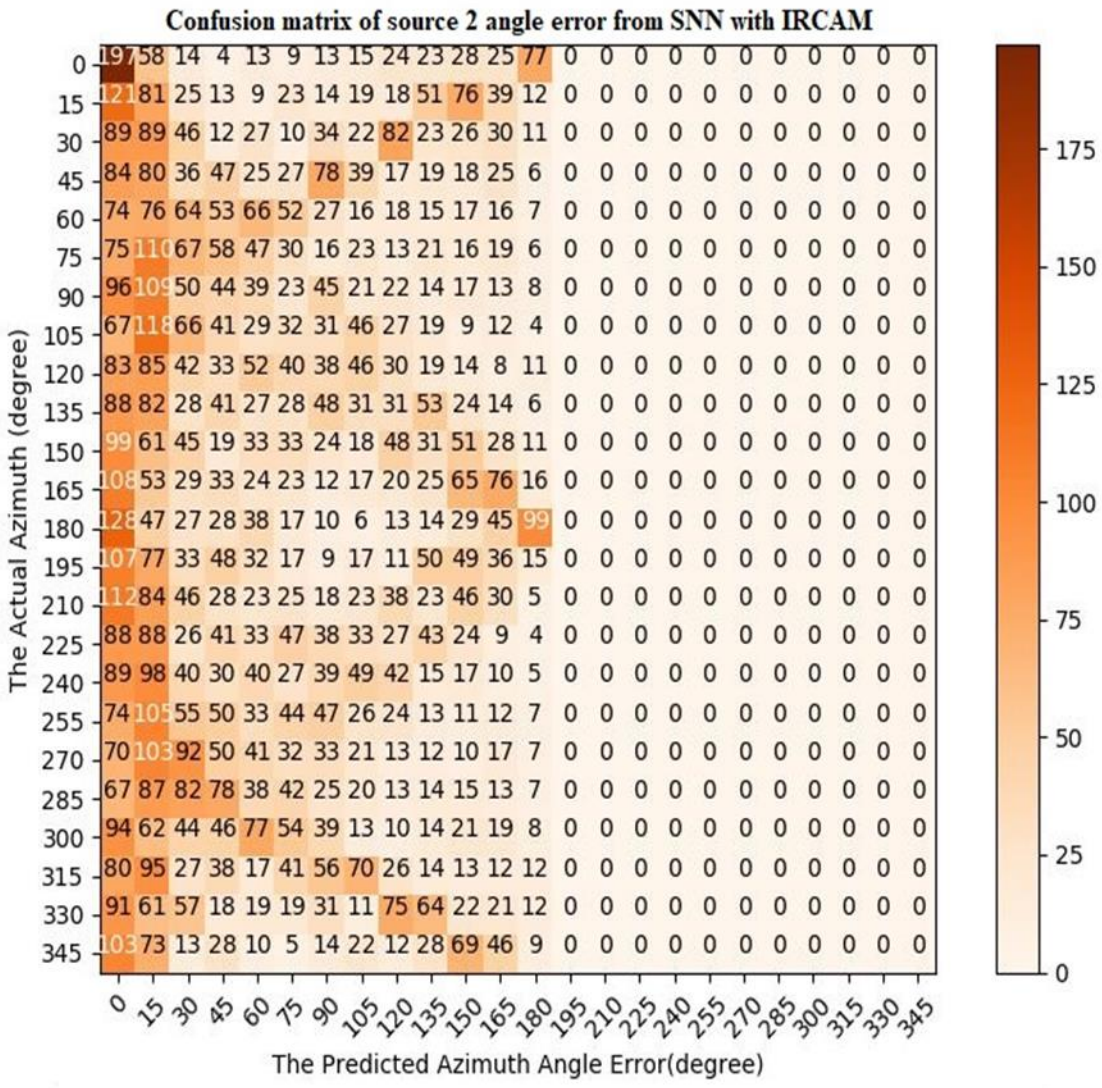


Figure 4.20: The source two azimuth angle errors from applying multisource localization based SNN on IRCAM with validation speakers.

Figures 4.21 and 4.22 illustrate the absolute angle error between the original and predicted locations for location one and location two that are predicted by SNN with KEMAR dummy head data set.

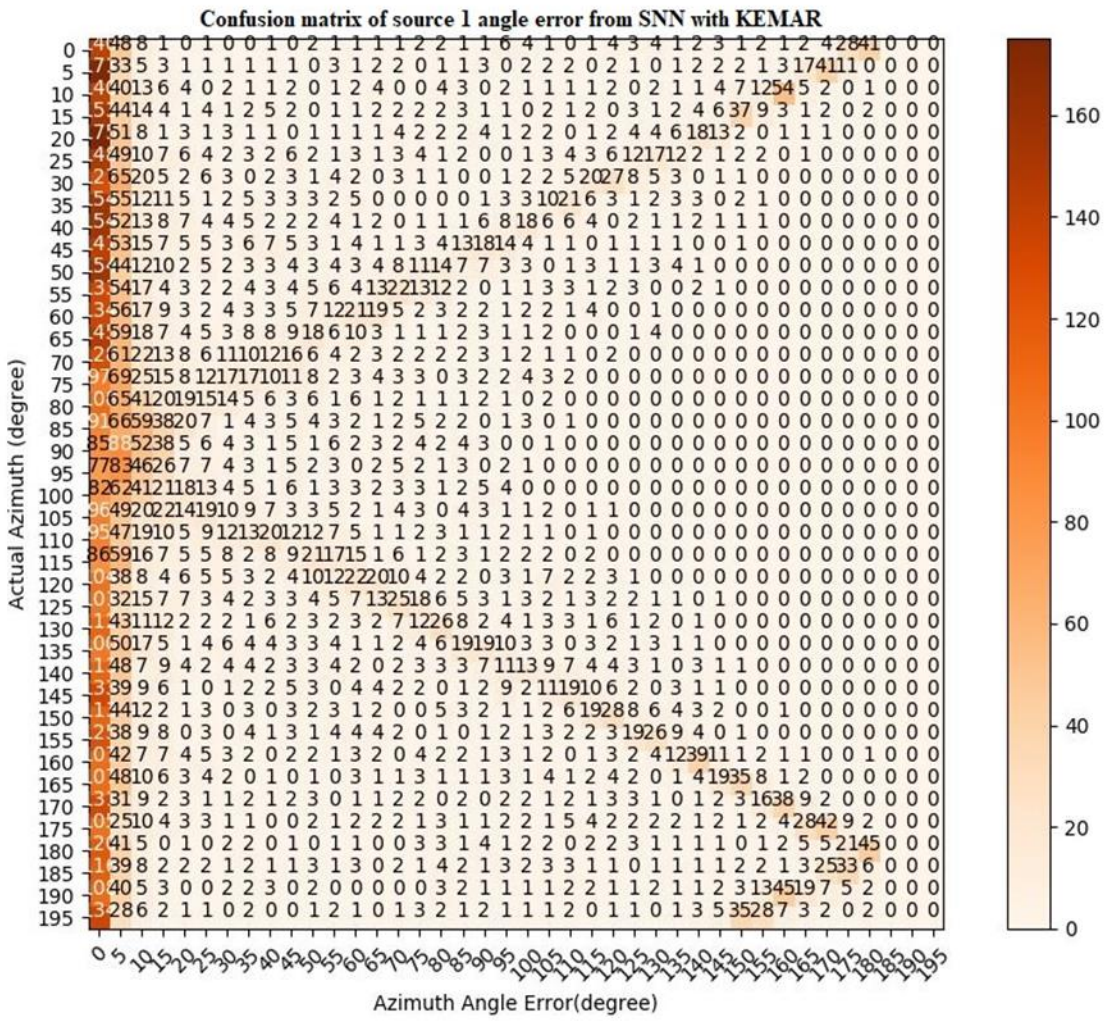


Figure 4.21: The source one azimuth angle errors from applying multisource localization based SNN on KEMAR dummy head with validation speakers.

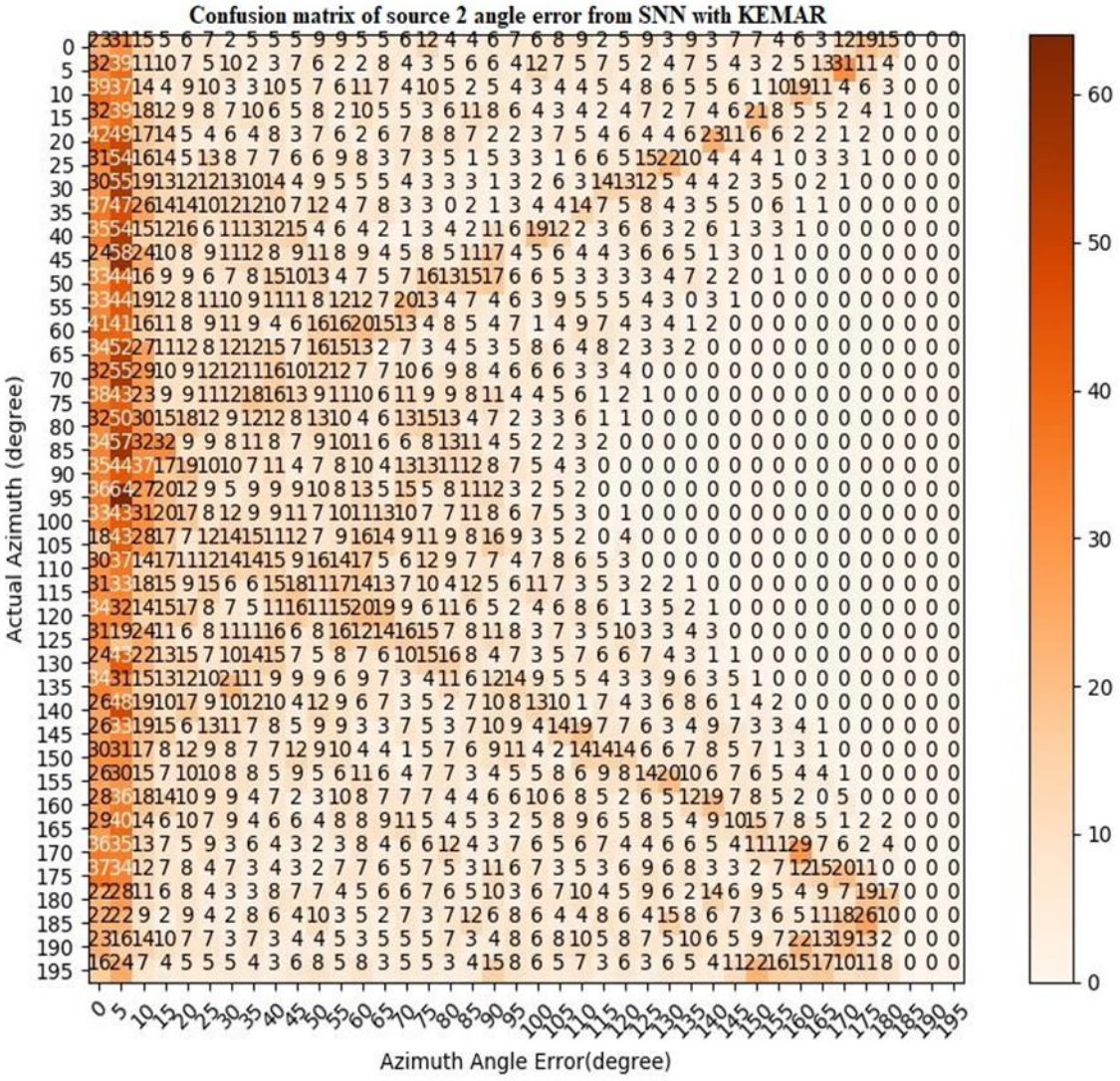


Figure 4.22: The source two azimuth angle errors from applying multisource localization based SNN on KEMAR dummy head with validation speakers.

Figure 4.23 and 4.24 show the distribution of angle errors of the two sources that results from applying the multisource localization model with IRCAM and KEMAR HRTFs. The plots are visualized as 11955 output points (angles) that result from applying the localization model based on SNN with validation data samples. The figures demonstrate that the errors have been increased in predicting source one and two from both HRTF data sets.

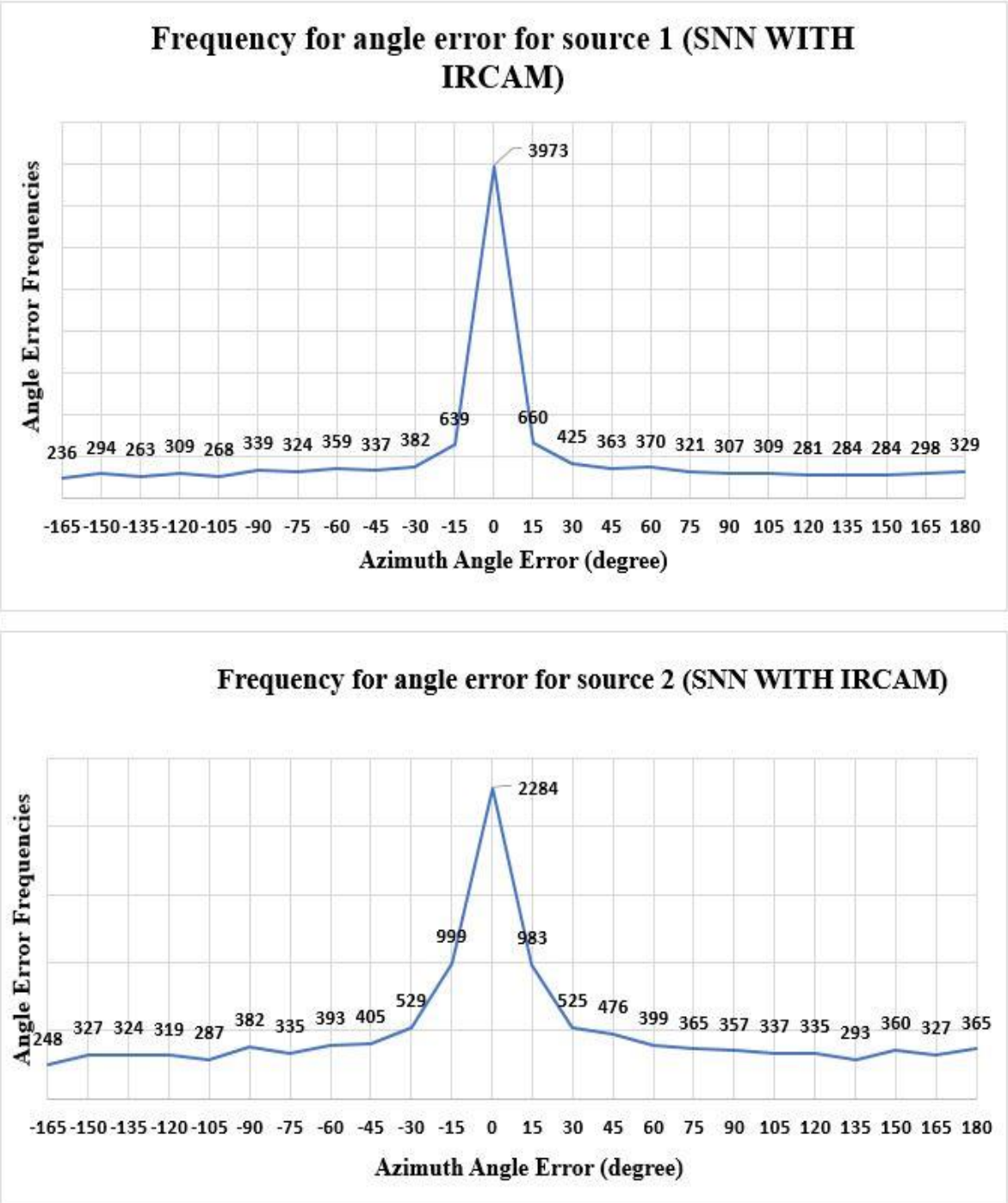


Figure 4.23: Bell shape explains the angle error frequencies for source one and source two from SNN with IRCAM HRTF and validation speakers

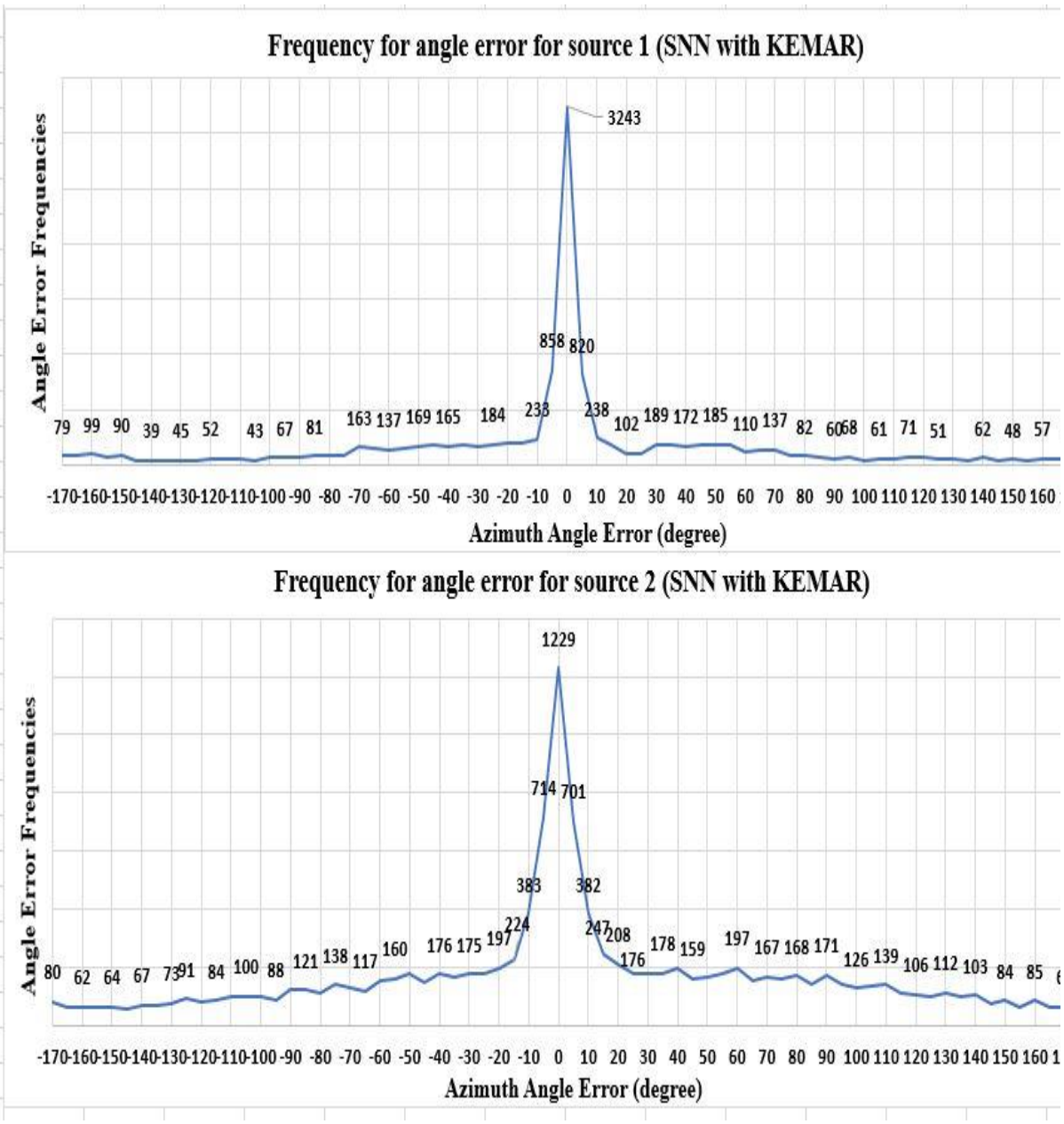


Figure 4.24: Bell shape explains the angle error frequencies for source one and source two predicted by SNN with KEMAR HRTF.

The localization outcomes demonstrated that the source two estimation angle error was high compared with source one estimation angle error. There were only 2284 angles predicted correctly out of 11955 angles applied from IRCAM dataset as shown in figure 4.23. While, figure 4.24 showed only 1229 outputs points (angles) that estimated correctly by applying the SNN based localization model with the KEMAR set.

However, it is obvious that the SNN based localization model was unable to process the spiking neural firing rate to accurately locate the two sources due to the ambiguity in the input signal results from mixing two sound signals. Although the method has an acceptance performance in detecting one source, it completely fails to locate the second source as proved in the figures 4.17 to 4.24. The SNN based localization model was extended to enhance it localization performance for multisource localization. The spiking neural based localization model output firing rates was processed using various machine learning methods including DNN and SVM. This novel idea has been tested and the results with different machine learning algorithms have been tested for single source localization as displayed in the following sections.

4.6 Sound source localization using hybrid model from SNN with machine learning methods

A novel idea was suggested to solve the multisource localization challenge when two different sound signals are emitted from two different locations at the same time. In previous sections, a SNN as a single sound source localization model has been investigated and tested with different input signals and under different conditions. The firing rates of the coincidence-neurons in the spiking neural network model provide information as the location of a sound source. Goodman used a winner-takes-all approach, where the azimuth and elevation of the neuron with the maximum firing rate is taken as the optimal prediction. This was shown to be accurate for single sound source localization, but the accuracy reduces for localization of multi sound signals that are emitted from two locations at the same time.

To improve the robustness of the prediction, the firing rates of all coincidence-detection-neurons is used to predict source locations. In the section, source localization consists of two complementary stages as explained in figure 4.25. Firstly, pre-processing which includes binaural feature extraction by using the firing rates from the SNN. Secondly, the localization problem is formulated as a classification problem where each class refers to an only source location. For evaluation process, the classification is carried out using distinct types of machine learning approaches included support vector machine (SVM), random forest, K-nearest-neighbour algorithm (KNN) and deep neural network (DNN).

Varied sizes of data sets have been generated to investigate the data size impact on the machine learning localization performance. These data were created using two types of input sound signals; white noise and speech samples. In the following sub-sections, the localization performance for different machine learning approaches with different size and types of generated data is examined. The training and validation data sets were generated using both KEMAR and IRCAM HRTF data sets to investigate the localization activity with different anatomical parameters.

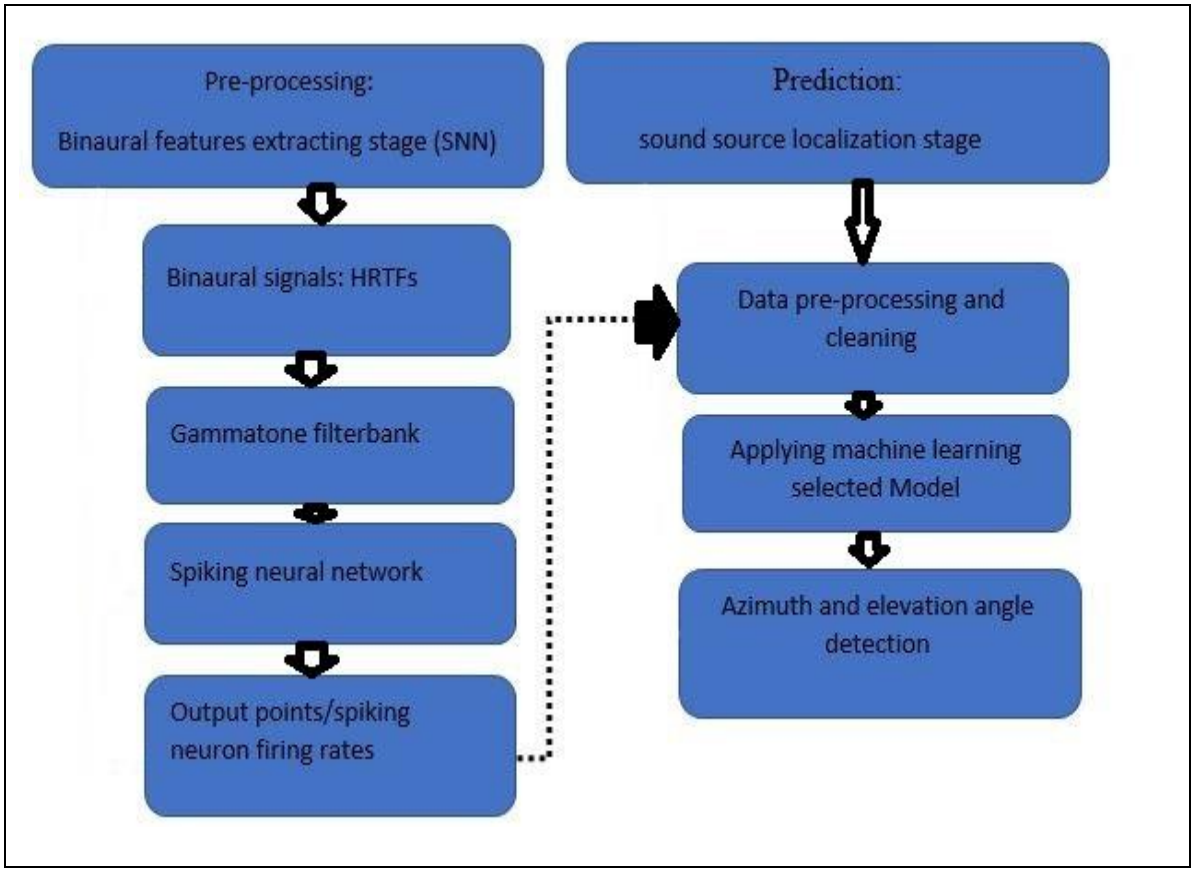


Figure 4.25: Single sound source localization by using integrated model from SNN as pre-processing method and machine learning algorithms.

4.6.1 Generate data from IRCAM and KIMAR with white noise input signal

The current stage of work can be summarized in the following steps: Firstly, a training dataset was generated. The IRCAM HRTF dataset, which has 187 azimuth and elevation angles, was convolved with 187 different instances of white noise (500ms duration). Likewise, The

KEMAR HRTF dataset, which has 710 azimuth and elevation angles, was convolved with 710 different instances of white noise (500ms duration). The response of a spiking neural network (embedded with the same IRCAM and KEMAR HRTF databases) to each of these white noise bursts is analysed and the firing rate of each coincidence-neuron calculated. Figures 4.26 and 4.27 show firing rates for each coincident neuron for a single source-location with IRCAM and KEMAR respectively. This results in 187 data points with 7480 dimensions for IRCAM and 710 data points with 28400 dimensions for KEMAR; the dimensionality is determined by the number of gamma-tone frequency bands times the number of locations in the HRTF data set. The firing rates in figures 4.26 and 4.27 are effectively the input feature set for a particular source location.

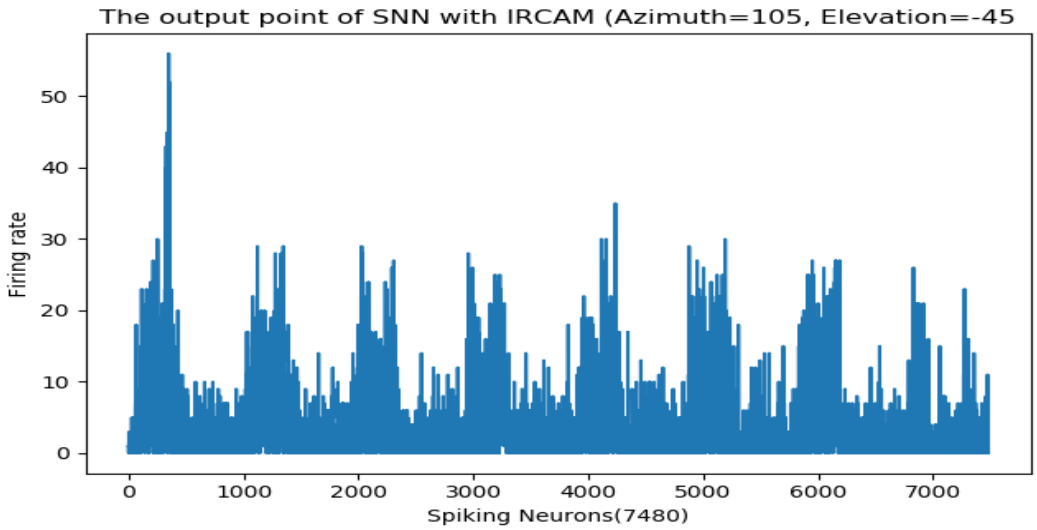


Figure 4.26: Example of the outputs points that used to generate the new data set which represent firing rate of coincidence neurons in the spiking neural network that was given input with data from the IRCAM HRTF database.

As the data represents angle and frequency, plots in figures 4.26 and 4.27 trying to show the differences in the firing rate levels for the IRCAM and KEMAR HRTF databases.

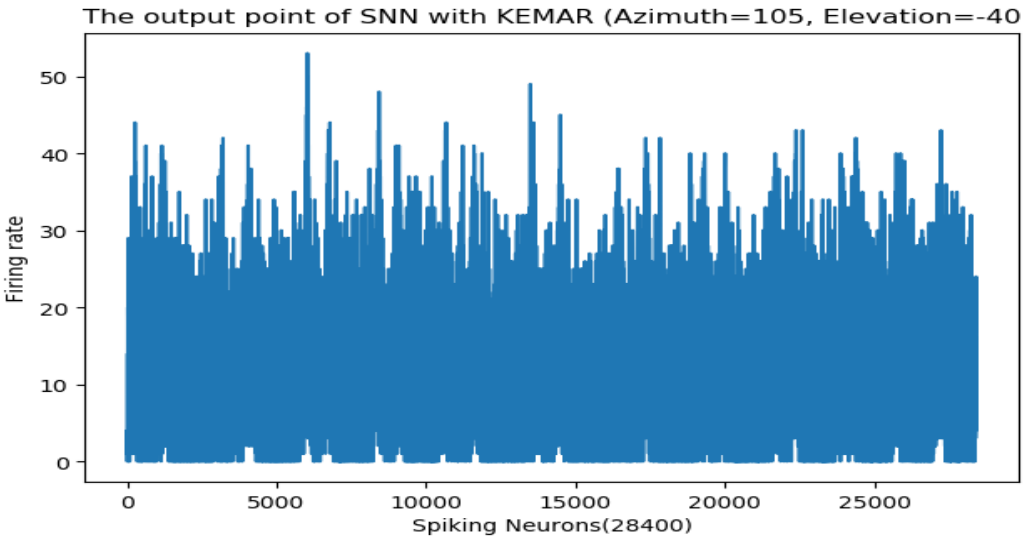


Figure 4.27: Example of the outputs points that used to generate the new data set which represent firing rate of coincidence neurons in the spiking neural network that was given input with data from the KEMAR HRTF database.

A second dataset is generated to validate the performance of the localisation algorithm; this dataset was identical but used different instances of white noise. The data are used to train and test selected machine learning techniques; support vector machine (SVM), K-nearest neighbour (KNN), and random forest (RF). These machine learning algorithms are selected to investigate their abilities in localizing different sound signal sources (azimuth and elevations angles). SVM has flexibility in terms penalties and loss functions; it is known to perform well as do many other machine learning algorithms when there are enough data for training phase (Demidova et al. 2016). SVM, with a linear kernel, a penalty parameter C=1, is implemented as a classifier technique to predict azimuth and elevation. For the supervised k-nearest neighbour algorithm, the performance is analysed with the number of neighbours (k), ranging from one to five. For the random forest classifier, the algorithm has been tested with different numbers of estimators to investigate the most suitable based on localization performance. The member of estimators has been varied in the range from 10 to 10000 (see appendix I). Also, a localization model based on deep neural networks (DNN) was tested for sound source localization. A DNN

with three hidden layers was applied for localizing single sound sources (chapter 5 has detailed description about the DNN structure and parameters).

The effect of training data size was investigated by generating a result from each location twenty times, using a new instance of noise for each data point. This results in 3740 data points from IRCAM and 14200 data points from KEMAR. The performance of each classifier is assessed by computing the signed angle error for azimuth and elevation. Classification performance is computed using 5-fold cross-validation accuracy.

4.6.2 Results and discussion

Figure 4.28 and 4.29 shows the localization accuracy of each of azimuth and elevation angle resulting IRCAM and KEMAR respectively (the elevation accuracy in the KEMAR data set is 10 degrees due to the resolution). The experimental results show that SVM performs the best with 78% accuracy followed by k-NN with 69% and the random forest with 49%.

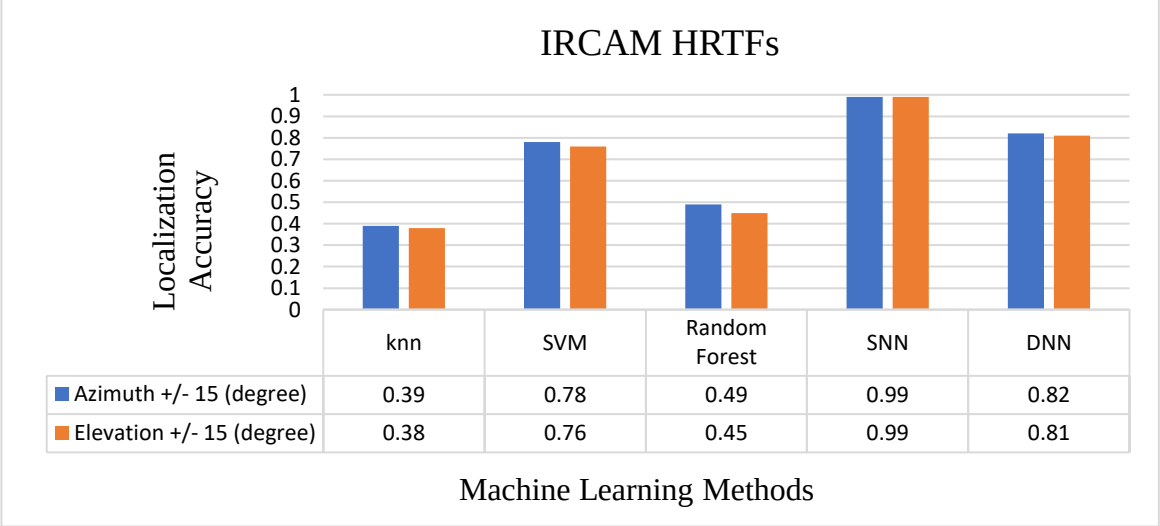


Figure 4.28: The localization accuracy for machine learning methods trained using only 187 output points that generated from trained the SNN with different instants of white noise convolved IRCAM HRTF.

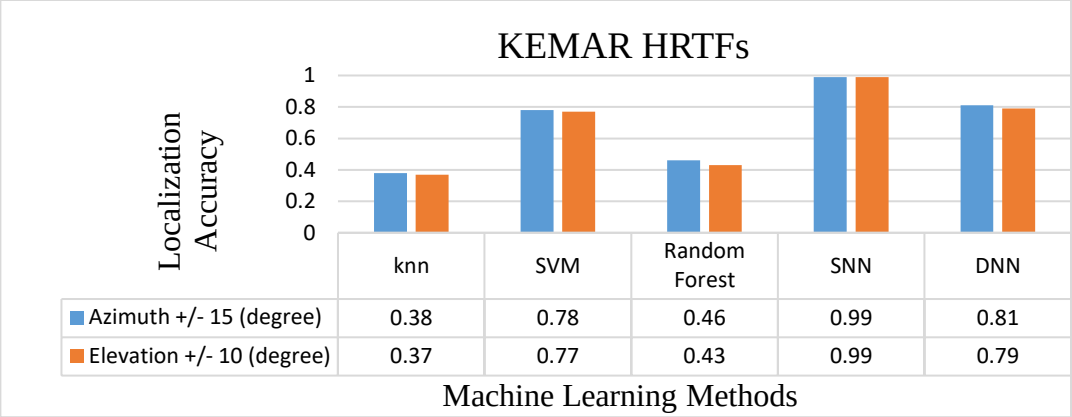


Figure 4.29: The localization accuracy for machine learning methods trained using only 710 output points that generated from trained the SNN with different instants of white noise convolved KEMAR HRTF.

By increasing the size of the training data set the performance of each classifier is improved and leads to enhancement of localization accuracy in both horizontal and vertical planes as shown in the figure 4.30 and 4.31. The localization problem has been processed by machine learning models as a multi-class classification task. And, most of machine learning methods present a less effectiveness when dealing with a big number of classes and required increasing in the computation cost to get better classification accuracy. So that, the machine learning methods presented an uneven classification performance as demonstrated in the experimental results.

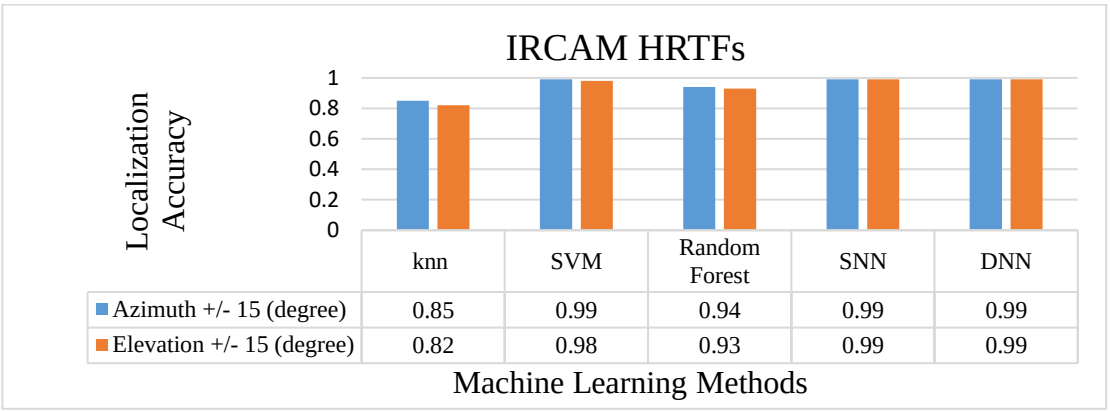


Figure 4.30: The localization accuracy for machine learning methods trained using only 710 output points that generated from trained the SNN with different instants of white noise convolved KEMAR HRTF.

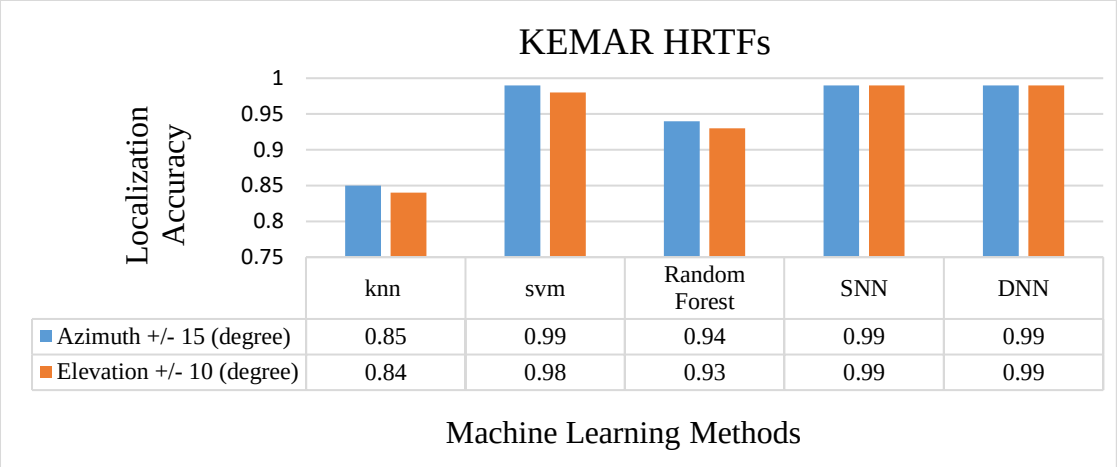


Figure 4.31: The localization accuracy for machine learning methods trained using data generated from each location twenty times represent different instants of white noise convolved KEMAR HRTF.

The outcomes discriminate between two types of performance for machine learning methods rely on the size of training data. At first, the machine learning models were trained by data generating from each location results in 187 data points for IRCAM and 710 data points for KEMAR. The effect of training data size was investigated by generating a result from each location twenty times, using a new instance of noise for each data point. This results in 3740 data points for IRCAM and 14200 data points for KEMAR.

4.6.3 Generate data from IRCAM and KIMAR with different speech samples

Another data set was generated from speech signals by using various speech samples. Anechoic speech samples from SALU-AC were convolved with binaural responses were applied to test and validate different machine learning algorithms for single source localization. The experiments were conducted on the speech samples from 100 speakers (50 Male and 50 Female) from the SALU-AC speech database (Al-Noori 2017). This experiment includes test different speech samples of different speakers (male and female) and various languages (native English, Arabic,). Each speech sample represents a full sentence with 10 second duration and belonged certain speaker. These sentences have been divided in to 20 chunks, each 0.5 second, to fit with model requirements which work with input signals with a duration of 0.5 seconds.

As previously mentioned, the output point of a SNN of each speech instance results in data with 7480 dimensions for IRCAM and 28400 dimensions for KEMAR; the dimensionality is specified by the number of gamma-tone frequency bands which is fixed in these experiments to 40 times the number of locations in the HRTF data set. The training and validation data were generated and results from each location twenty times, using a new instance of speech for each data point. This results in 3740 data points from IRCAM and 14200 data points from KEMAR for each speaker (100 speakers). At first, the data was divided to into two groups: a training speaker group (30 males and 30 females) and validation speaker group (15 males and 15 females). These generated data sets were used to train and validate the machine learning methods.

The performance of each classifier is assessed by computing the signed angle error for azimuth and elevation from IRCAM and KEMAR data set. Figure 4.32 and 4.33 explain the azimuth and elevation angle estimation accuracy by each machine learning approaches and their localization compared with SNN.

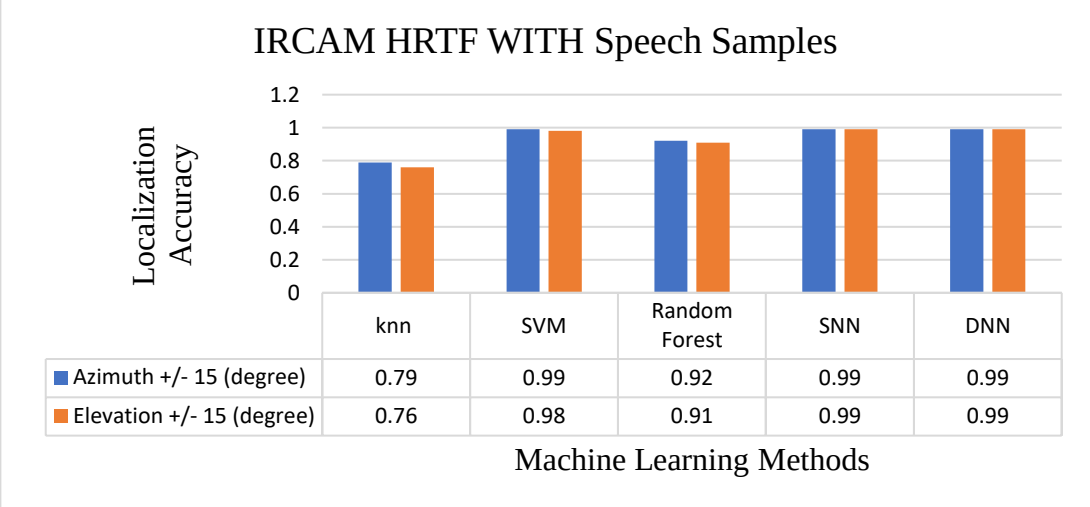


Figure 4.32: The localization accuracy for machine learning methods with big-generated-data with IRCAM HRTFs convolved with speech samples.

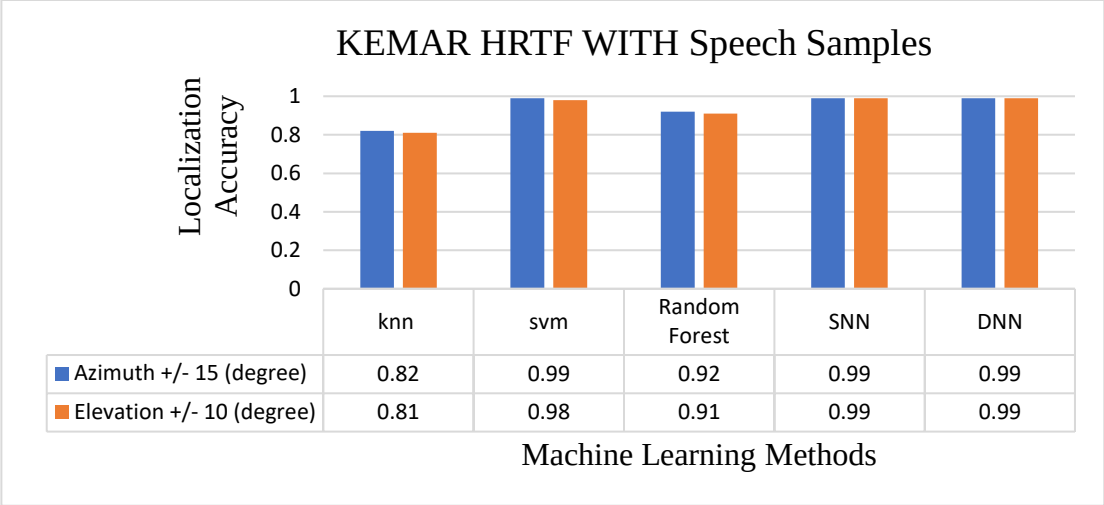


Figure 4.33: The localization accuracy for machine learning methods with big-generated-data with KEMAR HRTFs convolved with speech samples

The results demonstrate a high localization performance for SVM and DNN which is equivalent to the SNN performance. The localization problem here has been processed as a multi-class classification task. So that, increasing the number of classes that need to be classified resulted an increasing in the computation complexity that impact on the localization accuracy with KEMAR dataset. The machine learning performance for single source localization is an evidence of suitability of this novel idea in solving the multisource localization challenge.

The methods were trained with data that was generated from only one speaker (20 different speech instances for each location) and validated with different data generated from different speakers. With IRCAM, the machine learning methods appear to have an equivalent performance in estimating azimuth and elevation angles to the case using the data generating from the full range of speakers (100) or from only one speaker. Then, the number of different talkers had no impact in training. With KEMAR, the minimum number of speakers required to get an equivalent localization performance with full range of training speakers is 10 speakers. The differences in the localization performance with IRCAM and KEMAR is explained in figures 4.34, 4.35 and 4.36.

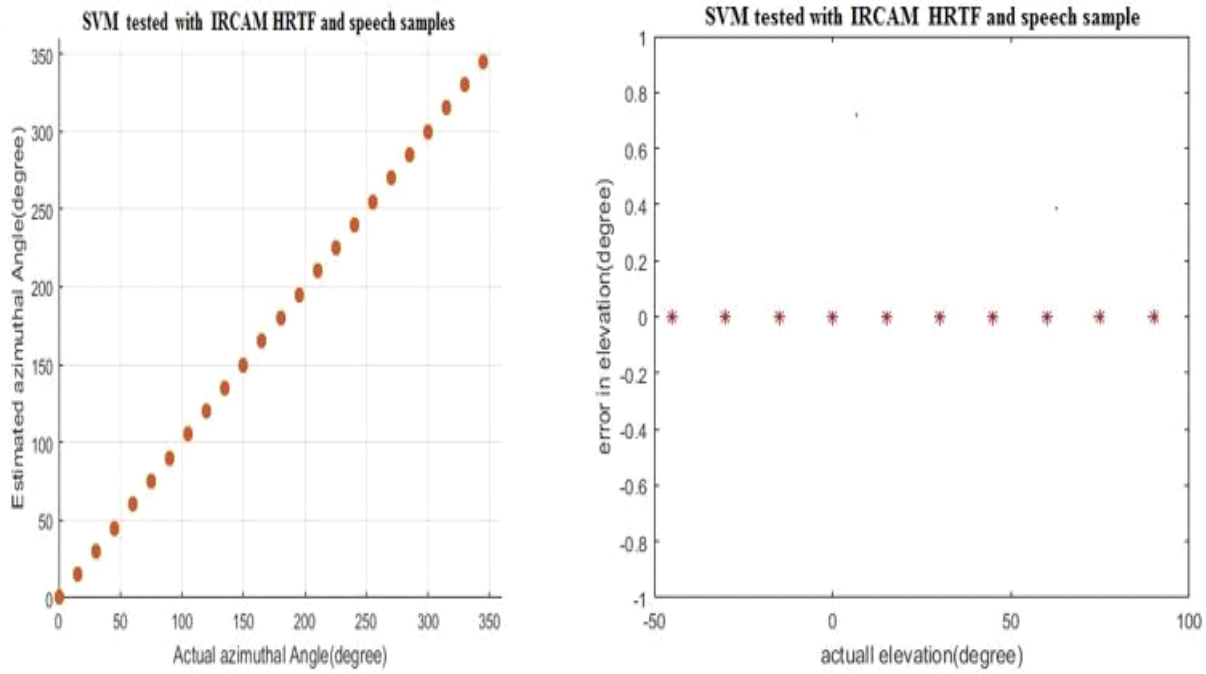


Figure 4.34: Single source localization model based on SVM performance with IRCAM HRTF data set and one speaker.

Figure 4.34 shows the SVM performance in estimation azimuth and elevation angles from IRCAM data set. The SVM was trained by using training data that was generated from 20 different speech instances from only one speaker. The x-axis in the left-side plot represents the actual azimuth angles which take range from 0˚ to 350˚ with 15˚ increment steps. The y-axis refers to the predicted azimuth angles. In the right-hand plot, the x-axis represents the actual elevation angles in the range of -45˚ to 90˚ with 15˚ increment steps. The y-axis refers to the error in predicted elevation angles.

Figure 4.35 shows the SVM performance in estimation of azimuth and elevation angles from KEMAR data set. Also, the SVM training data was generated using only one speaker. The x-axis in the left-side plot represents the actual azimuth angles which take range from 0˚ to 350˚ with 5˚ increment steps. The y-axis refers to the predicted azimuth angles. In the right-hand plot, the x-axis represents the actual elevation angles which range from -40˚ to 90˚ with 10˚ increment steps. The y-axis refers to the error in predicted elevation angles.

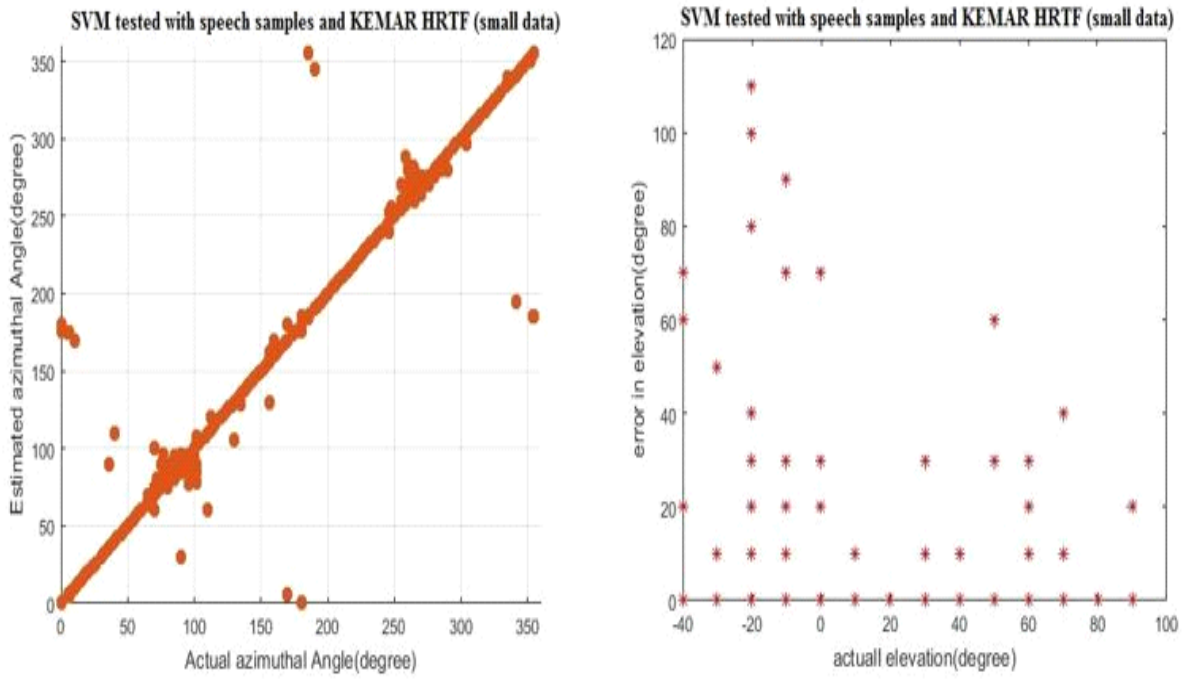


Figure 4.35: Single source localization model based on SVM performance with KEMAR HRTF data set and one speaker.

It is a notable that the angle error for estimating azimuth and elevation angles by SVM and KEMAR is high compared with IRCAM despite using the same size of training data. The localization performance for SVM with KEMAR data set is improved by increasing the size of training data set as shown in figure 4.36. The SVM was trained with data generated from 10 speakers and validated with data generated from one speaker (fresh data). There are two potential reasons for this relative difference in IRCAM and KEMAR performance. The first one is the differences in angles measurements between them where KEMAR measurements cover a wide range of locations in the vertical and the horizontal plane. These the diversity of locations required increasing in the learning examples for better localization performance of machine learning method. The second reason, the localization problem here has been processed as a multi-class classification task. So that, increasing the number of classes that need to be classified resulted an increasing in the computation complexity that impact on the localization accuracy with KEMAR dataset.

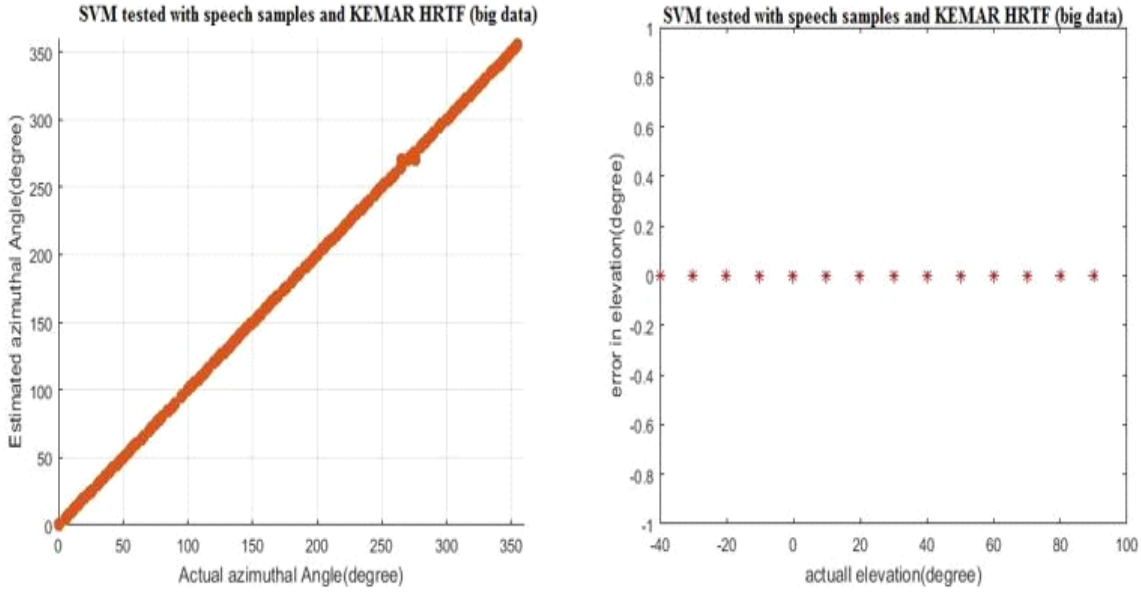


Figure 4.36: Single source localization model based on SVM performance with KEMAR HRTF data set and 10 speakers.

4.7 Chapter Summary

In this chapter, the localization model based on a spiking neural network presented by Goodman is reviewed and replicated with two HRTF data sets, KEMAR dummy head the IRCAM data set. The localization model has been tested with diverse types of input signals including Gaussian white noise, uniform white noise, pure tone modulated white noise and different speech samples that were collected in an anechoic environment. In addition, the localization model performance was investigated with single and octave frequency to demonstrate the effectiveness of localization cues on the localization model performance. Two localization related performance factors are examined, the input signal duration and number of gamma-tone frequency channels and their impact on localization model robustness are explained. The results explain the enhancement of the localization performance by increasing the input signal duration as well as the number of gamma-tone frequency bands. Furthermore, signal to noise ratio is shown to play a significant role in the robustness of the localization model. The model has been examined with different SNRs to identify the effect on performance in various levels of background noise. The outcomes show the variation of the effect of different SNRs on the performance of single sound source localization.

The support vector classifier as a multi-class classification function has been tested in processing the binaural signal that filtered through 40 gamma-tone channels to predict the incoming sound signal locations. Its performance has been compared with SNN based localization model. The spiking neural localization model has been tested to localize two sound signals that emitted from two different locations. The experimental results demonstrated that the SNN based localization model was unable to process the spiking neural firing rate to accurately locate the two sources due to the ambiguity in the input signal results from mixing two sound signals. A new idea has been suggested to improve the SNN based localization model to solve more complicated binaural hearing problems like multisource localization task. This idea is based on using SNN as a pre-processing method which includes binaural feature extraction, in the form of firing rates from the SNN. Finally, an implementation of various machine learning algorithms has been explained. Varied sizes of labelled data have been generated to train and validate the machine learning models. The localisation problem is formulated as a classification problem where each class represents a single source location. Classification is carried out using diverse types of machine learning methods. The results show differences in the performance of the various machine learning approaches in localizing single sound source. Also, they demonstrate some differences between IRCAM and KEMAR impact on the localization performance. These differences were handled by increasing the size of training data.

CHAPTER 5
MULTISOURCE LOCALIZATION MODEL BASED ON DNN UNDER CLEANAND NOISY CONDITIONS

Chapter Overview

This chapter explains the multisource localization model based on using DNN to process the SNN firing rates. The process includes the validation of the model using data generated from speech samples not used in training. Section 5.1 is a description of multisource localization structures and components. The process of generating a mixed signal and mixed data with background noise are explained in section 5.2. Section 5.3 shows the process of detecting number of sources. The *decreasing of the data dimensionality mechanism is explained in section 5.4. The detailed description of DNN that is applied for multiclass classification to solve the multisource localization problem is demonstrated in section 5.5. Section. The experiments to investigate the localization model performance with another machine learning method (SVM) are presented in section 5.6. Comparison between multisource localization model performance with localization model based SNN was showed in section 5.8. Section 5.9 examined the effect of multi-condition training using clean and noisy speech. The validation testing was carried out in emulated noisy conditions with controlled signal to noise ratios.*

5.1 Multisource Localisation Model

In the previous chapter source localization was shown to be accurate for single sound source, but the accuracy diminished for sound signals that emitted from two locations simultaneously. This chapter attempts to address this limitation. HRTFs with a wide range of azimuth and elevations were used to generate labelled data for training and validation. The SNN method is used as a feature extraction pre-processor which are then used as inputs to a learning algorithm trained to perform multisource localisation as a classification task. The advantage of this methodology is that by matching the HRTF integrated within the algorithm to the capture device (e.g. a dummy head), the impressive localisation ability demonstrated by the human auditory system may be captured. Accurate source location information as provided by this algorithm will enable many applications such as, virtual reality systems, augmented reality systems, human machine interaction and robotic applications along with security and monitoring. This technology could be used to enhance the performance of Hearing Aids. Within the hearing aid fitting process, the HRTF could be captured and embedded in the source localisation algorithm (Harder et al. 2015). Source location information could then be provided to the user via haptic or visual displays which would enhance the quality of life for people with hearing loss in one or both ears who may have difficulty in localising sounds. Multisource sound localization could also be used to enhance the tactical communication and protective systems in the military applications for example soldiers hearing protection devices (Joubaud et al. 2017).

The firing rates of the coincidence-neurons in the spiking neural network model provide information as to the location of a sound source. Goodman used a winner-takes-all approach, where the azimuth and elevation of the neuron with the maximum firing rate is taken as the optimal prediction. This was shown to be accurate for single sound source localization, but the accuracy reduces for localization of multiple sound signals that emitted from two locations at the same time. To improve the robustness of the prediction, the firing rates of all coincidence-detection-neurons which is known as the Spectro-temporal receptive fields are used to predict source locations.

In this work, the number of simultaneous sources is restricted to two. This limitation comes from the computational requirements when training the system to locate more than two sources. Training data is required with all possible combinations between pairs of locations in the HRTF

data set. Also, the data is created by finding all possible combinations between selected speakers (see section 5). However, increasing the number of locations, for example from 2 to 3, requires finding all possible combinations of three locations in the HRTF data set. This process will compound the size of training data, which increases the computational complexity and memory requirements. The multisource localisation model consists of two complementary stages: firstly, pre-processing which includes binaural feature extraction encoded as the firing rates from the SNN. Secondly, the localisation problem is formulated as a classification problem where each class represents a pair of source locations. Classification is carried out using a deep neural network. These two stages can be described in figure 5.1.

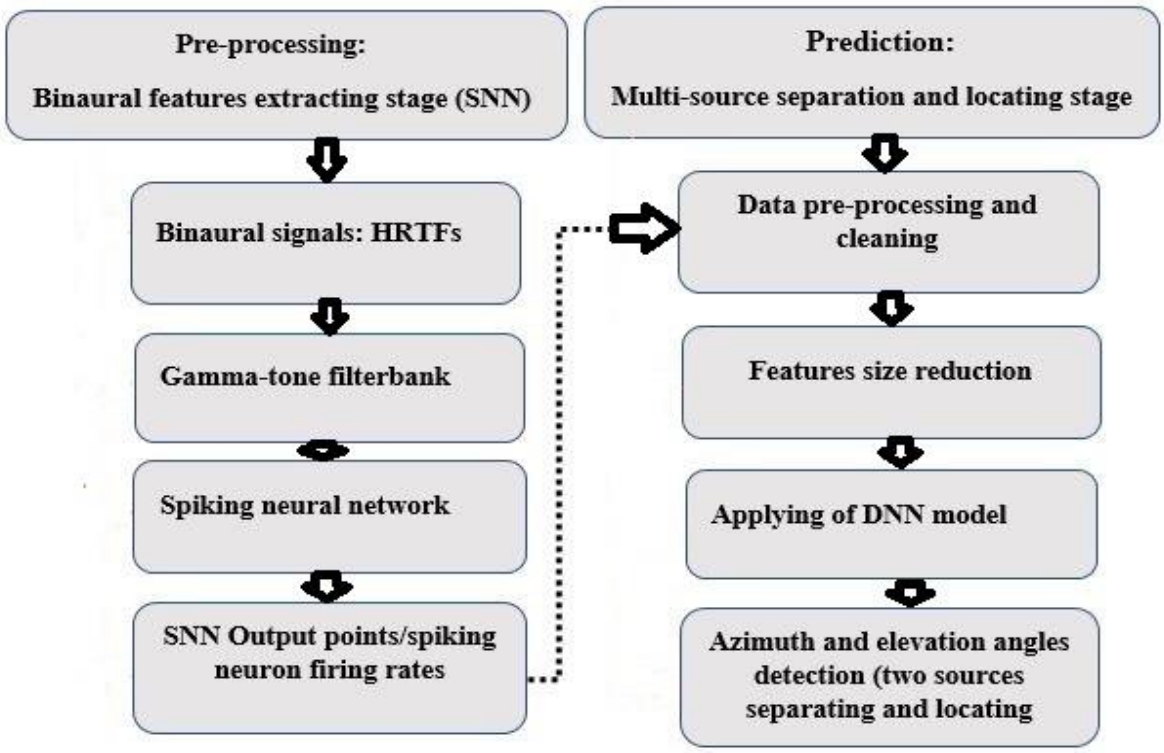


Figure 5.1: Stages of the multisource localization model, pre-processing step and prediction steps that include multi-classes multi-label classification using a DNN.

5.2 Mixing process and Data generated

A database of various speech samples, both male and female, from diversity of languages were used to generate the dataset (Al-Noori 2017). 17 speakers where chosen at random. Eight males and nine females were chosen to generate the training data. The validation data is created from

completely different speakers (two male and one female). Anechoic speech samples were convolved with binaural responses and used to train and validate the multisource localization model based on DNN. Each speech sample represents a full sentence with duration of 10 seconds. After removed the silence, these sentences were split into 20 chunks of 0.5 seconds each to achieve the model requirements.

All possible pairs of 17 speakers were simulated at all possible pairs of angles as defined by the IRCAM angular resolution. This produced one hundred and thirty-six combinations of seventeen speakers at 4032 angle pairs with each pair representing one class. Likewise, the generated data from KEMAR constituted 4800 angle pairs. Figure 3 shows the mixing process; each signal is convolved with the HTRF pair for the chosen angle and then are added together. Furthermore, to simulate the signal at the ears when two different sounds are emitted from two separated locations in noisy environment, a process is carried out as explained in figure 5.2. The training and validation noisy data were generated by using various speech samples of 500ms with different locations in elevation range (-15˚, 0˚, 15˚) by adding white noise of 500ms to the mixed two locations signal embedded in two speech signals that of 500ms. The noisy data that was generated by adding different levels of white noise of SNRs 10dB, 0db, and -10dB to the ear signals. This simulates diffuse noise due to both ears have different noise. The noisy training data was generated from different speech samples of 17 speakers (training speakers) with all possible combinations between locations from IRCAM data. The validation data was generated from different speech samples of 3 speakers.

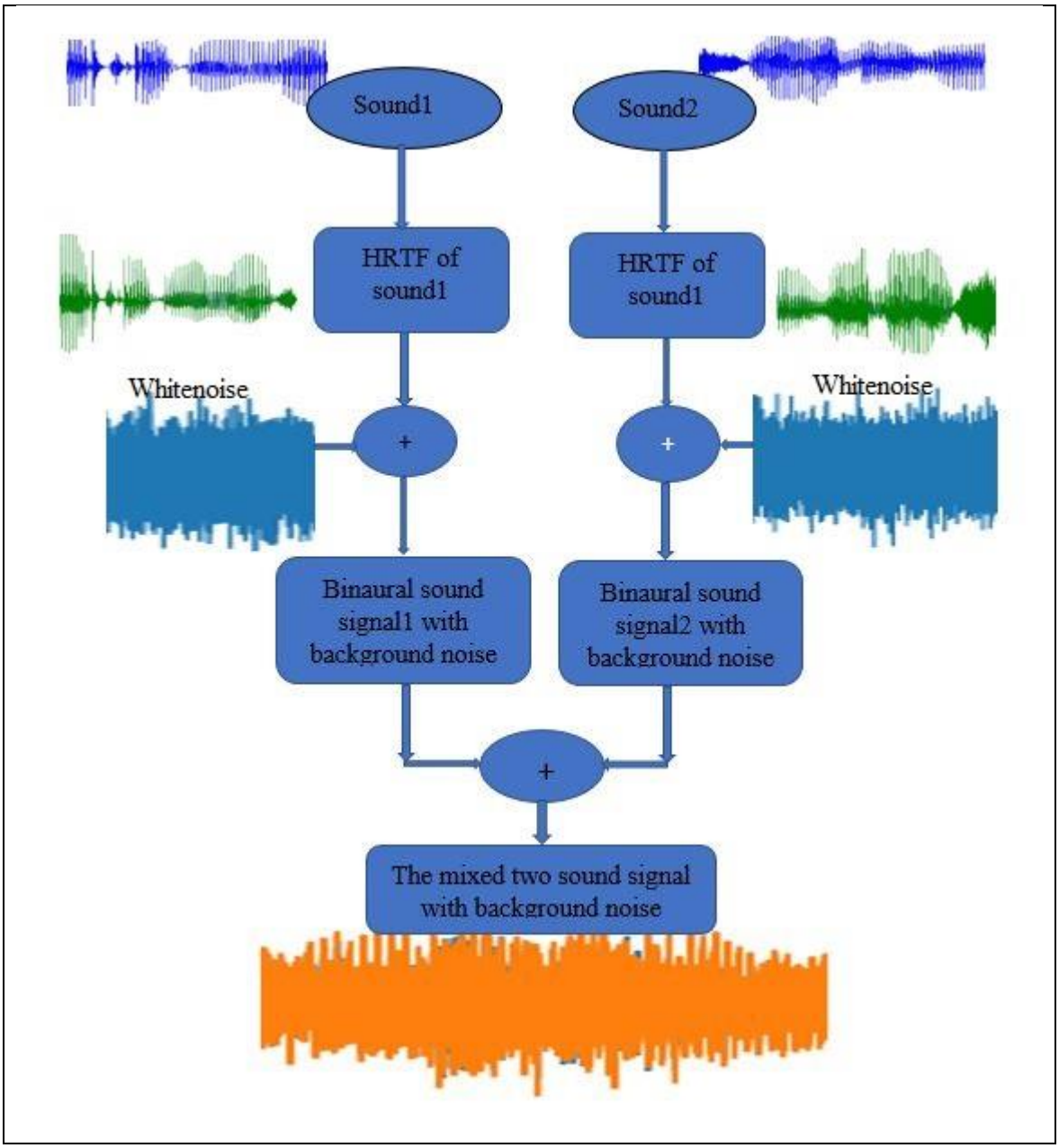


Figure 5.2: The mixing process for two different speech signals from two locations with added white noise after the convolution process to mimic the noisy environment.

5.3 Detecting the number of sources

Prior to localisation, the number of sources must be estimated. Once the number of sources is known, the appropriate localization model (single or multisource) can be selected in order detect the direction of the source or sources. Figure 5.3 shows the firing rates from all coincident neurons for one source and two sources resulted from applying SNN with 500ms speech signal.

This results in 7480 dimensions; the dimensionality is determined by the number of gammatone frequency bands (40) times the number of locations in the HRTF data set (187 in the IRCAM data set). Figure 5.4 shows the firing rates for the KEMAR set which has 710 locations, resulting 28400 dimensions.

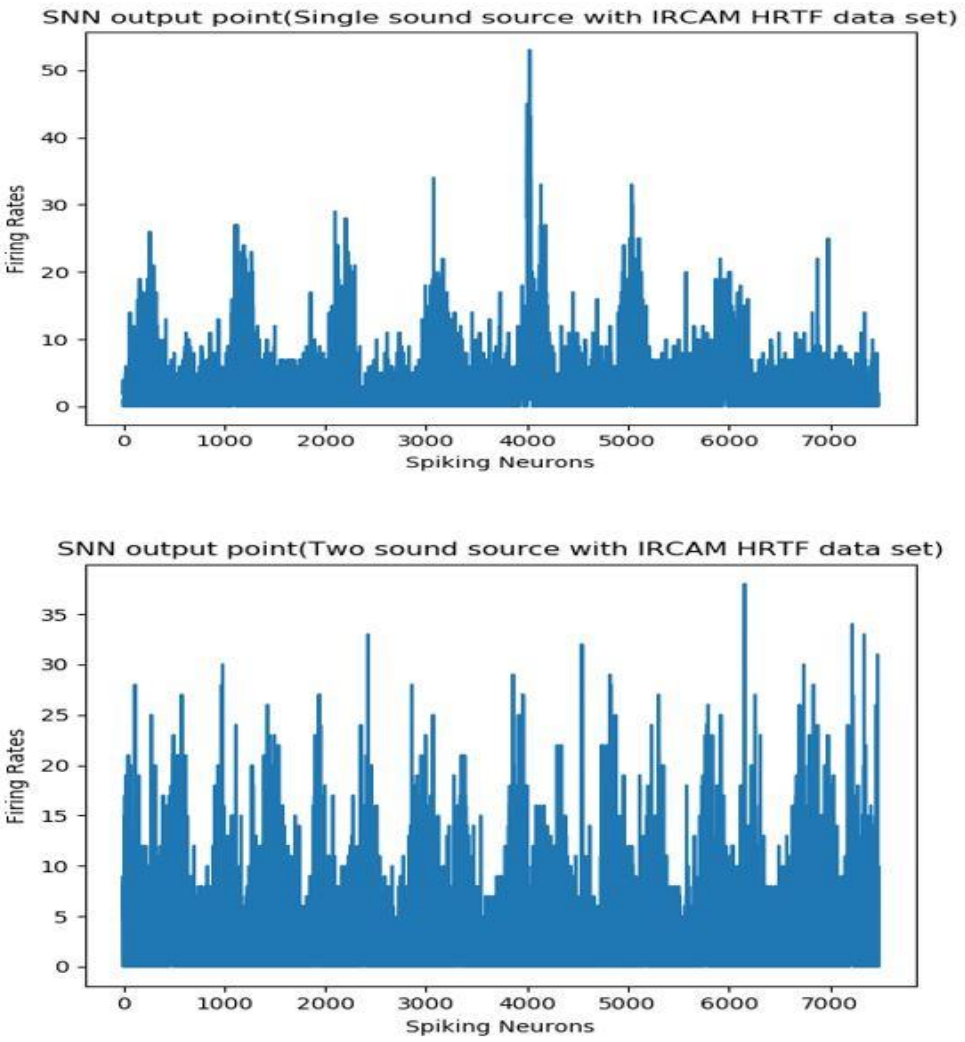


Figure 5.3: Spiking neural networks output points with IRCAM HRTF data set. Example of two types of spiking neural network (SNN) output vector that contains the firing rate for each individual neuron in the coincidence detection layers.

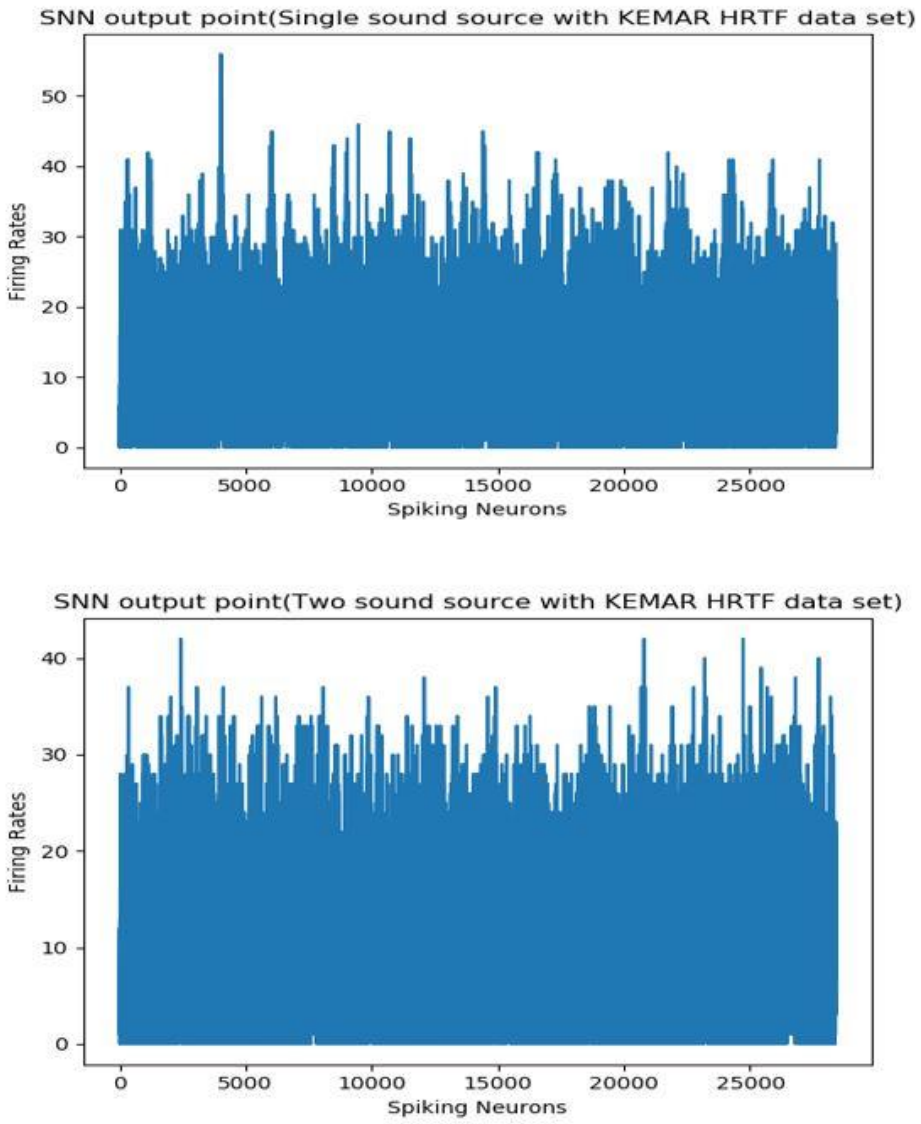


Figure 5.4: Spiking neural networks output points with KEMAR HRTF data set. Example of two types of spiking neural network (SNN) output vector that contains the firing rate for each individual neuron in the coincidence detection layers.

The firing rate is significantly higher for two source signals. Logistic regression method was applied to analyse the firing rates over the assemblage predict number of sources in the signal. It was used to estimate the number of sources in the signal based on observed firing rate characteristic. This method works to create a best fit logistic curve to separate between the two sources and one source signals.

The logistic regression was used to evaluate different types of signal and with two HRTF data sets. Moreover, the logistic regression prediction efficacy has been investigated in noisy environments and with noisy data at different level of background noise. Table 5-1 demonstrates the accuracy of predicting the number of talkers with two HRTF data sets (IRCAM and KEMAR). In addition, the results show the impact of background noise on the accuracy of predicting the number of sources. The noisy samples are generated from adding white noise of 500 ms added to speech samples convolved with binaural responses.

Table 5-1: Estimates of number of sources in diverse types of signal.

Type of inputs	Accuracy of predicted number of sources
Single source IRCAM (speech sample (female))	**99%**
Single source IRCAM (speech signal (male))	**99%**
Single source KEMAR (speech signal (female))	**99%**
Single source KEMAR (speech sample (male))	**99%**
Single source IRCAM with background noise	**93%**
Single source KEMAR with background noise	**93%**
Two sources data generated with IRCAM	**99%**
Two sources data generated with KEMAR	**99%**
Two sources with IRCAM with background noise (multi-condition noise)	**92%**

The table of results demonstrate that the number of sources in the input sound signal was predicted correctly with 99% for all different type of speech signals (female and male sound signal) for both HRTF databases. Whereas, the prediction accuracy for the number of sound signal sources has been reduced to 93% when the input signals were with background noise.

To analyse the input data pattern and vitalize the correlation between the single source and two sources firing rates, a principle components analysis PCA was applied and presented in figure 5.5. The model analyses the variance over the firing rate assemblages. The PCA is restricted to two components for visualisation. In this figure, the red points refer to the single

source signals and the green points refer to the two sources signals. It can be seen that the two are separable by a non-linear classifier.

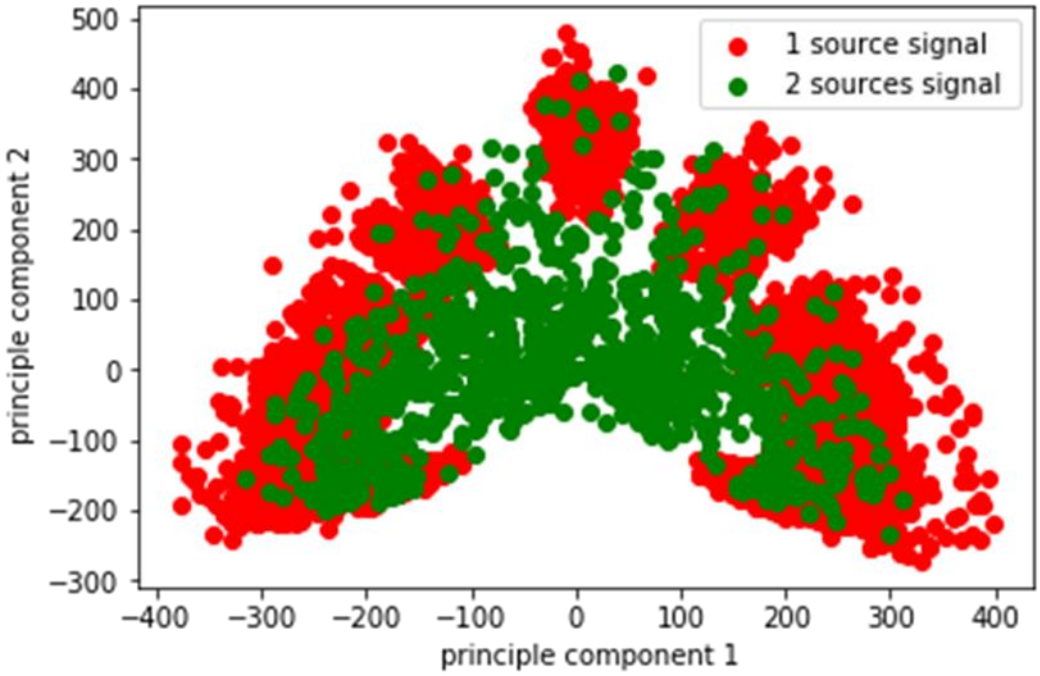


Figure 5.5: The PCA model used to visualize the correlation between the one source and two sources principle components.

5.4 Decreasing the Data Dimensionality

To reduce the memory requirements, the number of frequency bands was reduced. Figure 5.6 explains the procedure for feature dimensionality reduction. The firing rates over multiple bands are combined by averaging groups of frequency bands. Comparisons in performance were made when the average firing rates are evenly split into, four, two and one frequency band(s). According to the experimental outcomes, four gamma-tone bands were selected as this provides a good balance between reducing the memory requirements and supporting the localization performance when compared with the other tested bands, as discussed in the results section).

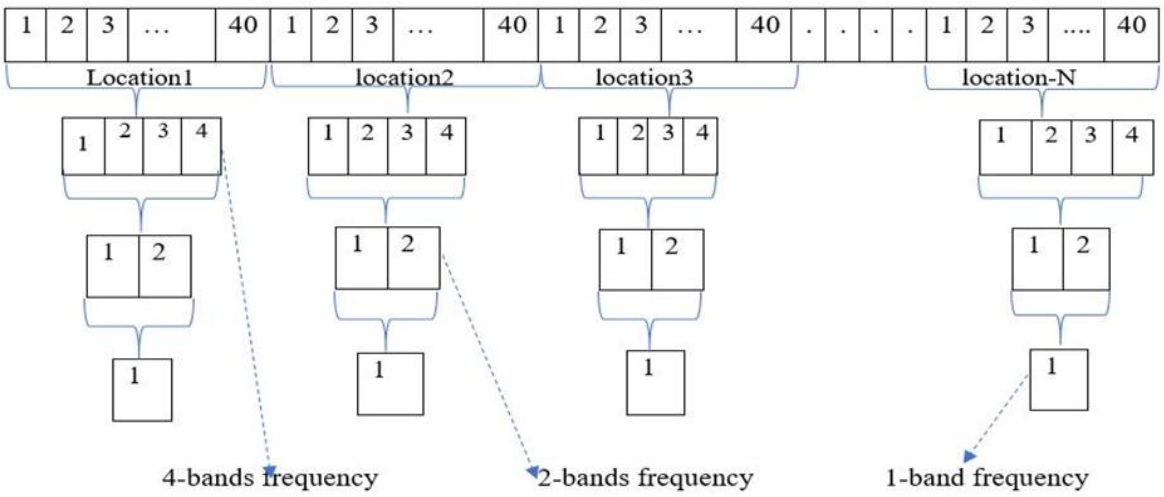


Figure 5.6: Gammatone frequency bands reduction process.

5.5 Multisource localisation by DNN

A supervised learning algorithm was used to solve the multi-sound source localization. A deep neural network was used to predict the source location of pairs of sources from the firing rate of the SNN. The classes were defined as every possible pair of source location for a given HRTF, in this respect a single classification operation yields the location of two sources. In the training phase, the deep neural network is trained to predict 4032 classes for the IRCAM data set and 4800 classes for the KEMAR data set. Each individual class contains two mixed sound instances (500ms) from two different speakers which are emitted from two various sources. These classes represent all possible sources mixed from the HRTF data sets. Each class has two labels, the labelling range starts from 0 to n, where n refers to the total number of locations in the HRTFs data set, so that the class headers take this sequence [((0, 1), (0, 2), …., (0, n)), ((1,0), (1,2), …, (1, n)), …… ((m, n))]. For example, first sequence represents all possible combinations between location 0 and other locations in the same HRTF data set. Similarity, this process is repeated for all locations (m, n). To decrease the memory requirement, the data is generated within the range of 7 elevation values (-45, -30, -15, 0, 15, 30, 45) for all azimuth ranges from the IRCAM set. Three elevation angles at -10˚, 0˚, 10˚ for azimuth ranging from 0° to 195° contribute to the data from the KEMAR set.

Initially, the size of the training data is checked for whether the training dataset provides a suitable amount of audio samples such that the classifier can learn notable features for each

class (pairs of locations). Training data are created from forty-six possible combinations for 10 speakers. The results were significantly bad and the localization accuracy for the both sources did not exceed the 30%. For better localization performance, the number of speakers is increased to 17. This produced a sufficient data size for classification task and the accuracy enhanced to reach 83%. The model was trained and tested by using the training data that was formed from 136 possible combinations for 17 speakers in the model's training phase for all possible location's combinations in the HRTF data set. The total size of the training data is 548352 rows (representing all possible location combinations for all possible speaker combinations), 7480 columns (representing the firing rate features resulted from applying spiking neural network on each location combinations). This becomes 548352 rows, 748 columns after applying dimensionality reduction of features by applying 4 bands of gamma-tone filtering rather than 40 bands. After setting the data size, the one-hot-encoder algorithm was applied to encode the classes and transform it from categorical form to binary form to match the machine learning supervised mapping requirements.

For critically analysing evaluation for the sources classification results, two accuracies were computed from the classification outputs. The accuracies were computed after rearranged the resulted locations to match the real time hearing process. The two sources are completely unknown then it is impossible to know which one the first source and which one is the second. So that, the initial locations predicted by the classifier have been processed by searching about the matched locations to bring them together. For example, the classifier predicts the two locations as (2,4) and the original locations were (4,3), the rearrangement locations process relies on reordered the predicted location to be (4,3) because the sources are predicted correctly but in wrong order.

5.5.1 Model description and parameters selection

In the previous section, the labelled training data preparation steps were described. In this section, the deep neural network topology and its parameters selection are illustrated. The deep learning neural network consists of five layers (an input and an output layer with three hidden layers) of nonlinearly-activating functions constructed as a deep neural network as shown in figure 5.7. The network is fully connected, each node in one-layer links with a specific weight w_{ij} to all neurons in the next layer. In this model, the input layer has 512 nodes and the

intermediate layers have 256 nodes in the first hidden layer and 64 nodes in the second one while the last hidden layer has 32 nodes. This structure arrived at after testing network parameters through the model selection process. All neurons in the input and hidden layers have a soft-plus activation function which is an analytic function defined as a smooth approximation to a rectifier. The soft-plus function produces output between 0 to infinity and it is mathematically described as follow:

$$softplus(\boldsymbol{x}) = \log(1 + e^{x}) \qquad (5.1)$$

The soft-max activation function, which is popular form multiclass classification methods, is applied to the output layer. The soft-max function outputs represent the probability distribution over all possible output classes. The number of neurons in this layer equals the number of classes. Here the number of classes relies on the angular resolution of the HRTF. The soft-max function formula is explained in equation 5.2 where an n-dimensional vector x of qualitative real values to an n-dimensional $\sigma\ (x_j)$ vector of real values in the domain of (0, 1) that sum up to 1 (Chung et al. 2016).

$$\sigma(x_j) = e_i^x(\sum_{i=0}^{n} e^{x_i})\ (for\ all\ j = 1\ to\ n) \qquad (5.2)$$

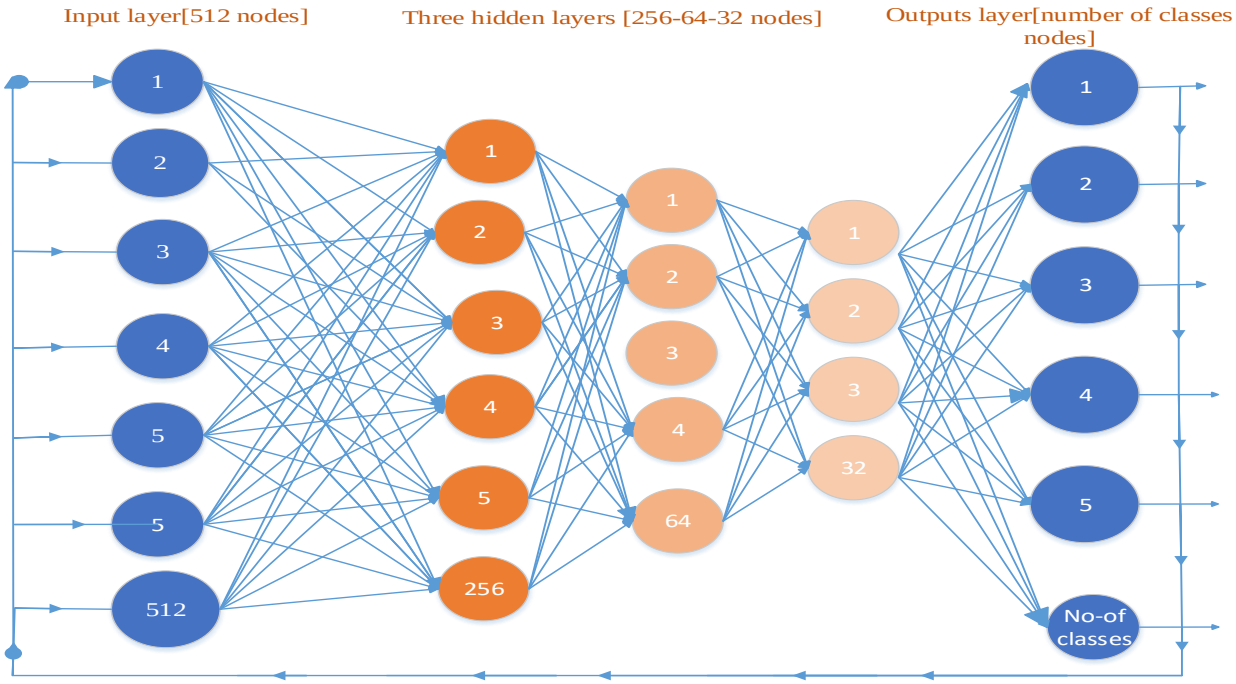


Figure 5.7: The deep neural network structure of the multisource sound localization model.

Various hyper-parameters were investigated to reach the final deep neural model structure. The number of hidden layers, number of nodes in each layer and the type of activation function was investigated to find the best performing model. The model selection process has been done experimentally to test number of model parameters for example, the number of hidden layers and the type of activation functions for each layer. The key parameter in setting the deep neural networks is the number of hidden layers. Table 5-2 shows the result of multiple trials that applied a various number of hidden layers. The findings demonstrate that preferable localization performance was achieved from the DNN with 3 hidden layers, the higher number of hidden layers were not able to enhance the localization accuracy, therefore; a DNN with 3 hidden layers is used in the in all experiments. Furthermore, the localization model with less or more hidden layers severed from unstable localization performance. The resulting model structure is shown in figure 5.6. The bottle neck shape of the suggested deep neural network overcomes an overtraining problem and the nonstable performance.

Table 5-2: The number of hidden layers in the deep neural network.

Number of hidden layers	Source one estimation accuracy ±15˚	Source two estimation accuracy ±15˚
1	**0.765**	**0.756**
2	**0.813**	**0.783**
3	**0.836**	**0.796**
4	**0.825**	**0.773**
5	**0.791**	**0.772**
6	**0.754**	**0.731**

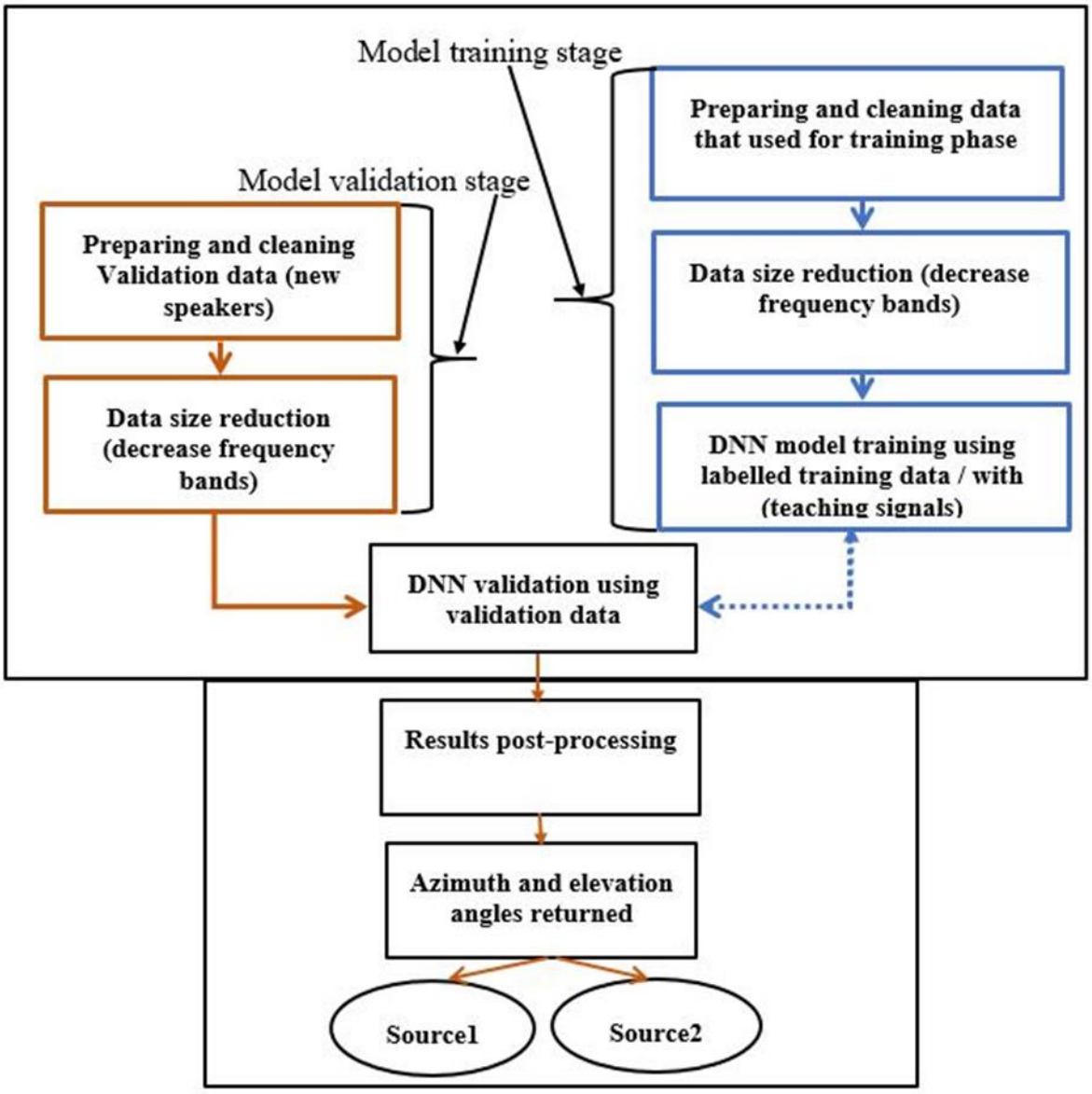


Figure 5.8: Multisource sound localization model training and validation stage.

5.5.2 Experimental results and discussion

In the previous subsections, the multisource localization model based on a DNN was described. In the following, the performance of the multisource localization model with IRCAM and KEMAR HRTF datasets are examined as shown in following experimental results. The confusion matrix is used to study the model performance in estimating the source directions. Also, the absolute angle error and signed angle error are figured to compute the localization accuracy for estimating each source.

Experiment 1: Comparing different gamma-tone frequency bands.

As previously mentioned, the number of frequency bands of gamma-tone filter bank is reduced for memory necessities. To investigate the optimal bandwidth to carry out the multisource localisation model, gamma-tone frequency bands were combined and the impact on reducing resolution on localization performance is reported in Table 5.3. The best localization accuracy is achieved when the 40 bands gamma-tone filter bank is reduced to four. The resultant feature dimensionality at this band is 748 and this will be fixed for all experiments in this chapter.

Table 5-3: Different gamma-tone bands impact on the multisource localization performance.

Gammatone Frequency Bands	Number of input features	Source one estimation accuracy ±15°	Source two estimation accuracy ±15°
1 band	**187**	**0.609**	**0.403**
2 bands	**374**	**0.786**	**0.641**
4 bands	**748**	**0.908**	**0.895**

The figures 5.9, 5.10 and 5.11 demonstrates the accuracy of localisation that results from predicting source one and source two using multisource localization model based DNN. This experiment was performed using training and validation data generated from IRCAM that used to train and validate the multisource localization model based on DNN. The signed angle error was computed for quantitative evaluation of the multisource prediction results. Figures 5.9, 5.10 and 5.11 demonstrate the frequency distribution of angle errors of the two sources that results from applying multisource localization model with IRCAM HRTF date set at each gamma-tone frequency bands.

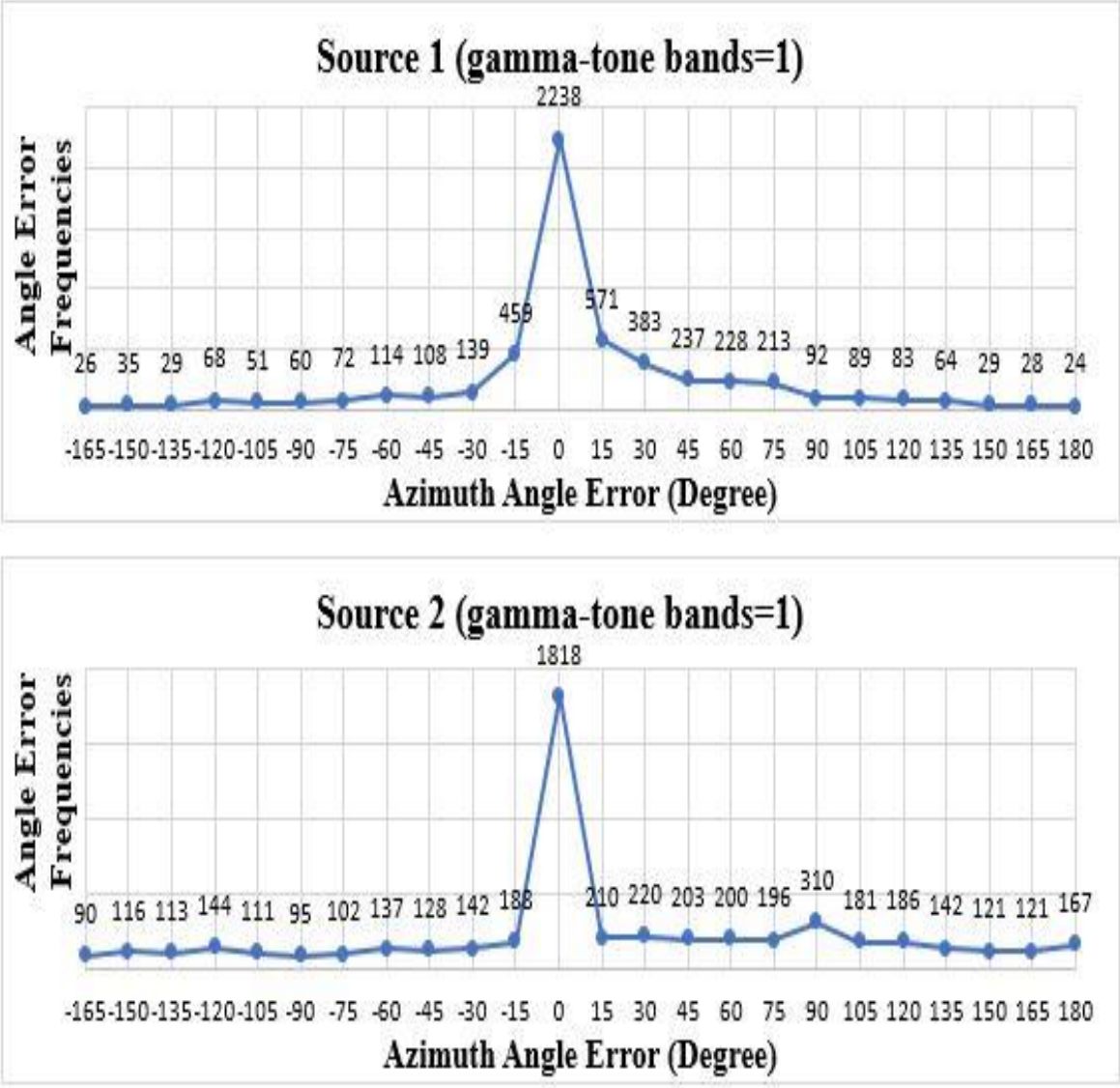


Figure 5.9: Angle error frequencies for source one and two with band=1.

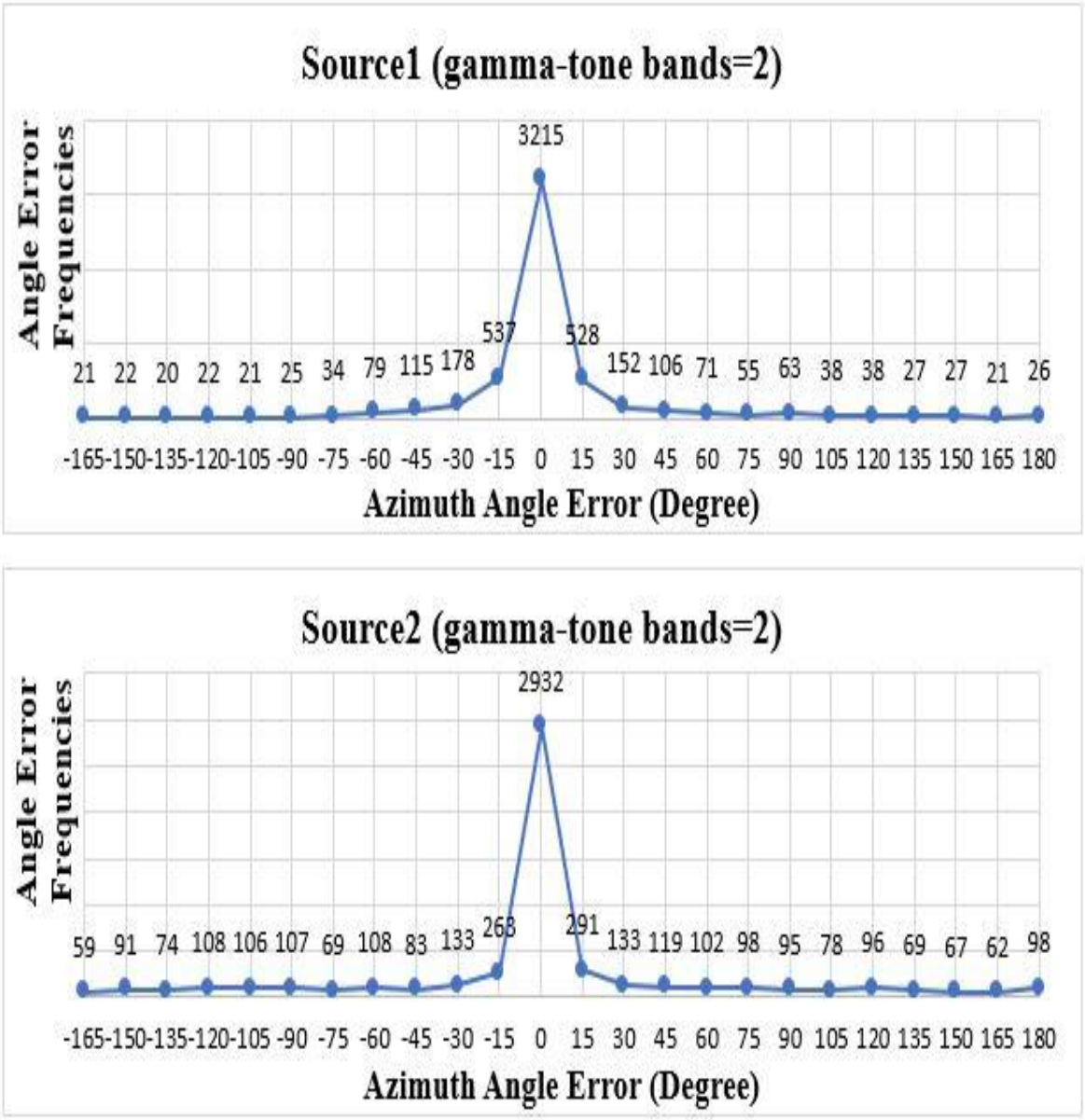


Figure 5.10: Angle error frequencies for source one and two with band=2.

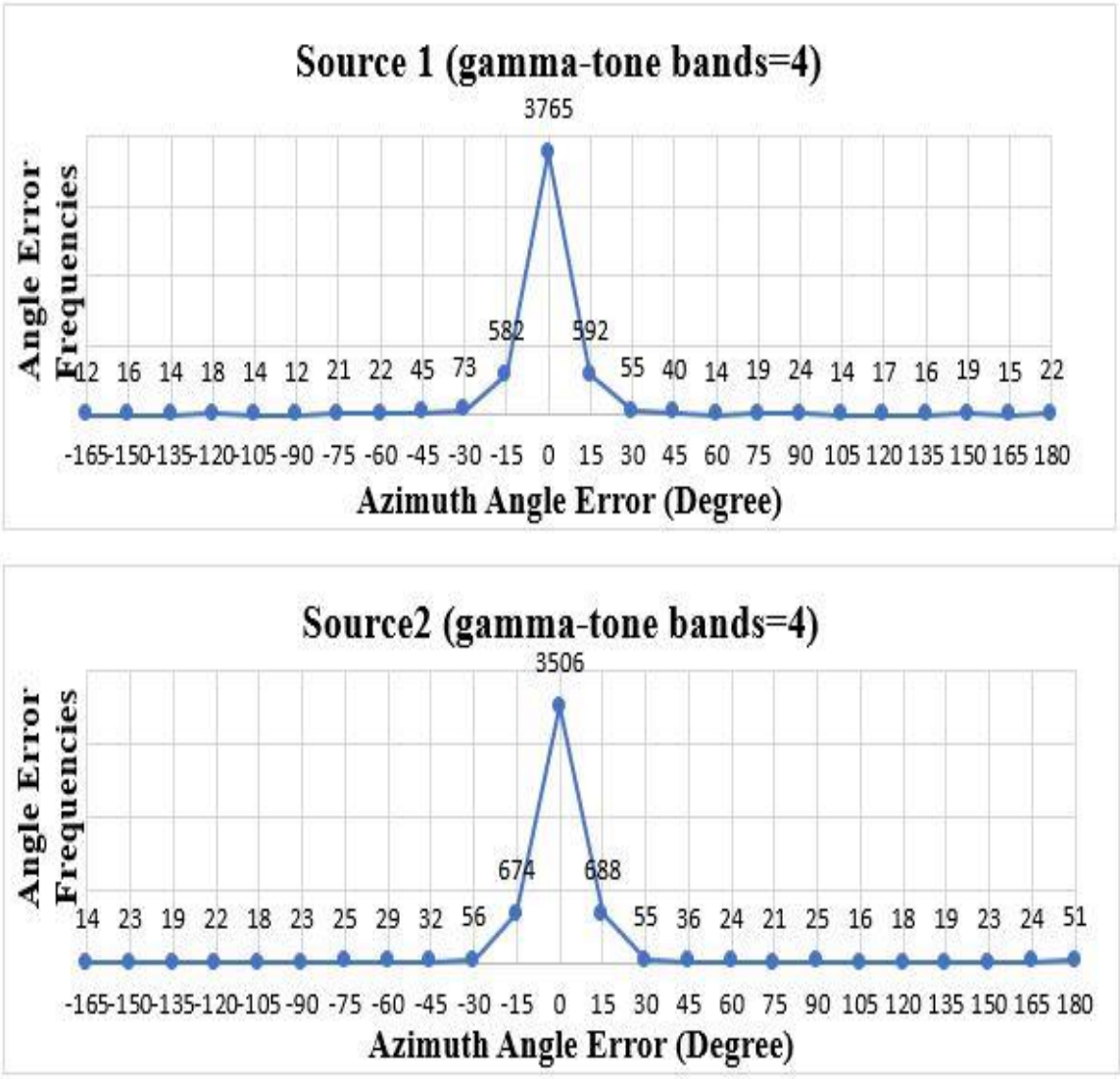


Figure 5.11: Angle error frequencies for source one and two with band=4.

In previous figures, the x-axis refers to the angle errors range from -165˚ to 180˚ with 15˚ increments. While, the y-axis represents the frequencies of each angle error from the total number of samples that used in this plot. Number of validation samples that used to plots is 5440. When the original angle is predicted correctly the angle error will be 0˚, meaning there is no difference between the original and predicted angle. For example, figure 5.11 demonstrates that the most locations are predicted correctly at 0˚ angle error 3765 times for source one and 3506 times for source two out of the total number of outputs samples. Also, it is notable that most of the error is frequent at ±15˚ from the actual angle. And, this represent the best

localization accuracy have been achieved when the 40-band gamma-tone filter bank is reduced to four compared with the other bands that showed in the figures 5.9 and 5.10.

Experiment 2: the multisource model trained and validated with IRCAM HRTF

The multisource localization model tested with data that generated using IRCAM dataset. Figures 5.12 and 5.13 show the confusion matrix plots of predicted azimuth of source one and two; both figures are resulted from the IRCAM data set with validation speakers.

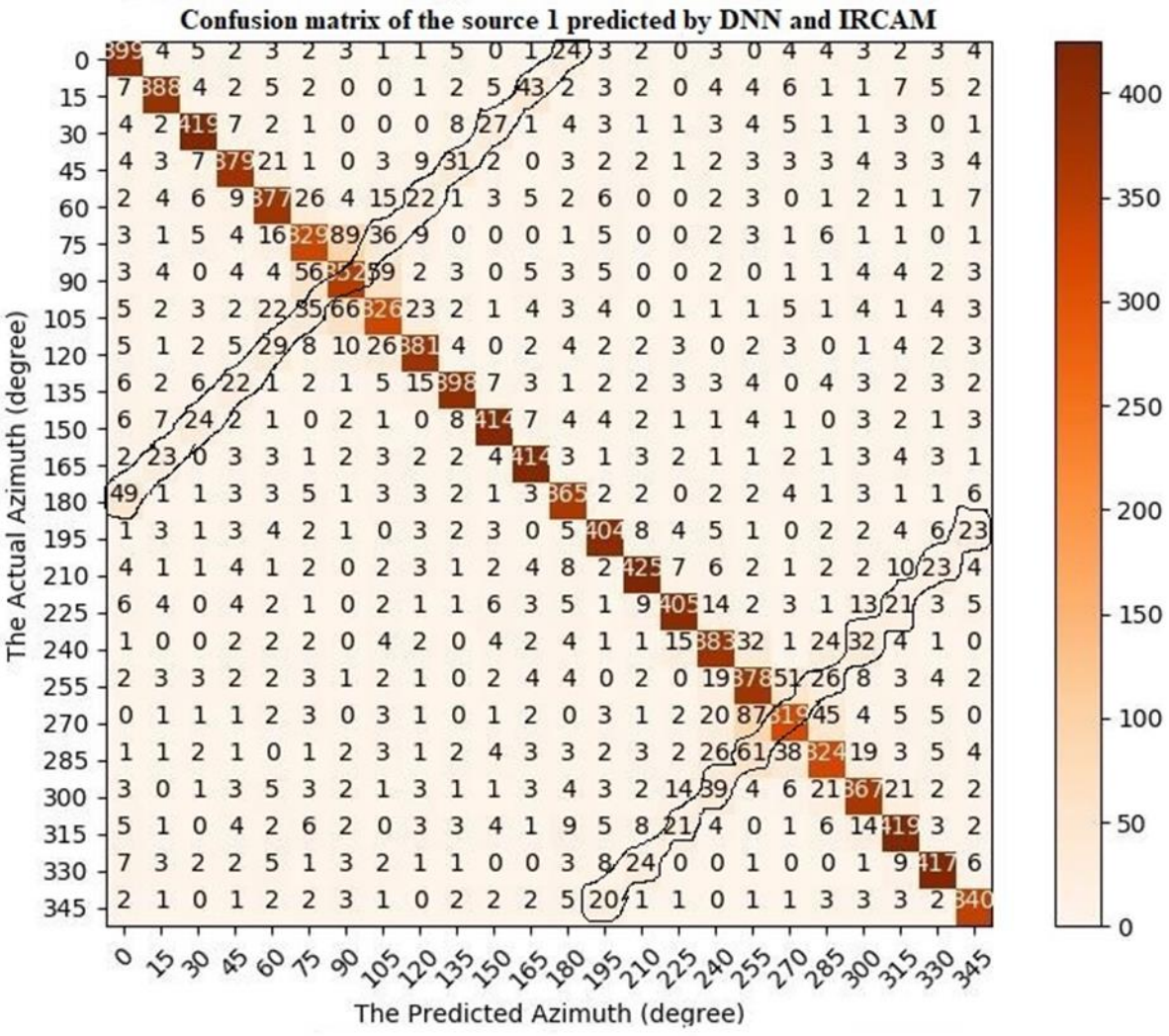


Figure 5.12: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation model on IRCAM HRTFs and validation speakers.

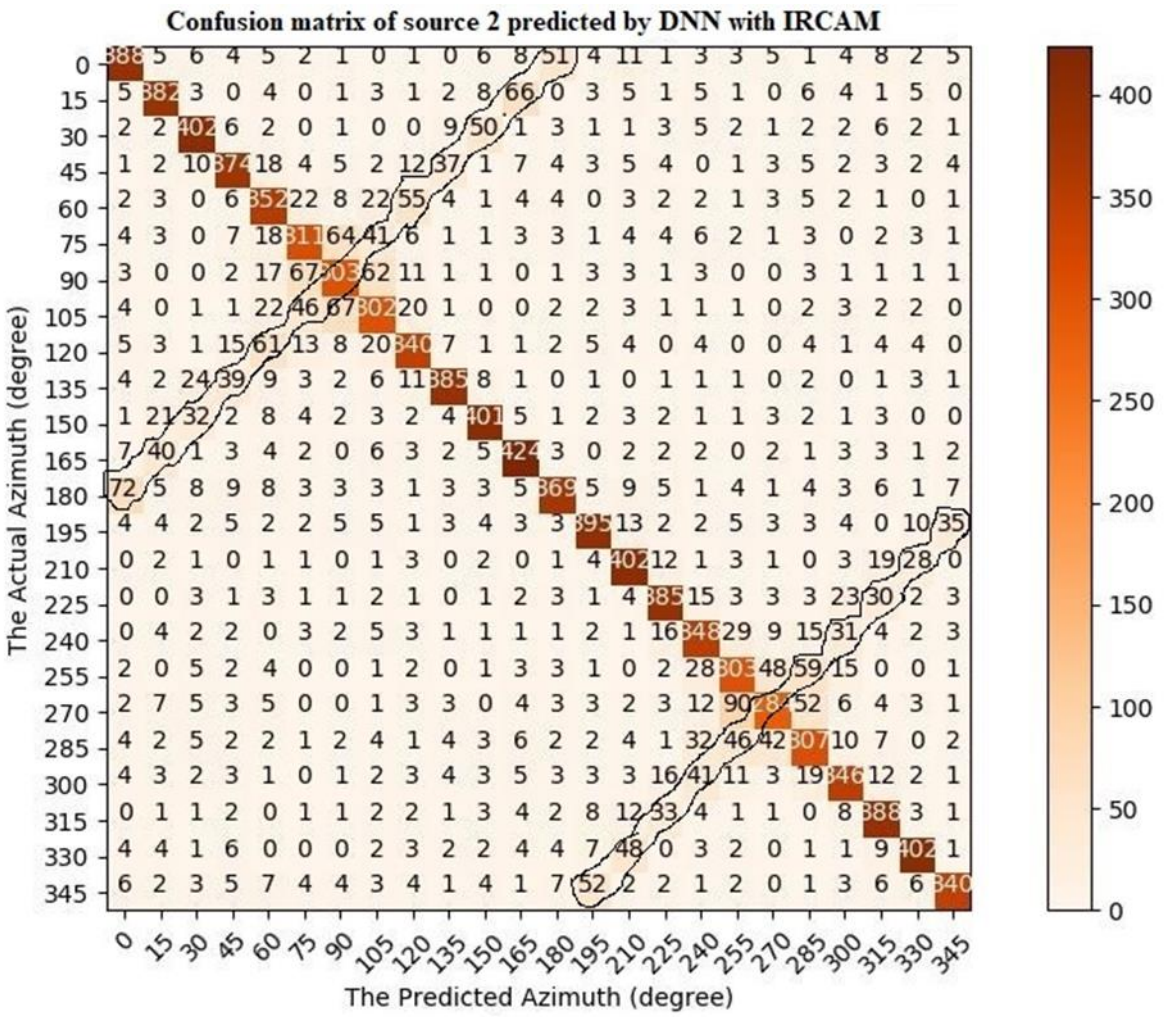


Figure 5.13: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation model on IRCAM HRTFs and validation speakers.

In these confusion matrix plots, the x-axis refers to the predicted azimuth angles in degrees and y-axis refers to the actual azimuth angles. Azimuth angles have the range from 0˚ to 345˚ in 15° increments. The diagonal line refers to the number of angles that predicted correctly from the entire number of validation samples 11955 fresh samples that are used to validate the multisource localization model in total. The front-back confusion appears clearly in the source1 and source2 confusion metrics. The error points that represents the front-back confusion in the above figures have been marked. The error points bounded between 180˚ and 0˚ on the y axis that represent the locations in the front side. While, the error points between 195˚and 345˚ represent the locations in the back side. It's clear that the error points that represent the front-back confusion are symmetrical along the front and back sides. Furthermore, the absolute angle

error between the original and predicted locations are computed to evaluate the model performance in detecting the two sound sources in the input signal. Figure 5.14 and 5.15 illustrate the confusion matrix plots of estimation angle errors for source one and source two.

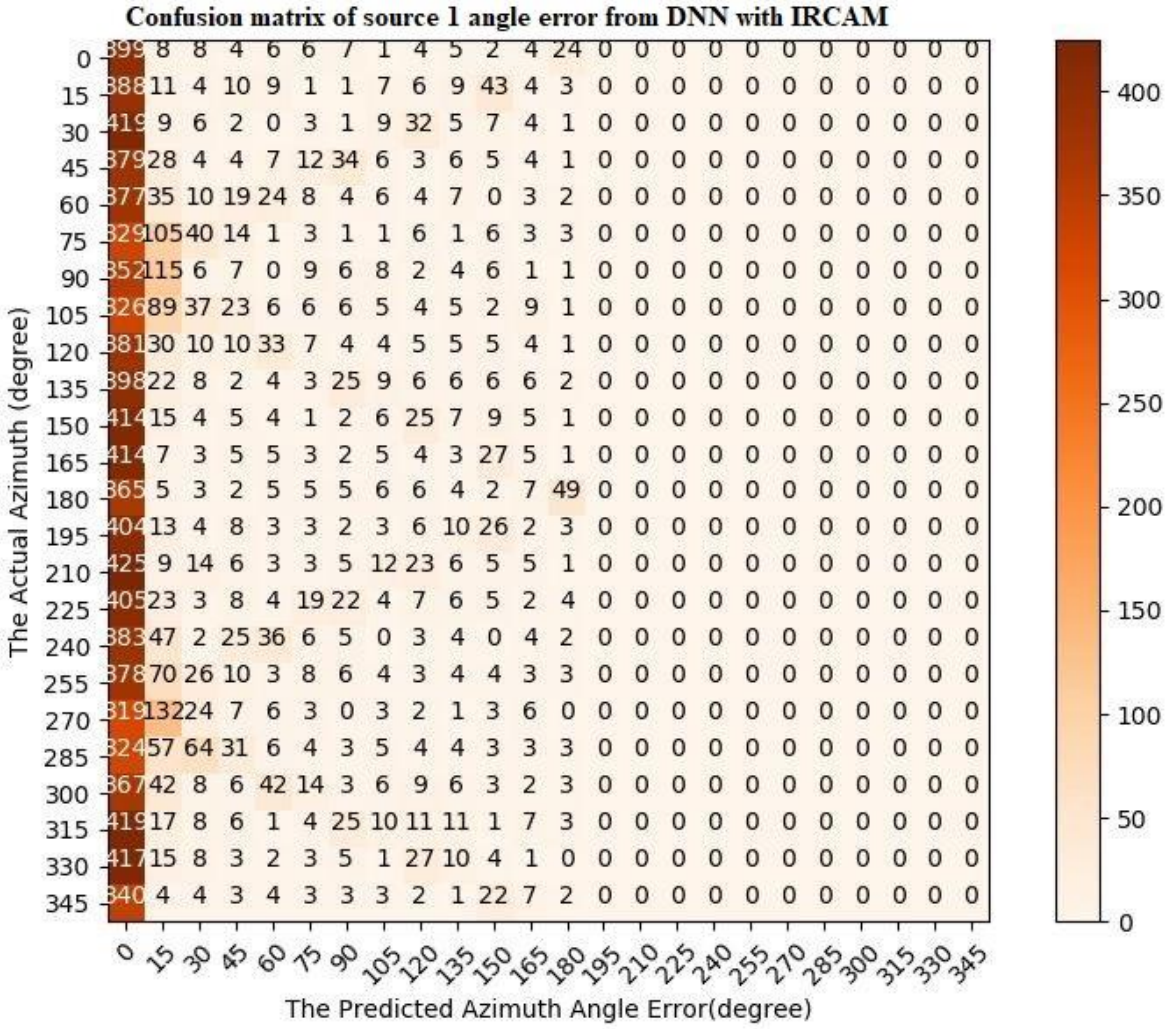


Figure 5.14: The sources one azimuth angle errors from applying multisource localisation model on IRCAM HRTFs with validation speakers.

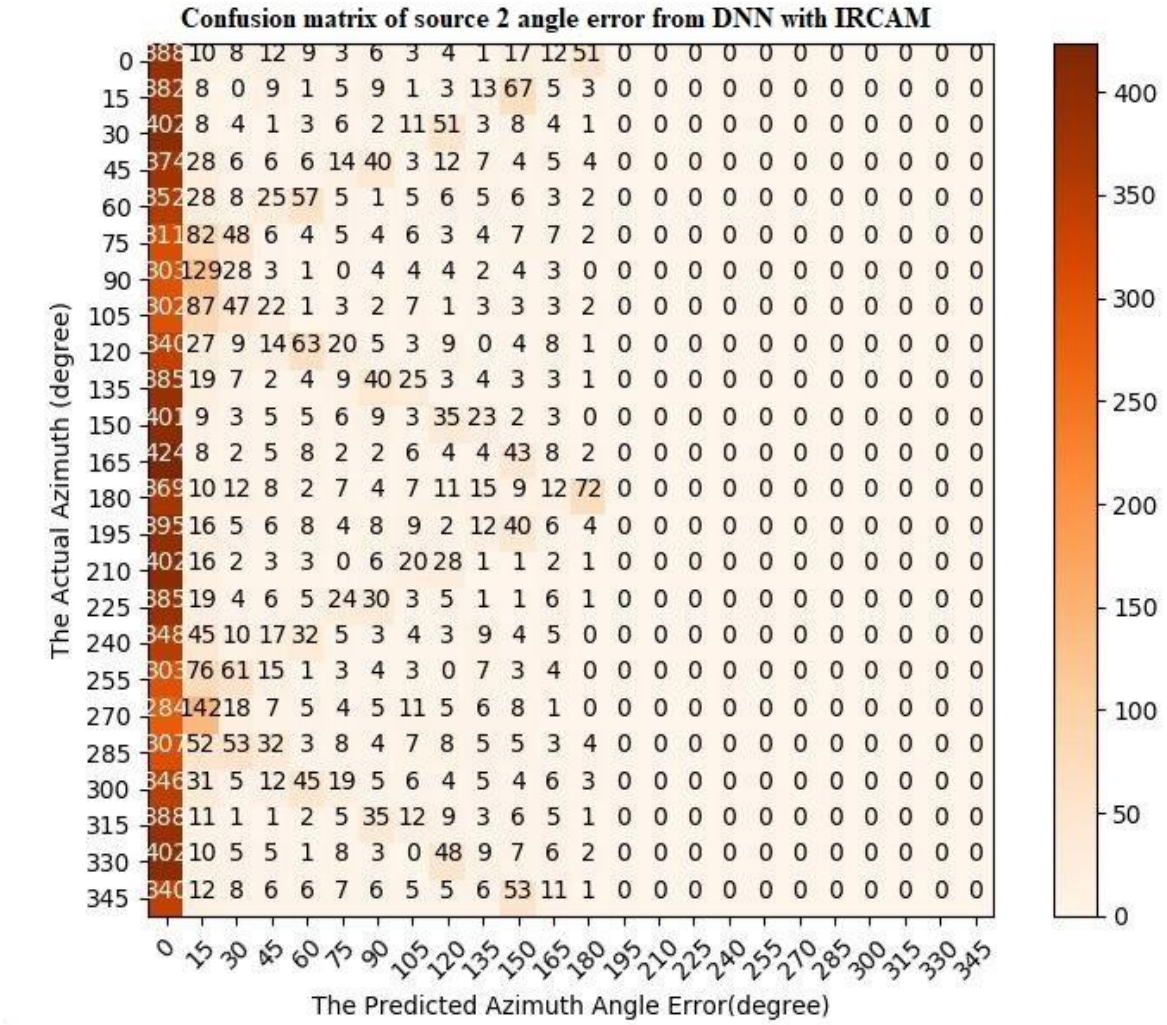


Figure 5.15: The sources two azimuth angle errors from applying multisource localisation model on IRCAM HRTFs with validation speakers.

In these figures, the x-axis refers to the predicted azimuth angle error in degrees while the y-axis refers to the actual azimuth angles. Obviously, the confusion matrix plots of angle error plainly show that the maximum error angle is 0˚ and most of error at 15˚ away from the actual angle. The front-back confusion is also demonstrated in the confusion matrix of azimuth angle error and it is clearly represented in increasing the error in angle 180˚ particularly in the angle error plot of source two. The angle error plots demonstrate the symmetrical angle errors along the front and back sides that take the shape of the sigma symbol (Σ) which refers to the

front back confusion. Figure 5.16 and 5.17 demonstrate the confusion matrix plots of elevation angles prediction performance for both sources. The elevation angles have range from -45° to 45° in 15° increments. In these confusion matrix plots, the x-axis refers to the actual elevation angles while the y-axis refers to the predicted elevation. Generally, the model appears to have a good effectiveness at estimating elevation angles, as clearly shown in the diagonal lines for the source1 and source2 confusion matrix plots. These plots included all azimuths from IRCAM data set at range 0˚ to 345˚.

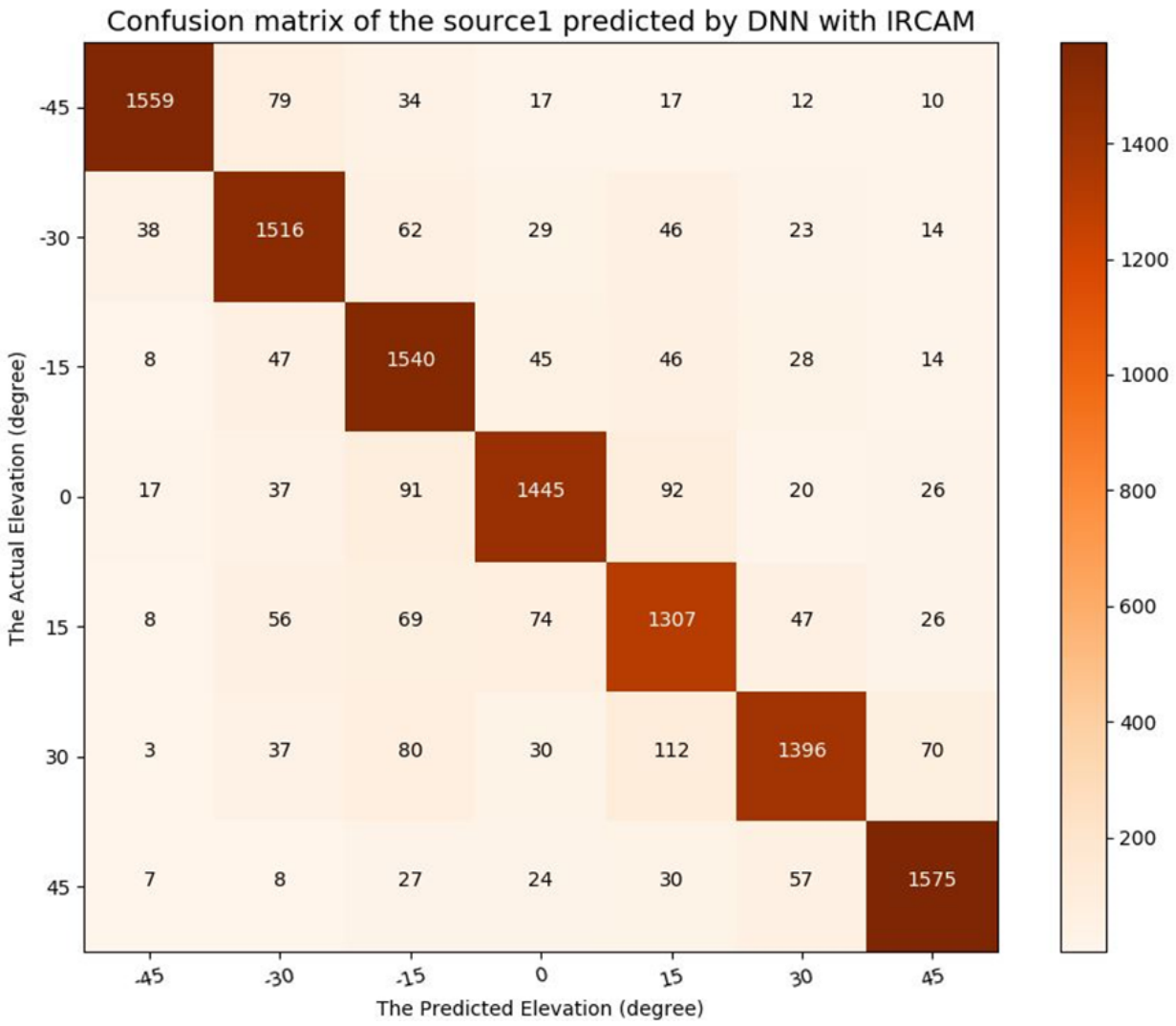


Figure 5.16: The confusion matrix plot for the source one elevation angles predicted by multisource localisation model on IRCAM HRTFs and validation speakers.

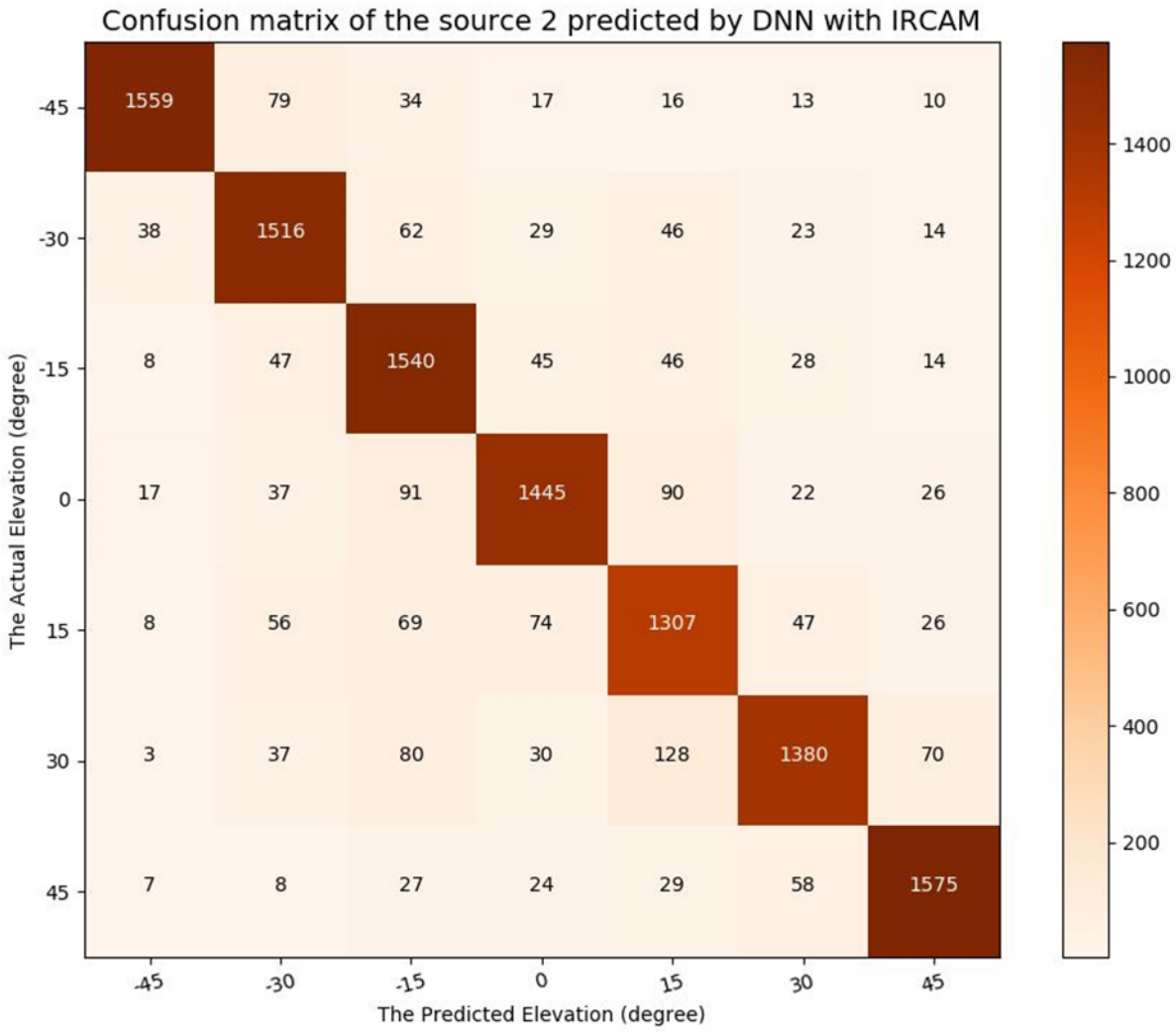
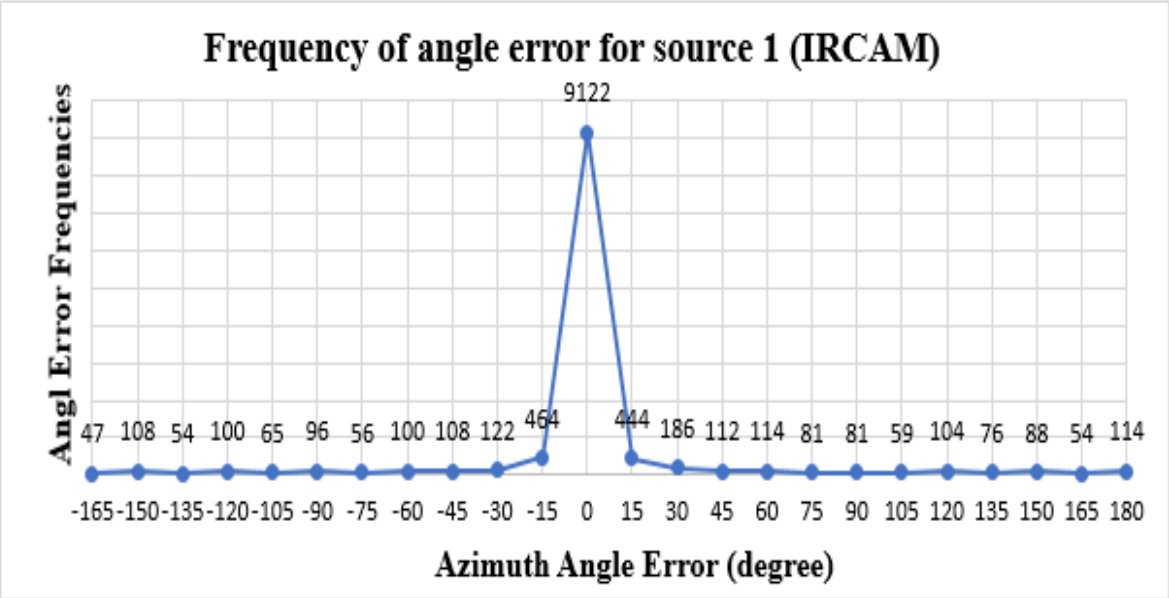


Figure 5.17: The confusion matrix plot for the source two elevation angles predicted by multisource localisation model on IRCAM HRTFs and validation speakers.

The signed angle error was computed for quantitative evaluation of the multisource prediction results. Figure 5.18 explains the frequency distribution of angle errors of the two sources that results from applying multisource localization model with IRCAM HRTF date set. The x-axis refers to the angle errors range from -165˚ to 180˚ with 15˚ increments. While, the y-axis represents the frequencies of each angle error from the total number of samples that used in this plot. Number of validation samples that used to plots is 11955. When the original angle is predicted correctly the angle error will be 0˚, meaning there is no difference between the original and predicted angle. The figure demonstrates that the most locations are predicted correctly at 0˚ angle error 9122 times for source one and 8633 times for source two out of the

total number of outputs samples. Also, it is notable that most of the error is frequent at ±15˚ from the actual angle.



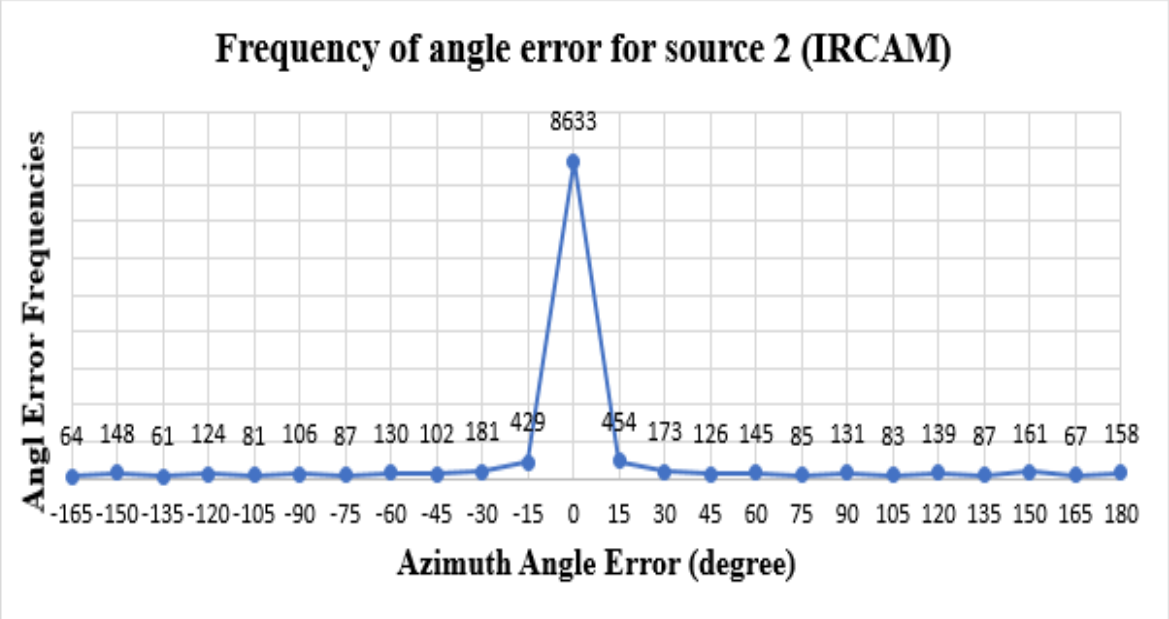


Figure 5.18: Bell shape explains the angle error frequencies for source one and source 2 predicted by DNN with IRCAM HRTF and validation speakers.

Experiment 3: the multisource model trained and validated with KEMAR HRTF data set

In previous experiment, the multisource localization model performance with IRCAM HRTF was investigated. This section looks at applying the multisource localization model with the KEMAR dataset. The main advantage for this test is to investigate the localization model with different HRTFs anatomical parameters and different measured environment. Furthermore, KEMAR has a relatively large number of measured angles which provide variety of locations for generalize and extend the testing process. As mentioned in chapter 3, KEMAR HRTF refers to the dummy head while IRCAM HRTF refers to the human male subject. Both datasets contained distinct sets of azimuth and elevation measurements. The KEMAR dummy head dataset has 710 locations with unequal increments between azimuth angles along vertical plane. The minimum distance between locations in KEMAR HRTF data is 5° along azimuth angles and 10° along elevations angle (horizontal plane). In contrast, the IRCAM HRTF data set has 187 locations with regular increments by 15° in both vertical and horizontal planes. Figures 5.19, 5.20 show the confusion matrix plots of predicted the azimuth of source one and source; both figures are from the KEMAR data set with validation speakers.

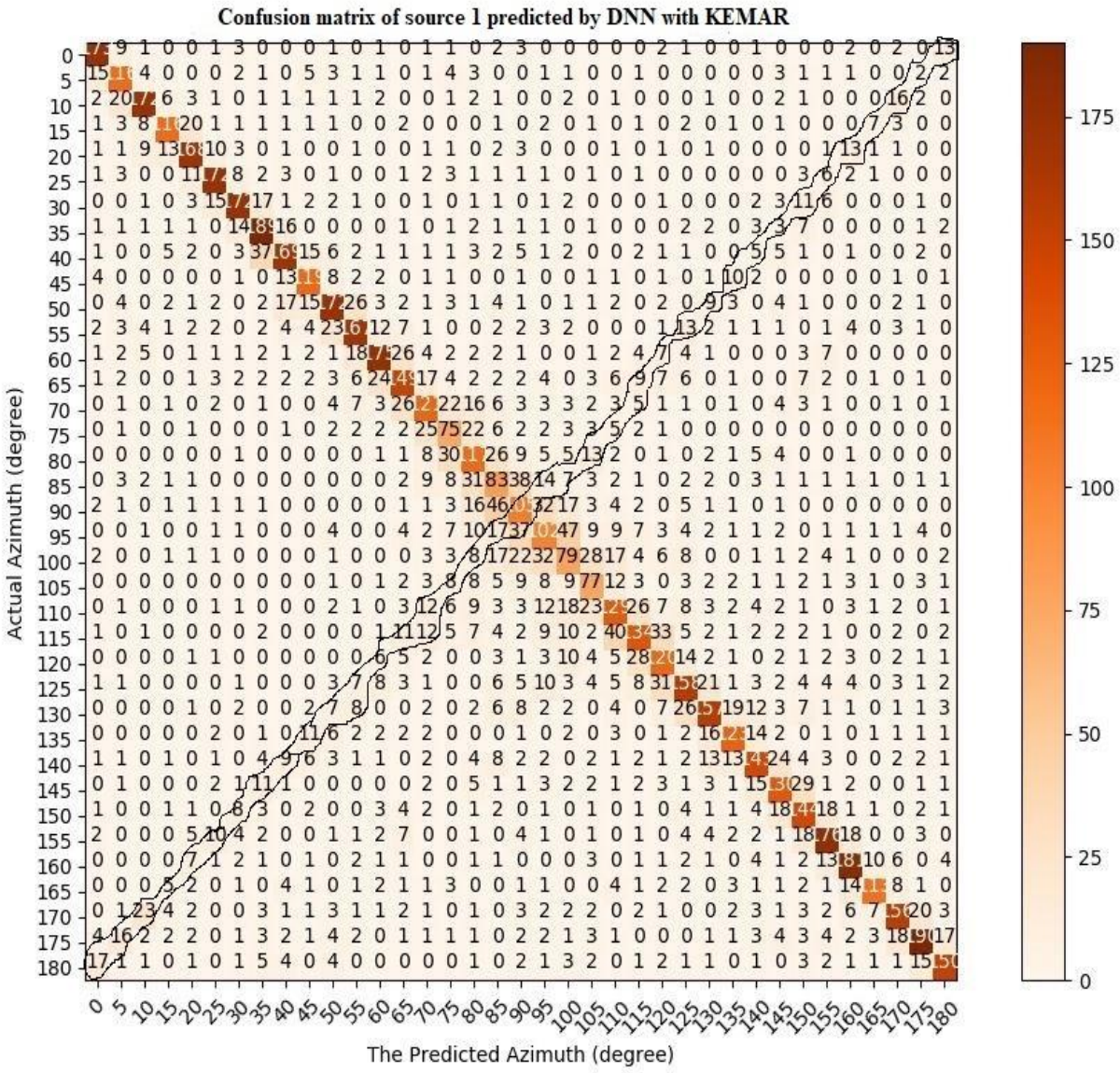


Figure 5.19: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation model with KEMAR HRTFs and validation speakers.

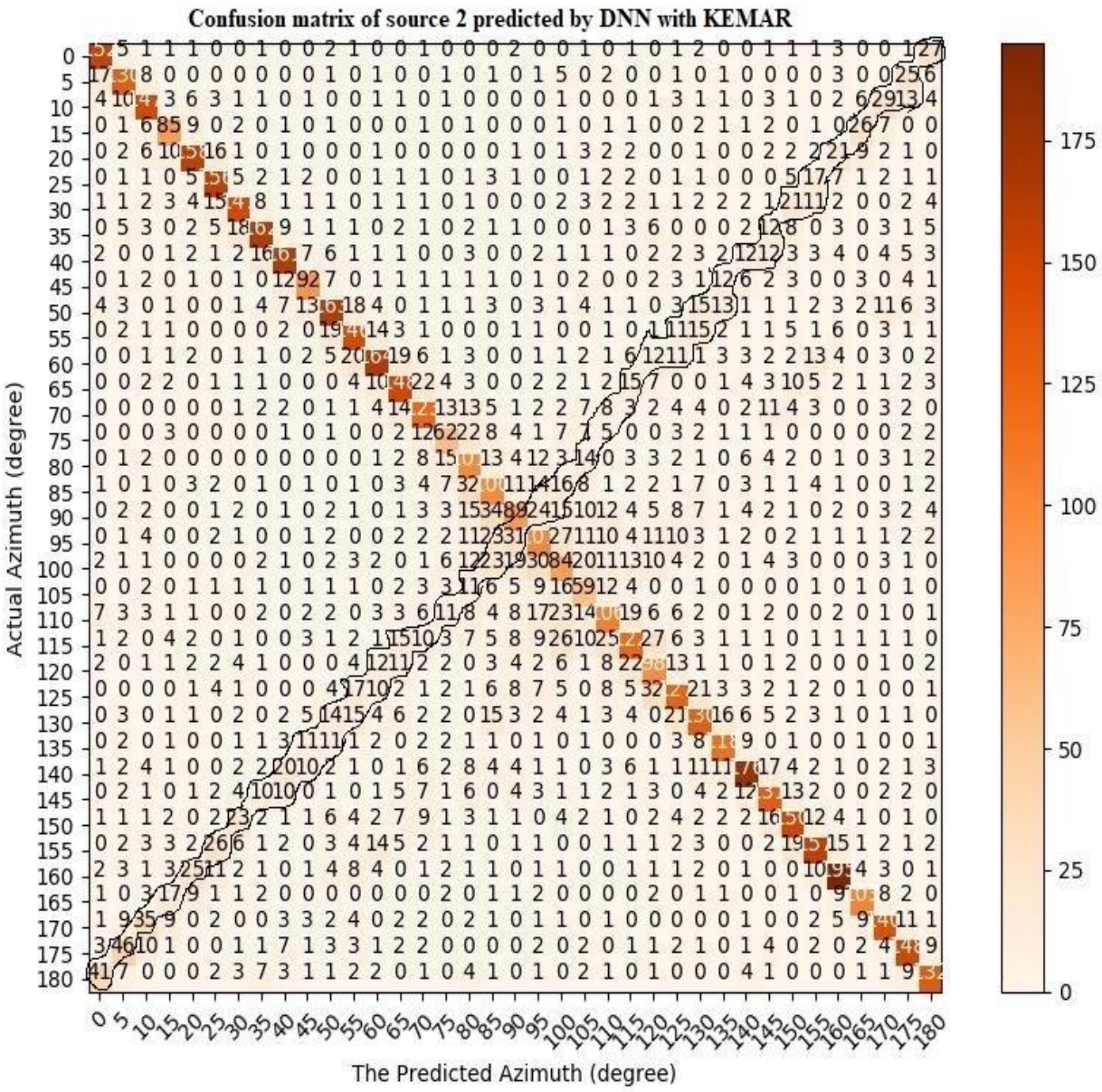


Figure 5.20: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation model with KEMAR HRTFs and validation speakers.

In these confusion matrix plots, the x-axis refers to the predicted azimuth angles and y-axis refers to the actual azimuth angles. Azimuth angles range from 0˚ to 180˚ at 5° increment steps. The total number of samples that were used in this test is 11955 output points resulted from model validation stage. The figures 5.19 and 5.20 show only the azimuth angles between 0˚ to 180˚ from. These angles were used to train and validate the DNN with KEMAR HRTF

database. The front-back confusion illustrated in the source1 and source2 confusion metrics in the error points have been marked between 0˚ and 180˚.

Figures 5.21 and 5.22 explain the absolute angle error between the original and predicted locations.

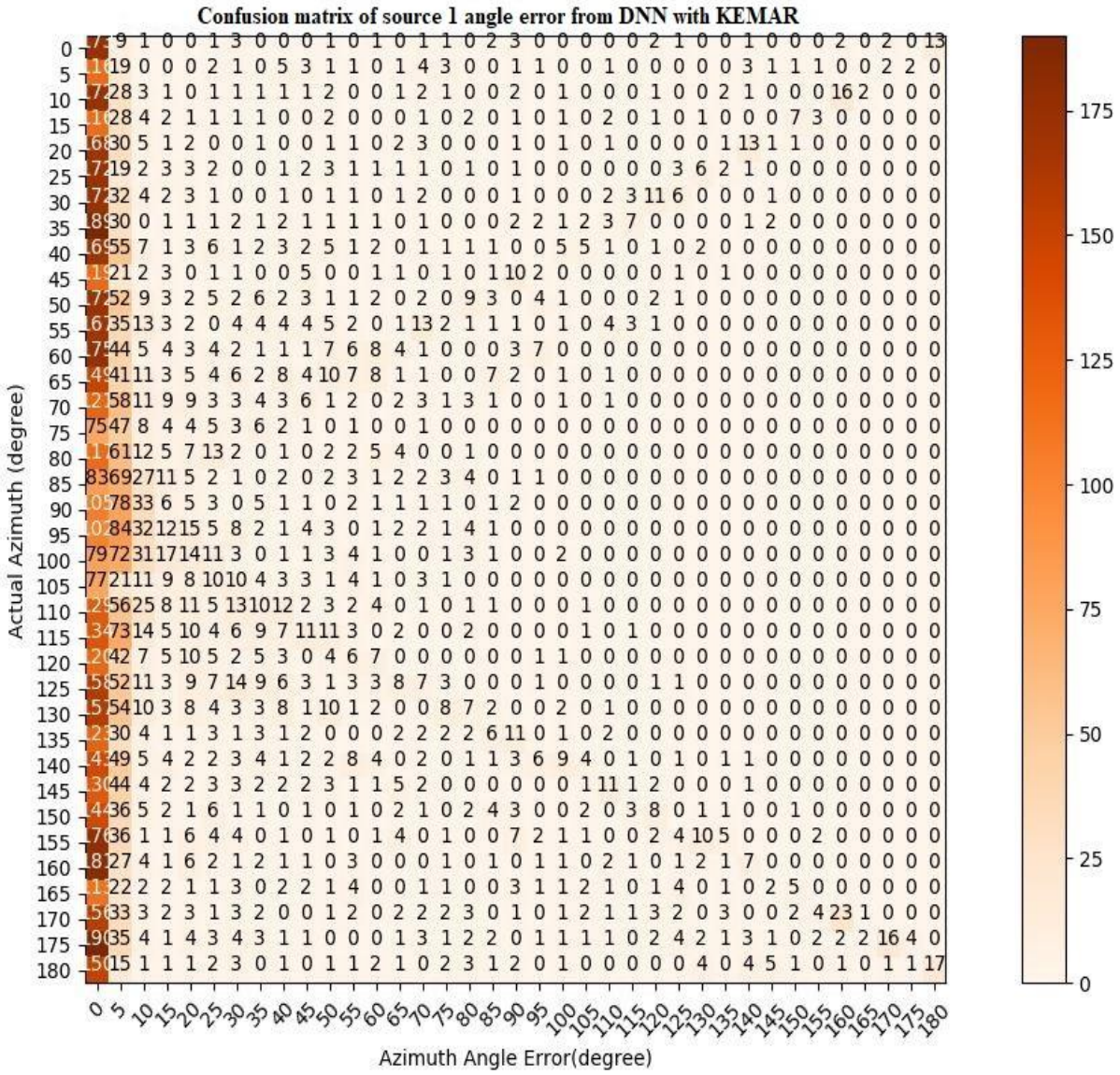


Figure 5.21: The source one azimuth angle errors from applying multisource localisation model on KEMAR dummy head with validation speakers.

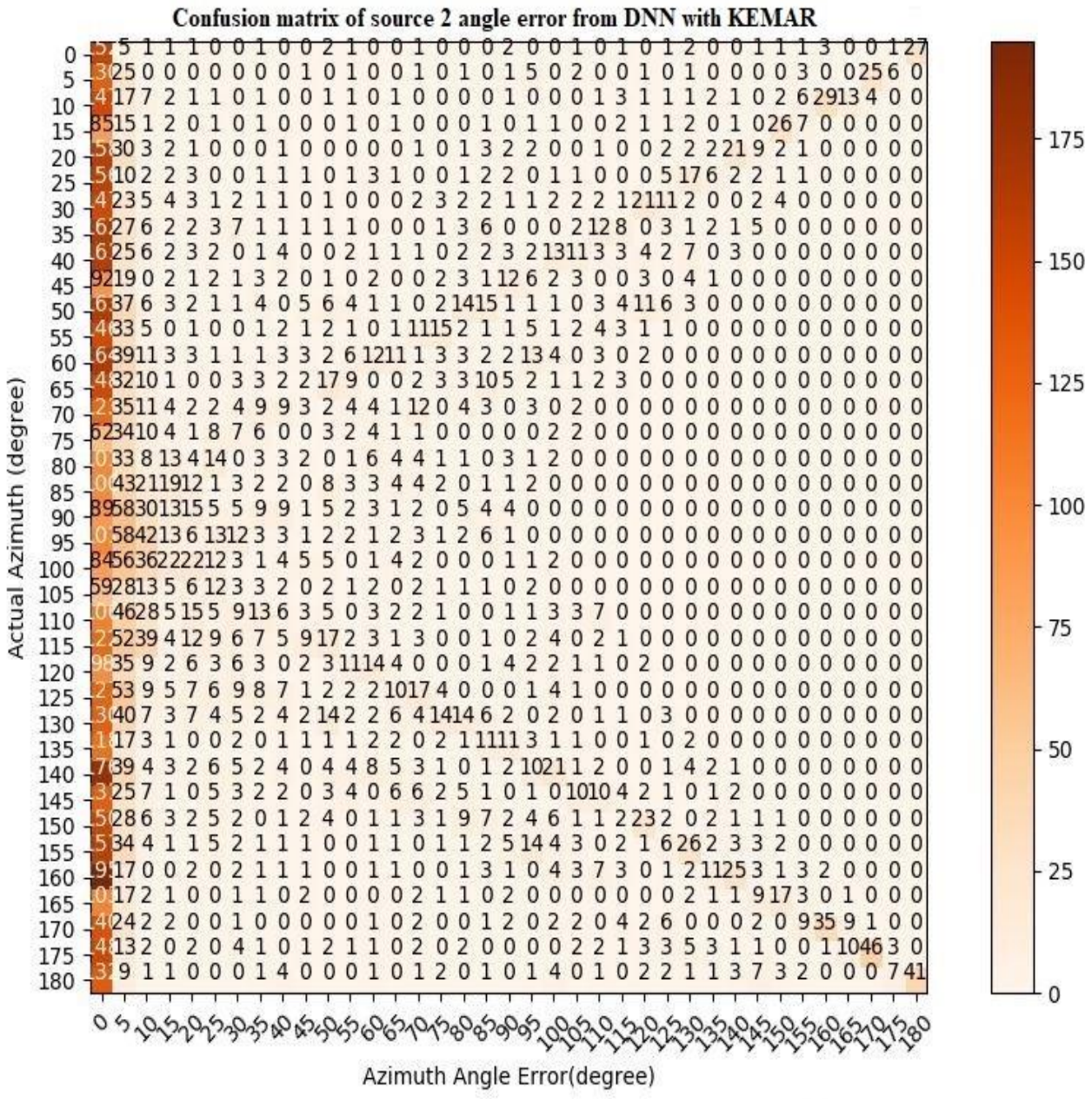


Figure 5.22: The source two azimuth angle errors from applying multisource localisation model on KEMAR dummy head with validation speakers.

In these plots, the x-axis refers to the predicted azimuth angle error in degrees while the y-axis refers to the actual azimuth angles. The confusion matrix plots of angle error demonstrate that the most angles have been predicted correctly with 0˚ angle error. Then, the angular errors increased at 5˚ away from the actual positions. For example, in the figure 5.21, sources from

angle 180˚ have been predicted correctly 150 times with 0˚ angle error when the source was the location one. For location two, the sources with angle 180˚ have been correctly predicted only 132 times with 0˚ angle error as established in figure 5.22. Also, the angle error plots demonstrate the symmetrical angle errors between 0˚ and 180˚ that take the shape of the symbol (<) which refers to the front back confusion.

Figure 5.23 and 5.24 demonstrate the confusion matrix plots of three elevation angles (-10°, 0°, 10°) for source one and source two that predicted from applying the multisource localization model with KEMAR HRTFs. In these confusion matrix plots, the x-axis refers to the predicted elevation angles while the y-axis refers to the actual elevations.

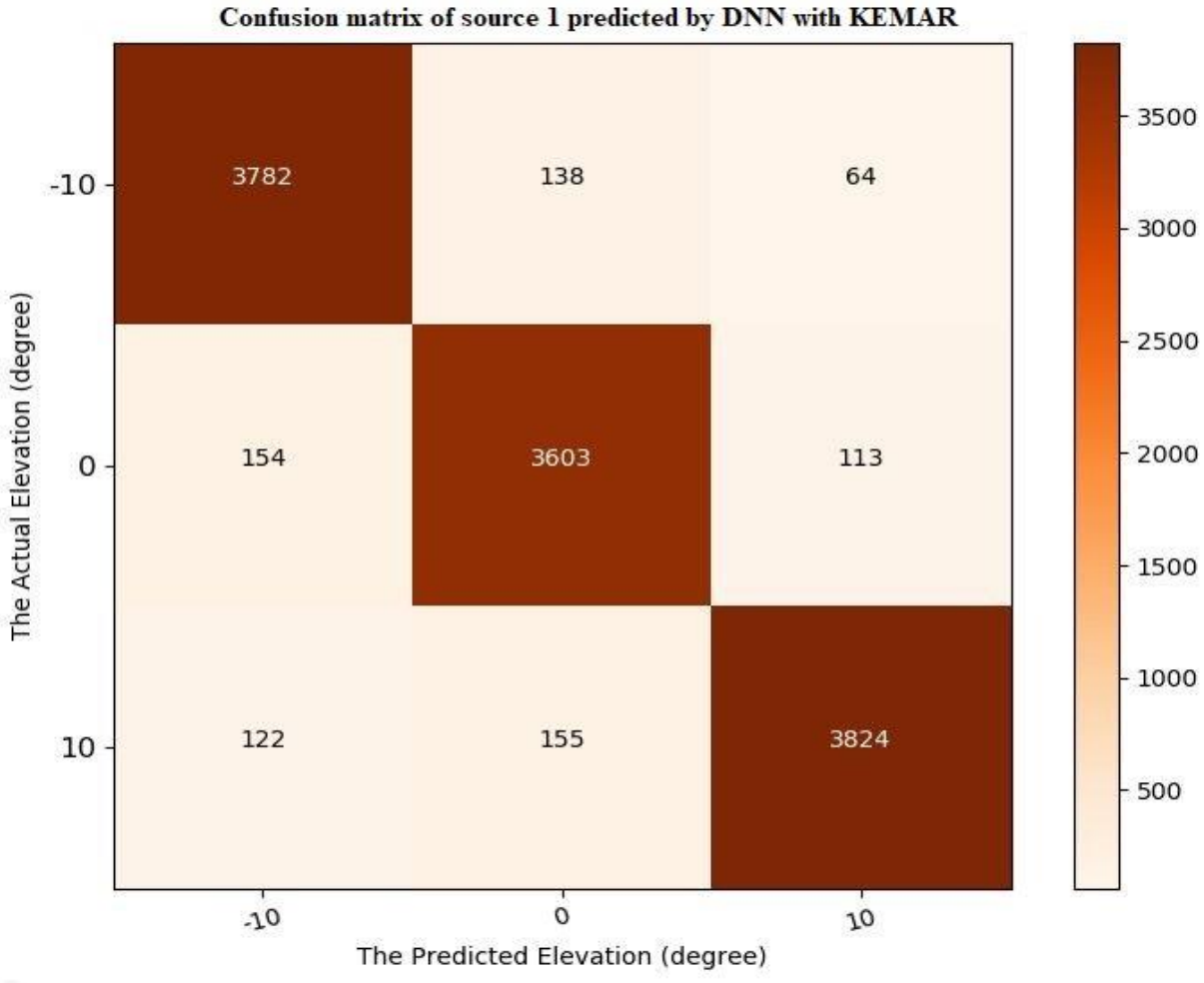


Figure 5.23: The confusion matrix plot for the source one elevation angles predicted by multisource localisation model with KEMAR HRTFs and validation speakers.

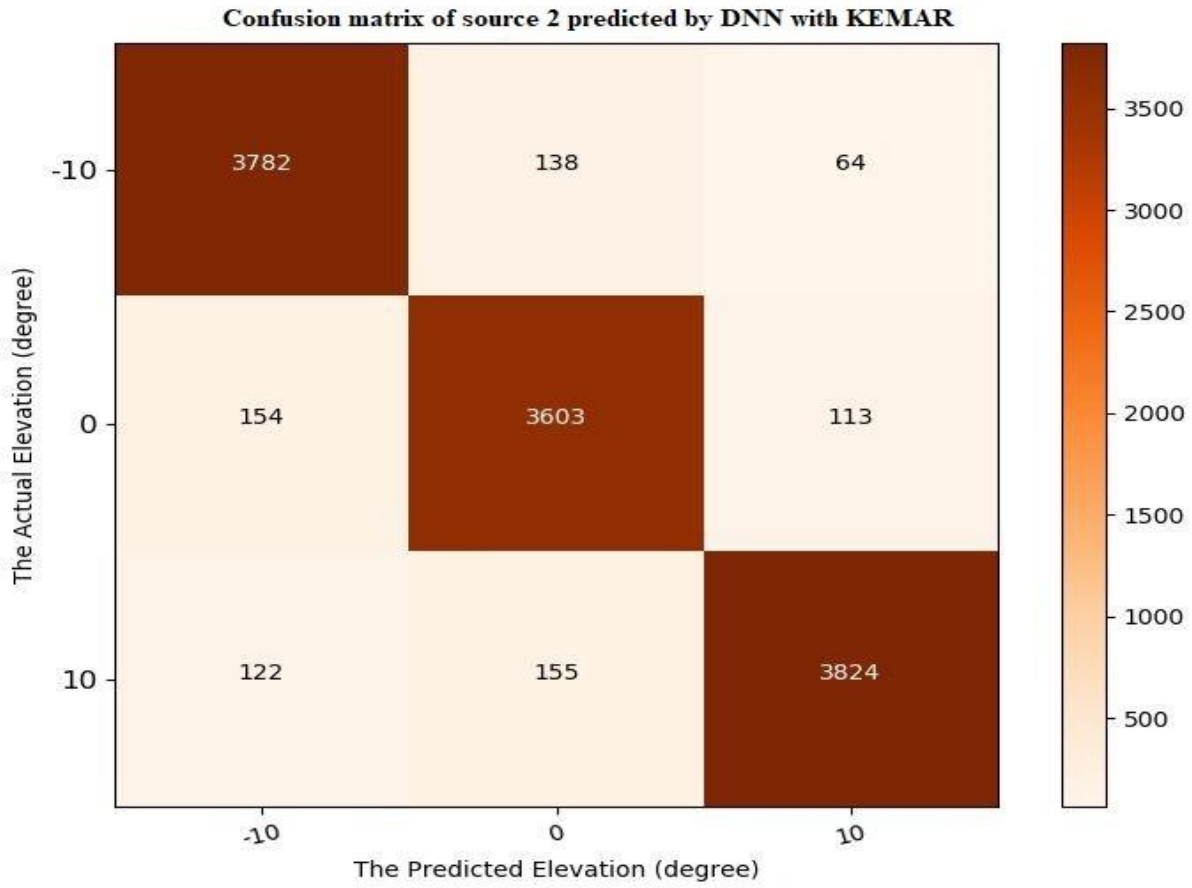


Figure 5.24: The confusion matrix plot for the source two elevation angles predicted by multisource localisation model with KEMAR HRTFs and validation speakers.

The signed angle errors between the actual and predicted angles for source one and source two have been computed. Figure 5.25 explains the frequency distribution of angle errors of the two sources that results from applying multisource localization model with KEMAR HRTFs. The x-axis refers to the angle errors range from -165˚ to 180˚ with 5˚ increment step. While, the y-axis represents the frequencies of each angle error from the total number of samples that used in this plot. Number of validation samples that used to plots is 11955. The figure demonstrates that the most locations are predicted correctly with higher peak at 0˚ with 6965 output points for source one and 6405 output points for source two out of the total number of outputs samples. Furthermore, the source one plot shows that the most angle error is at -5˚ with 1086 output points and at +5 with 1006 output points. Then, at -10˚ with 250 output points and +10 with 214 output points. Besides, the number of angles that predicted with angle error -15˚ is 127 output

points and with +15 is 118 output points. For source two, the angle error was relatively high with most error at the angles ±5˚, ±10, ±15, and +180. The growing in the error angle at +180˚ refers to the front-back confusion phenomenon.

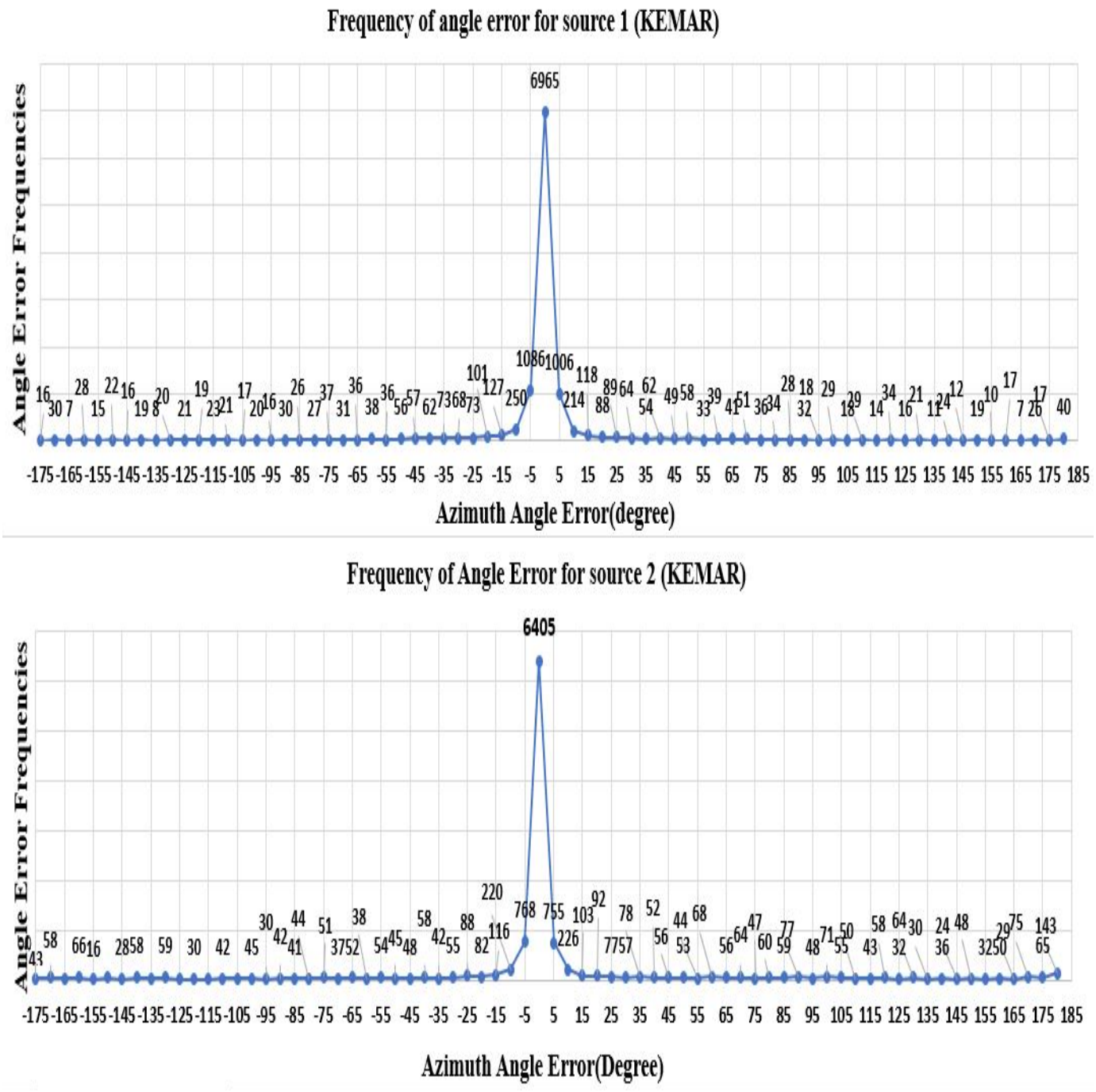


Figure 5.25: Bell shape explains the angle error frequencies for source on and source two predicted by DNN with KEMAR HRTF and validation speakers.

Consequently, after running the multisource localization model on the data that was generated using various angles from the IRCAM and KEMAR data sets, the model exhibited a reliable ability to detect the sources in an incoming signal and separate between them independently.

5.6 Test the Multisource sound localization performance in individual elevation angles using DNN and SVM.

In the previous chapter, the support vector machine has been investigated in localizing the single sound source two times; first by processing the gamma-tone filter bank features to predict the sound source. And secondly, a support vector classifier used to process the SNN firing rate features to predict the incoming signal location. In the second method, two different sizes of data have been generated and the results demonstrated that the support vector classifier was able to solve the single source localization problem when enough training data was available. In the following section, the support vector machine (SVM) as multiclass classifier function was examined to examine its effectiveness to sort out a multisource localization problem. The multisource localization model based SVM with linear kernel is applied to test its ability in detecting and separating the two sources in the incoming signals.

In this experiment, training data has been generated from all possible pairs of 17 speakers (9 males and 8 females) which are simulated at all possible pairs of angles each elevation separately. The data is created individually for each single elevation angle along 24 angles in the horizontal plane resulting in 576 classes for each elevation value. Similarly, the validation data was generated from all possible pairs of 3 speakers (2 males and 1 females). The SVM classifier was trained and validated to predict the 576 classes that represents different source combinations. Table 5-4 shows the estimation results of azimuth angles in each elevation level from applying the SVM and DNN with IRCAM dataset. The results demonstrate an acceptable localization performance from the SVM with a limited number of classes (two source combinations). DNN performed well in localizing the both sources. It is clearly noticed in the elevation -45˚ localization, the SVM appeared a good localization performance in estimating only one source, but it is degraded in recognising the second source. The SVM has a similar performance in estimating both sources for the rest of the elevation angles. However, this is a reasonable matter due to the ITD ambiguity of signals that are emitted from the downward

directions affected by the acoustic shadow of body, torso and shoulder. Otherwise, using DNN solves this issue as shown the table 5-4 where DNN has good performance at all elevation levels.

Table 5-4: The azimuth estimation Accuracy in each individual elevation level from SVM and DNN with IRCAM HRTF data set.

Elevation Angle	Source one estimation accuracy ±15˚ (SVM)	Source two estimation accuracy ±15˚ (SVM)	Source one estimation accuracy ±15˚ (DNN)	Source two estimation accuracy ±15˚ (DNN)
-45	**0.776**	**0.539**	**0.926**	**0.917**
-30	**0.590**	**0.581**	**0.924**	**0.919**
-15	**0.589**	**0.591**	**0.917**	**0.913**
0	**0.596**	**0.573**	**0.921**	**0.918**
15	**0.590**	**0.592**	**0.919**	**0.915**
30	**0.593**	**0.591**	**0.927**	**0.912**
45	**0.582**	**0.570**	**0.925**	**0.913**

The figures 5.26, 5.27, 5.28 and 5.29 illustrates the SVM performance at levels 0˚, -15˚, -30˚ and -45˚. These figures show the frequency distribution of angle error that results from predicting source one and source two using SVM with IRCAM HRTF date set. The range of angle errors is from -165˚ to 180˚ with 15˚ increments. These results are from applying SVM model to predict 576 classes at each elevation using validation speakers. The total size of validation data is visualized in the following figures is 5440 samples. In each plot, the x-axis represents the angle error degrees that resulted from compute the signed differences between the actual angles and predicted ones. The y- axis refers to the frequency angle error for each angle. Generally, the SVM kernel was linear so it makes sense performance was poor as the task is a non-linear problem. The confusion matrix plots for source one and source two prediction performance resulted from applying DNN with data generated at individual elevation are shown in appendix II.

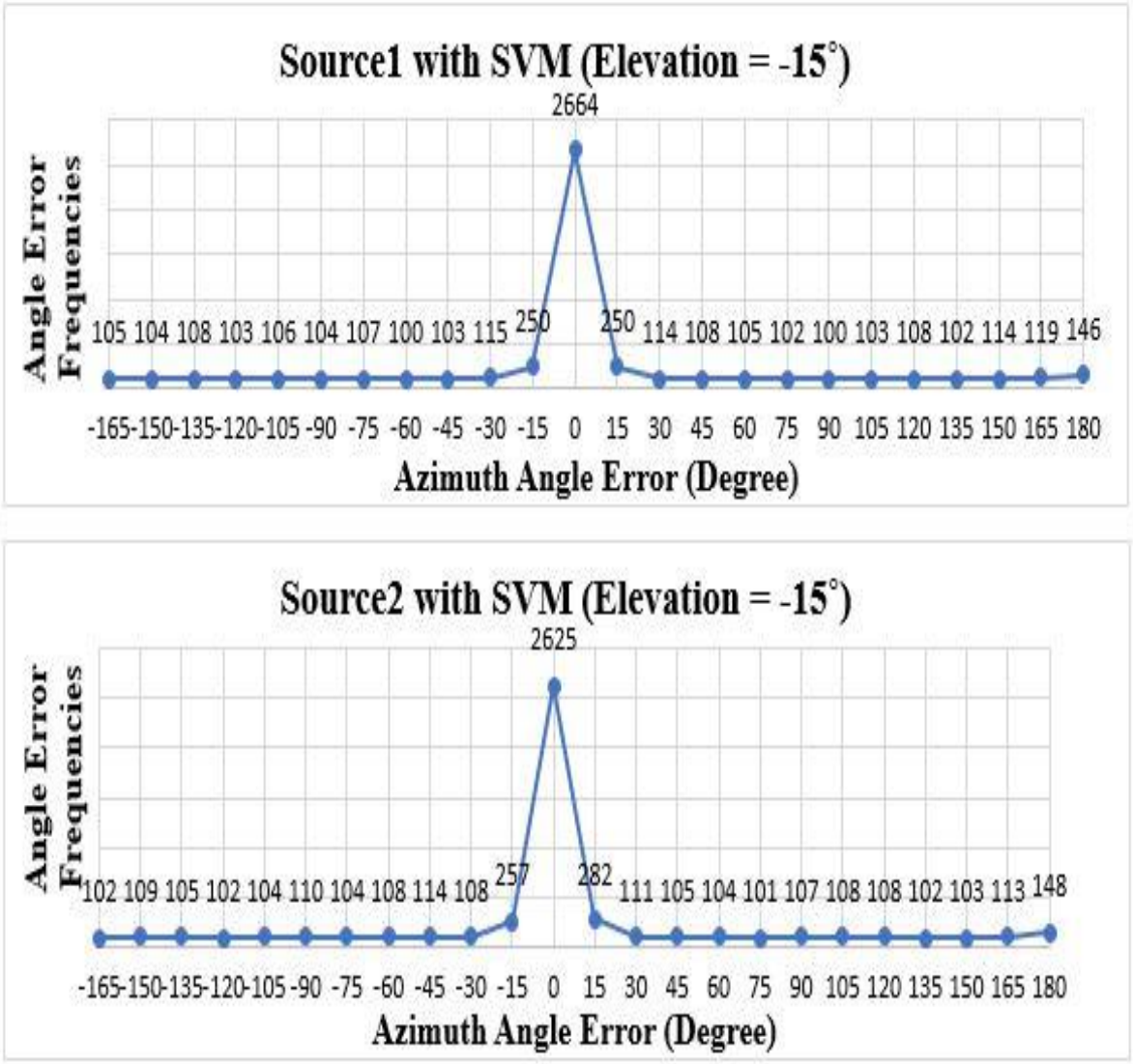


Figure 5.26: Bell shape showing the angle error frequencies for source one and source two predicted by SNN with KEMAR HRTF.

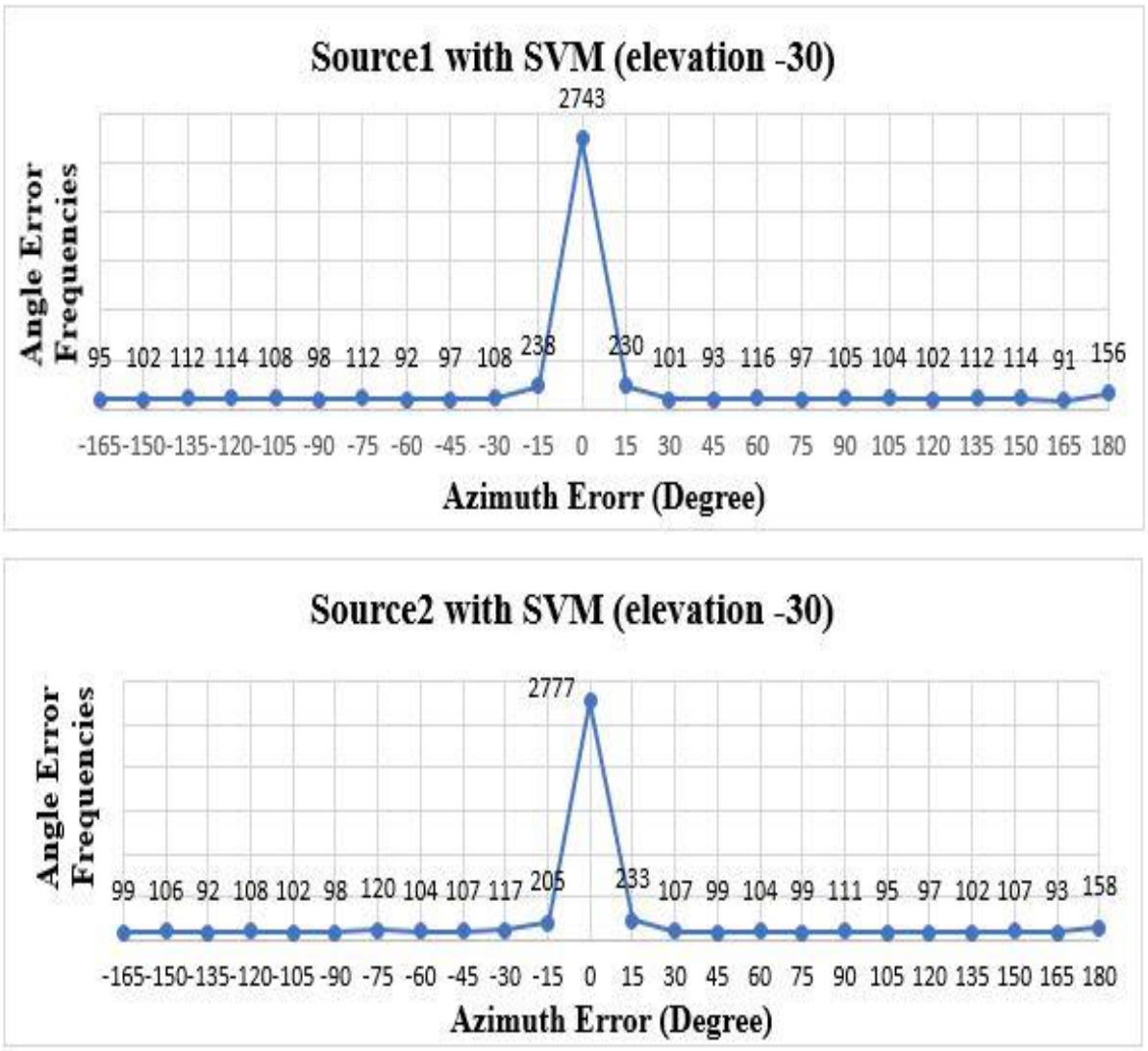


Figure 5.27: Bell shape showing the angle error frequencies for source one and source two predicted by SNN with KEMAR HRTF.

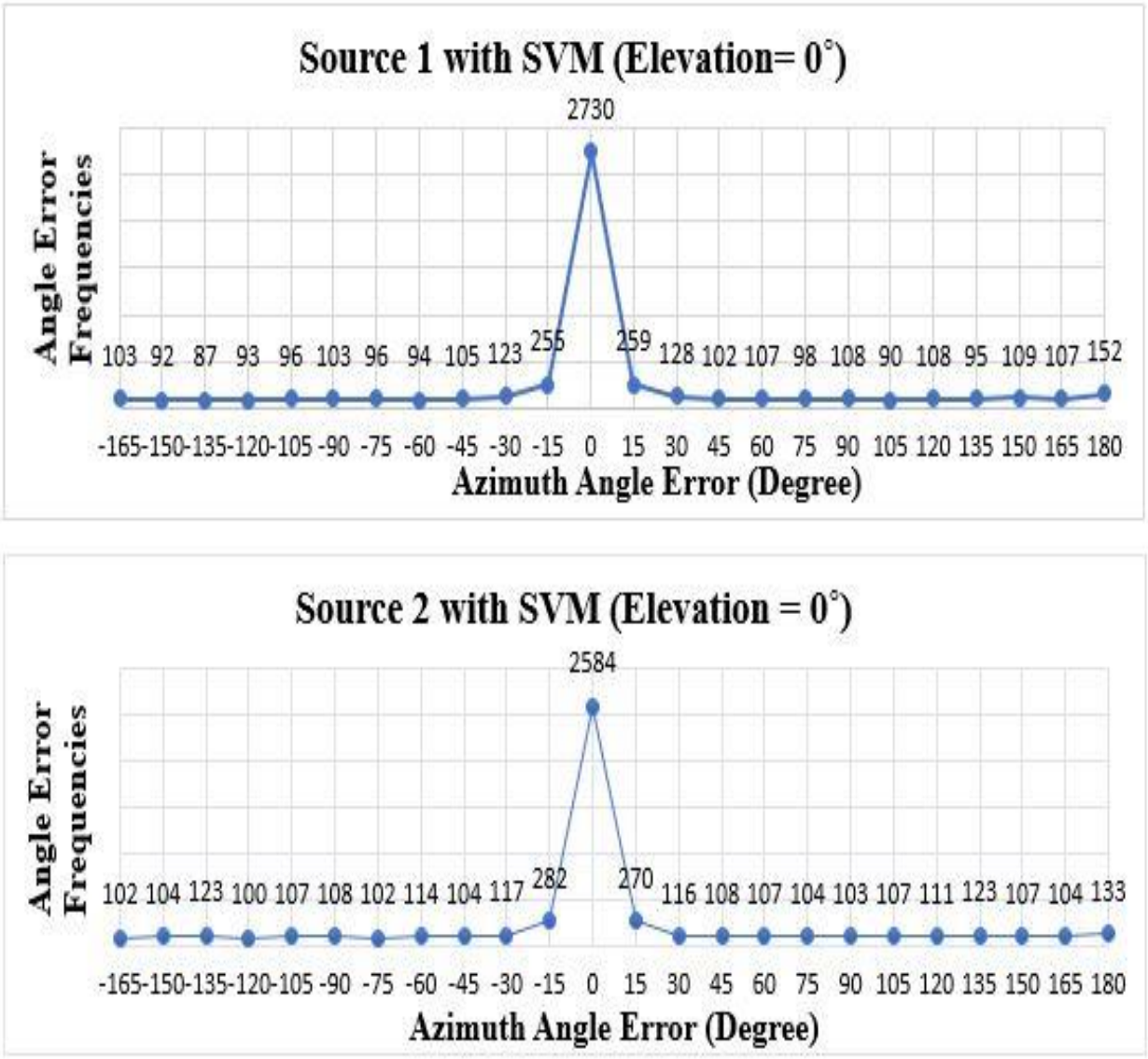


Figure 5.28: Bell shape showing the angle error frequencies for source one and source two predicted by SNN with KEMAR HRTF.

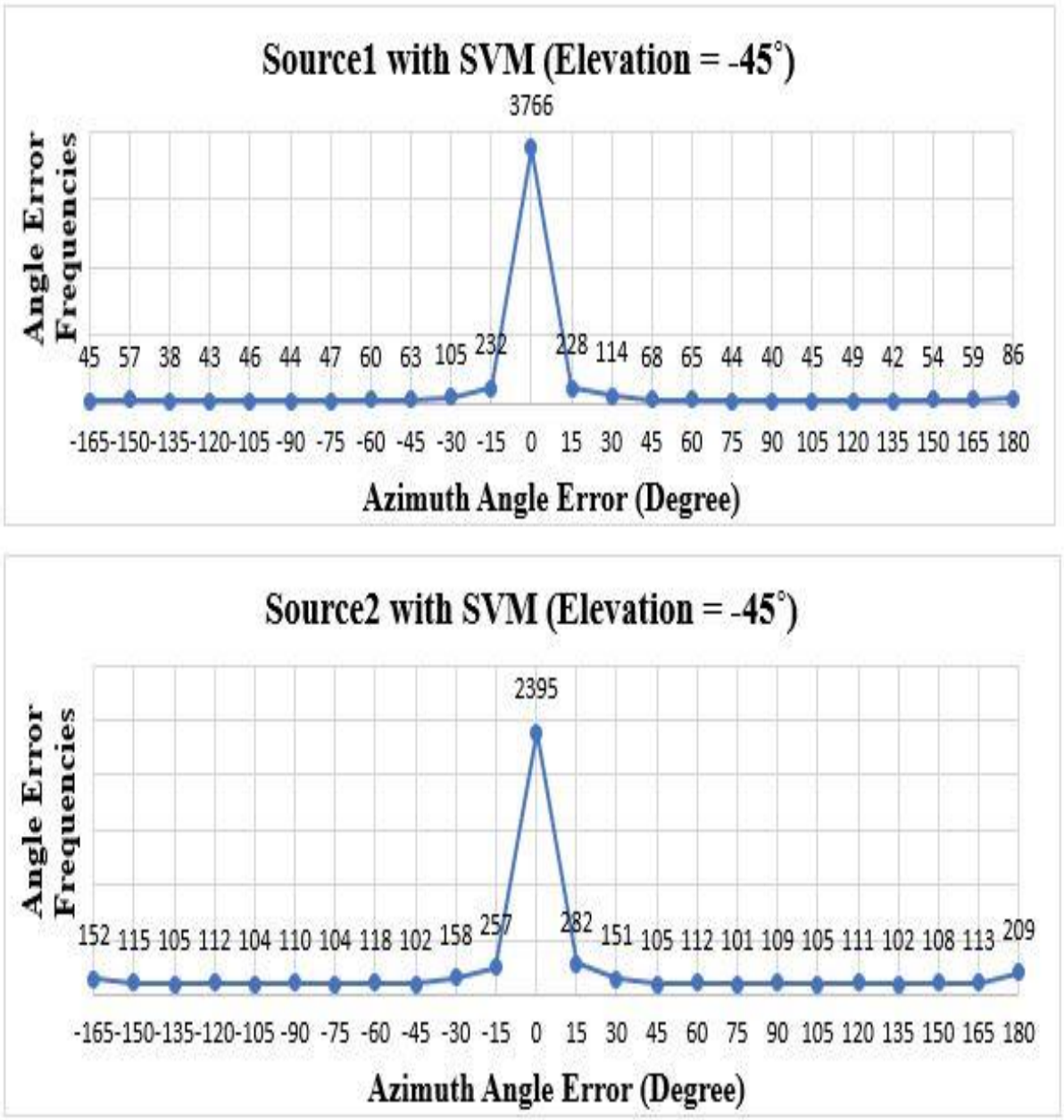


Figure 5.29: Bell shape showing the angle error frequencies for source one and source two predicted by SNN with KEMAR HRTF.

5.7 Comparison between machine learning methods and SNN for the multisource localization.

In this section, the estimation accuracy of localization models by applying SNN and DNN with each type of HRTF data is demonstrated in table 5-5. The final localization judgment accuracies are analysed and computed with angle error ±15˚ also when the back-front (FB) confusion error

is discarded during data analysis. The front back confusion can be resolved by tilting the head through the hearing process. Furthermore, the localization accuracy with ±5˚ and with ±10˚, which is special case when applying KEMAR dummy head data set, are described in the following table for both localization methods. The results in tables 5-5 and 5-6 represent the full data that generated using full locations in elevation range from -45˚ to 45˚ in IRCAM data set while represent the KEMAR data set in elevations -10˚, 0˚, 10˚.

Table 5-5: Comparison between DNN and SNN for multisource localization with KEMAR and IRCAM HRTF data sets.

Sound Sources Localization Method	Localization Accuracy (+/-15˚)	Localization Accuracy (+/-15˚) (without FB Confusion error)	Localization Accuracy (+/-5˚)	Localization Accuracy (+/-10˚)
DNN with IRCAM source1	0.885	0.956	___	___
DNN with IRCAM source2	0.864	0.962	___	___
DNN with KEMAR source1	0.891	0.972	0.857	0.894
DNN with KEMAR source2	0.8515	0.870	0.763	0.800
SNN with IRCAM source1	0.441	0.516	___	___
SNN with IRCAM source2	0.356	0.449	__	___
SNN with KEMAR source1	0.472	0.596	0.401	0.442
SNN with KEMAR source2	0.32	0.408	0.221	0.285

Table 5-6 shows separates the data into azimuth estimation accuracy ±15˚ and elevation estimation accuracy ±15˚ for IRCAM and ±10˚ for KEMAR. The accuracies in the table 5-5 represent the general localization accuracy (average of azimuth and elevation prediction accuracy to represent the locations estimation accuracy). While the table 5-6 demonstrate the actual prediction accuracy of individual azimuth and individual elevation separately to show the localization model performance at each plane (horizontal plane and vertical plane).

Table 5-6: The azimuth and elevation estimation accuracy from DNN and SNN for multisource localization with KEMAR and IRCAM HRTF data sets.

Sound Sources Localization Method	Azimuth estimation accuracy (+/- 15˚)-IRCAM	Elevation Angle estimation accuracy (+/- 15˚)-IRCAM	Azimuth estimation accuracy (+/- 15˚)-KEMAR	Elevation angle estimation accuracy (+/- 10˚)-KEMAR
Source 1 with DNN	0.838	0.932	0.817	0.986
Source 2 with DNN	0.796	0.932	0.719	0.984
Source 1 with SNN	0.441	0.516	0.479	0.492
Source 2 with SNN	0.356	0.449	0.325	0.381

To compare the SVM performance with multisource localization model based on DNN and SNN, SVM is extended to predict the sources from data that was generated from three elevation levels (-15˚, 0˚,15˚). There was an attempt to train the SVM using the full range of data that have been used to train and validate the DNN, but the SVM fails in processing this massive size of training data due to the memory requirements. However, in this experiment, the SVM classifier was trained to predict 1728 classes generated from all possible location combinations at three elevation levels for hundred and thirty-six possible combinations of 17 speakers. Table 5-7 demonstrates the localization performance of multisource localization by using three different machine learning algorithms (DNN, SVM, and SNN). The results are from validation stage when a fresh data presented for these models. Validation data consists of all possible locations' combinations of in the elevation angles (-15˚, 0˚, 15˚) of IRCAM with three possible combinations of three speakers.

Table 5-7: Azimuth and elevation angles estimation accuracy by three localization models (DNN, SVM and SNN).

Sound Sources Localization Methods	Azimuth Angle estimation accuracy (+/- 15˚)-IRCAM	Elevation Angle estimation accuracy (+/- 15˚)-IRCAM
Source 1 with DNN	0.918	0.932
Source 2 with DNN	0.892	0.932
Source 1 with SVM	0.594	0.873
Source 2 with SVM	0.567	0.818
Source 1 with SNN	0.441	0.516
Source 2 with SNN	0.356	0.449

However, the spiking neural based localization model output firing rates was processed using various machine learning methods including DNN and SVM. This novel idea has been tested and the results with different machine learning algorithms have been tested foe single source localization as displayed in the following sections. The experimental findings established that the machine learning concepts able to solve the multisource localization problem when there is an appropriate data. Obviously, the best localization performance is for DNN compared with other machine learning approaches (SVM, SNN). As shown in the table of results. The DNN can learn important patterns in the data to enable successful localisation performance. In addition, the non-linearly separable data, needs non-linear learner. So that the SVM with linear kernel shows a poor localization performance.

5.8 Multisource source localization model with multi-conditions noise

In previous sections, different localization methods were examined to determine the sound sources emitted from different locations and different speakers simultaneously. All prior experiments were done using clean data that simulate ideal environments. In this section, the multisource localization model performance is investigated in noisy conditions. One of the most challenging in the field of sound source localization in general and, more specifically, in binaural hearing, is the noisy environment. This part of our work considers the problem of multisource localization in real-world like conditions when the input speech signals are corrupted by unknown noise levels. Three series of tests with various amounts of background noise were applied to examine how the multisource localization model performs in a number of

noisy situations. The background in these three experiments consisted of no added noise, added background white noise and added directional noise.

Experiment 1: No added noise (clean environment)

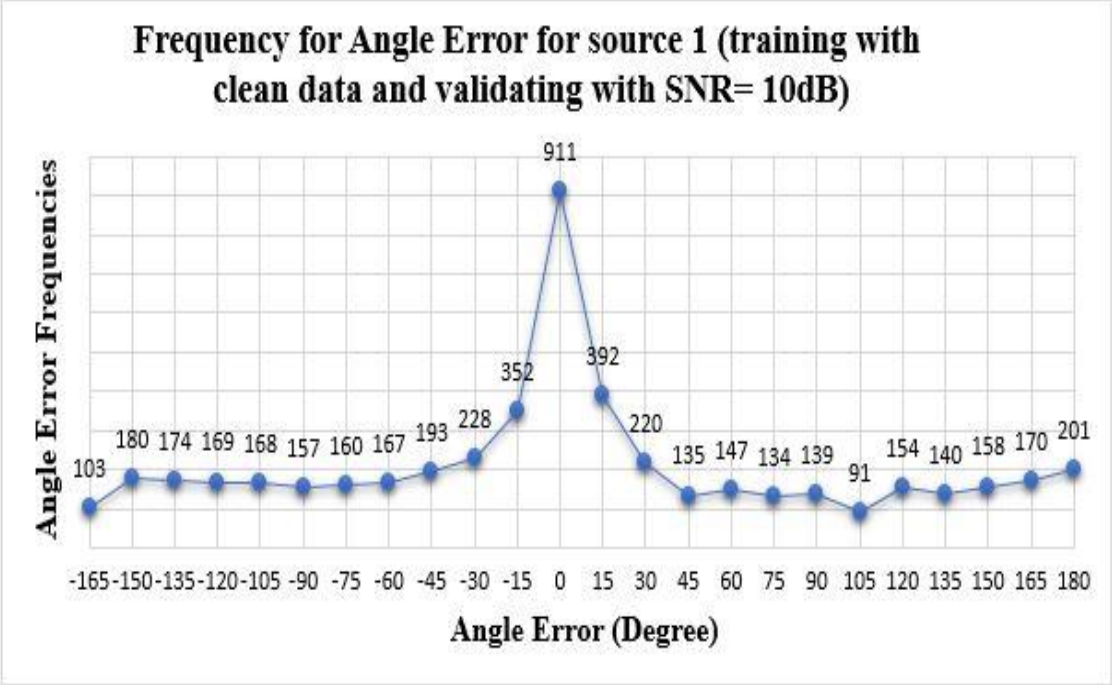
In this experiment, the multisource localization model has been trained by using the full range of training data set that generated from IRCAM and 17 speakers with no added noise. This data has been generated using clean speech samples that were collected in anechoic environment. In the testing stage, the multisource localization model was tested by adding noise with signal-to noise-ratios varying between 10dB, 0dB, -10dB. The testing data was generated by using various speech samples of 500ms belonging to three speakers with different locations in elevation range (-15˚, 0˚, 15˚) adding white noise of 500ms to the mixed two locations signal embedded in two speech signals that of 500ms. The model was trained with clean data and validated with controlled SNRs to investigate the impact of noisy environments on localization model performance. Table 5-8 shows the localization accuracy for estimating source one and source two in each SNR. The results demonstrate a high reduction in localization performance of multisource localization model due to the model has no previous knowledge about these levels of noise through training stage.

Table 5-8: Training the multisource localization model with clean data and validating the model with noisy data over various SNRs separately.

SNR dB	Source one estimation accuracy ±15˚	Source two estimation accuracy ±15˚
10	**0.39**	**0.30**
0	**0.23**	**0.20**
-10	**0.20**	**0.17**

Figures 5.30, 5.31 and 5.32 show the resolution accuracy hat results from predicting source one and source two using multisource localization model based DNN. This experiment was done using data generated from mixing locations of three elevation levels (-15˚, 0˚, 15˚) of IRCAM date set with validation speakers. In each plot, the x-axis represents the angle error degrees that resulted from compute the signed differences between the actual angles and

predicted ones. The y- axis refers to the frequency angle error for each angle. These plots represent 5441 output points resulted from model validation stage.



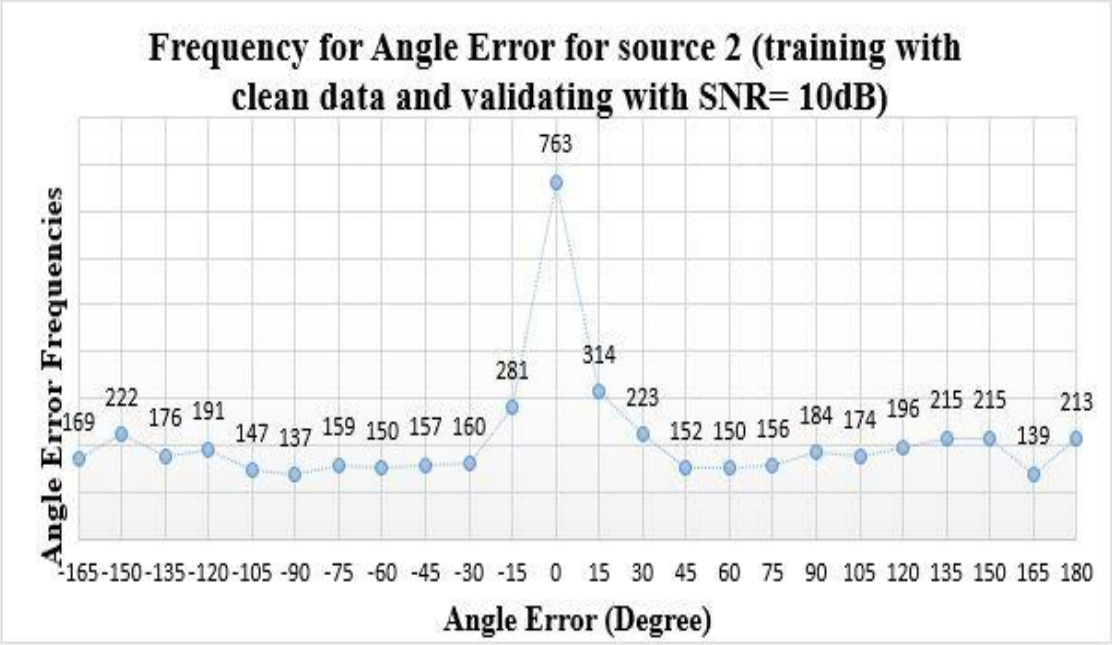
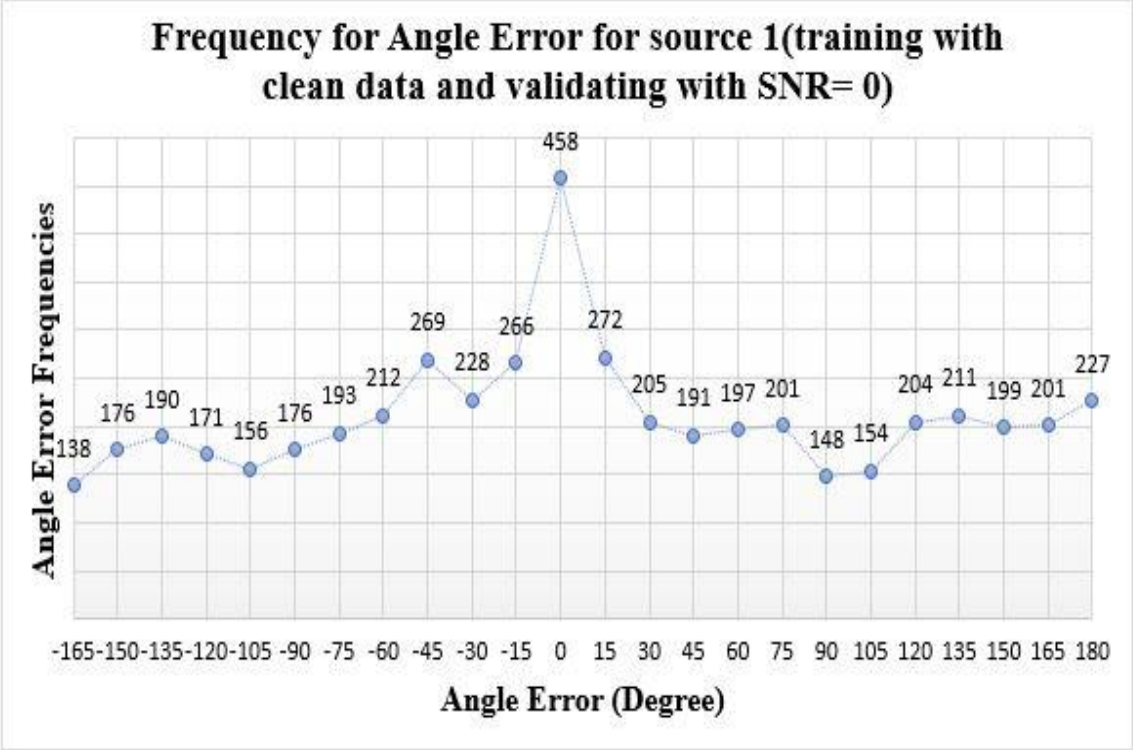


Figure 5.30: Angle error frequencies for source one and two predicted by DNN trained with clean data and validated in noisy condition with SNR = 10dB.



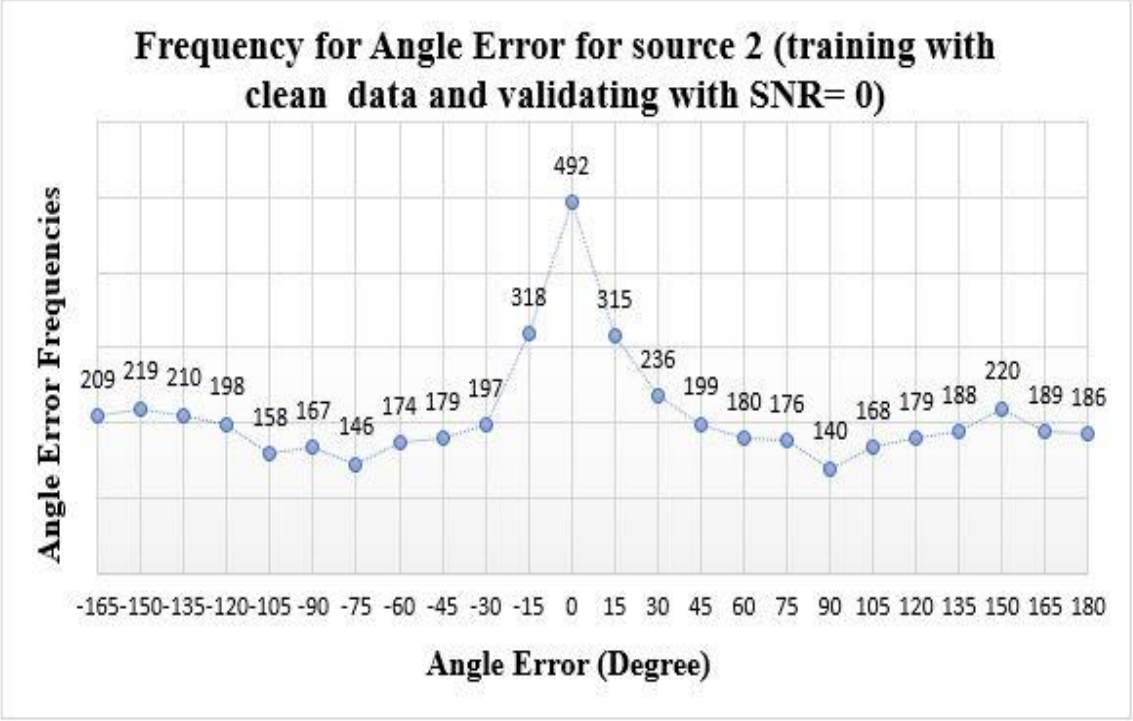
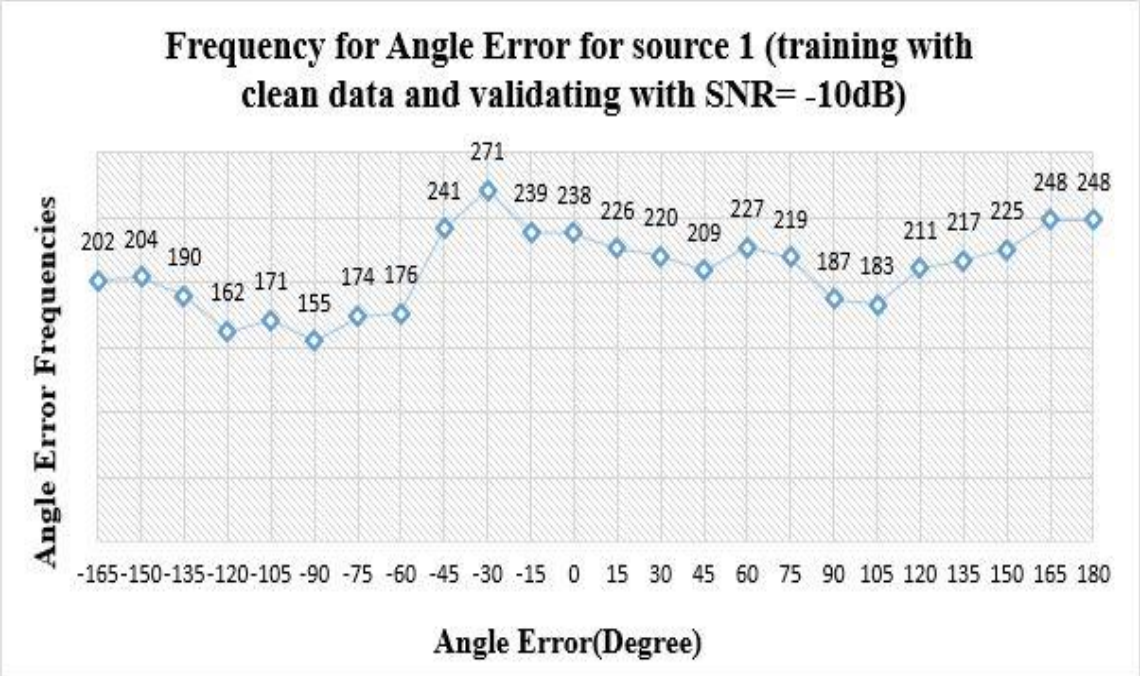


Figure 5.31: Angle error frequencies for source one and two predicted by DNN trained with clean data and validated in noisy condition with SNR = 0dB.



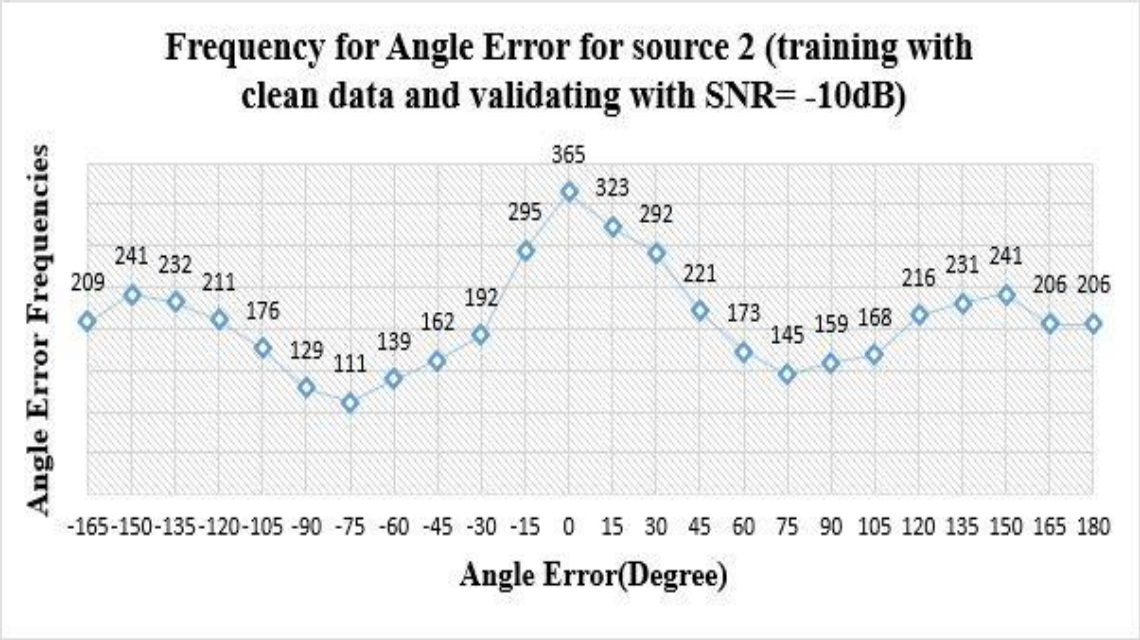


Figure 5.32: Angle error frequencies for source one and two predicted by DNN trained with clean data and validated in noisy condition with SNR = -10dB.

The experimental findings showed the localization performance when the model was trained with clean data and tested with noisy data at different SNRs. It demonstrates the high impact of noisy environments on localization model performance due to there was no previous knowledge about them which caused a highly reducing in the localization performance compared with using a pure testing data. However, figures 5.30, 5.31 and 32 showed the frequency of angle error points of the localization model at each SNR. The localization performance reduced at a low SNR to reach a very poor performance at -10dB of SNR as shown in figure 5.32.

Experiment 2: Added background white noise

To enhance the model localization effectiveness, the model was being trained with noisy data that was generated by adding different levels of white noise of SNRs 10dB, 0db, and -10dB to the ear signals. This simulates diffuse noise due to both ears have different noise. The noisy training data was generated from different speech samples of 17 speakers (training speakers) with all possible combinations between locations from IRCAM data. Firstly, testing the model under single noise condition when SNRs were used individually to train and validate the multisource localization model and the results are reported in table 5-9 (see appendix II).

Table 5-9: Training and validating the multisource localization model on the same noise level separately.

SNR dB	Source one estimation accuracy ±15˚	Source two estimation accuracy ±15˚
10	**0.75**	**0.67**
0	**0.60**	**0.54**
-10	**0.45**	**0.39**

Secondly, testing the model under multiple background noise conditions where the model has been trained with noisy data with various level of background noise and validated with controlled SNRs. The source one and source two estimation accuracy at each noise level are demonstrated in table 5-10.

Table 5-10: Training the multisource localization model with All SNRs and validating the model with noisy data over various SNRs separately.

SNR dB	Source one estimation accuracy ±15˚	Source two estimation accuracy ±15˚
10	**0.71**	**0.65**
0	**0.59**	**0.53**
-10	**0.44**	**0.34**

Figures 5.33, 5.34 and 3.35 show the multisource localization performance to predict source one and source two under different level of background noise at three different SNRs. In each plot, the x-axis represents the angle error degrees that result from computing the signed differences between the actual angle and predicted ones. The y- axis refers to the frequency angle error for each angle. This result is from IRCAM of angles in three elevation range (-15˚, 0˚, 15˚) with validation speakers which resulted 5441 output points.

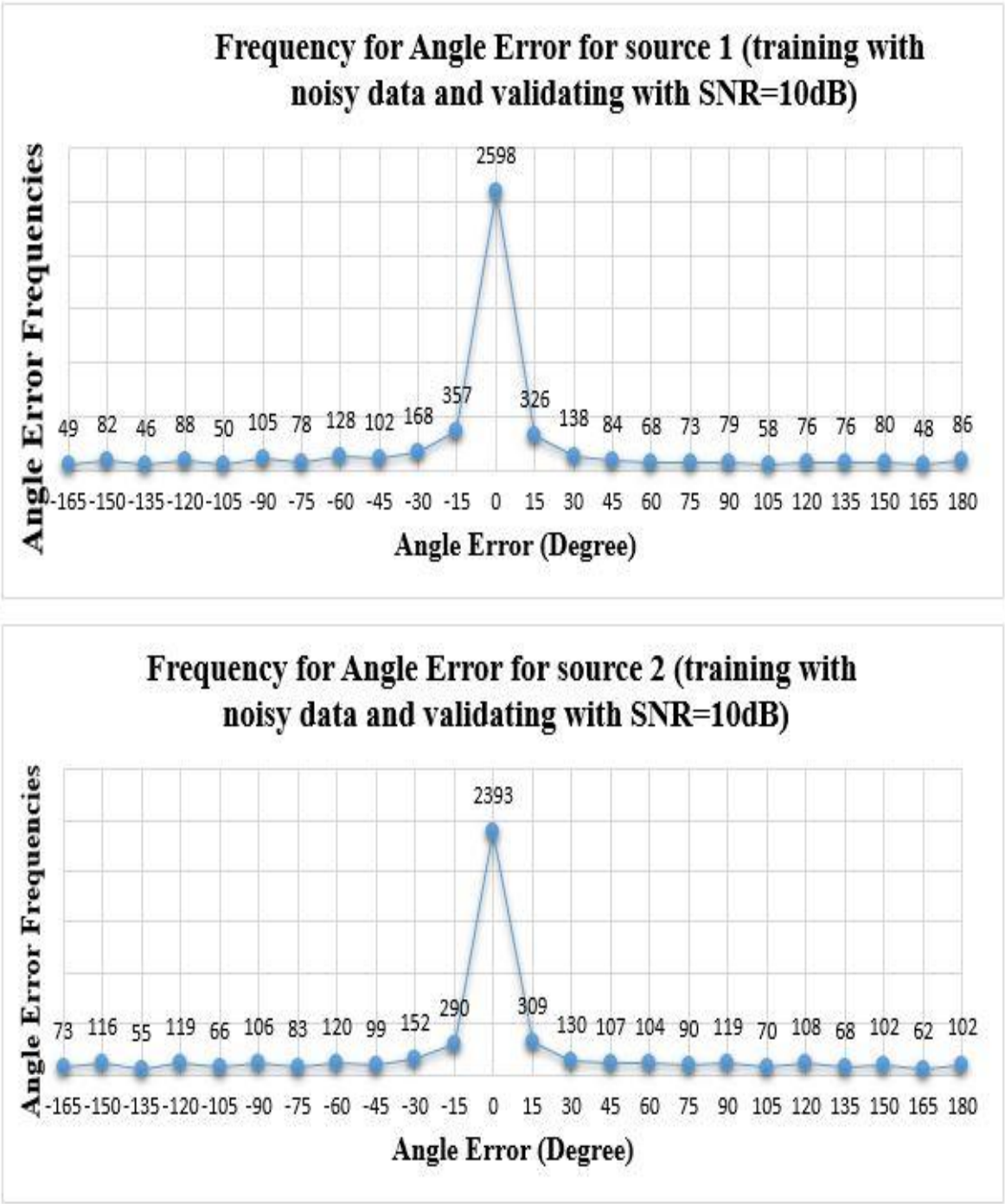


Figure 5.33: Angle error frequencies for source one and two predicted by DNN trained with noisy signal data and validated in noisy condition with SNR = 10dB.

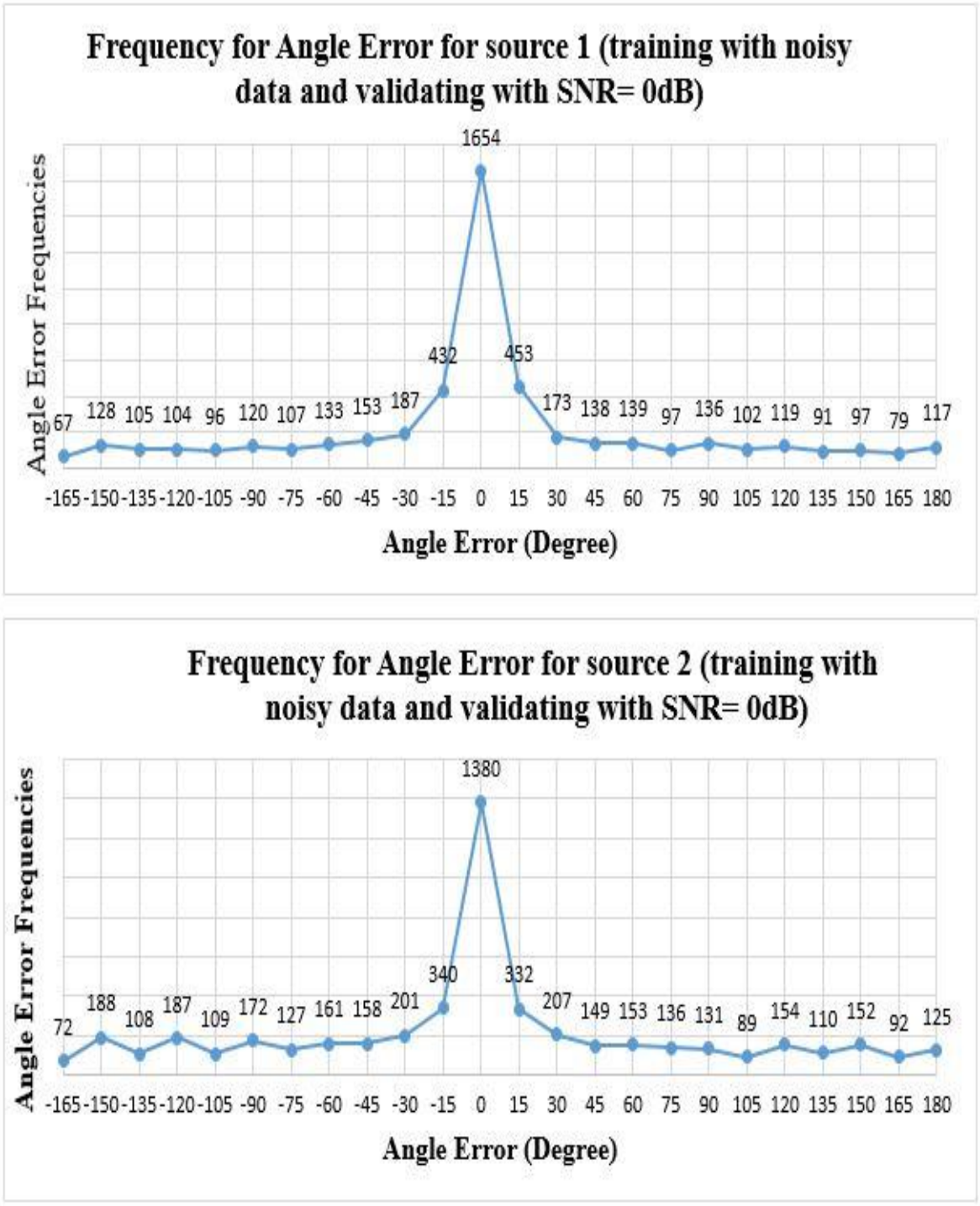
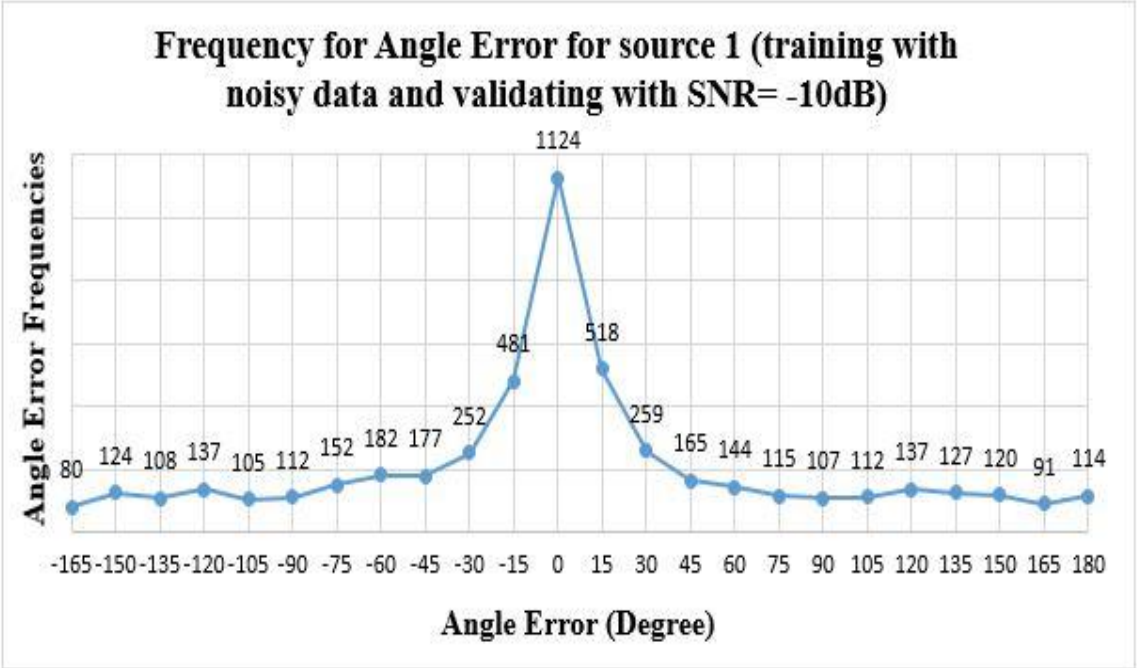


Figure 5.34: Angle error frequencies for source one and two predicted by DNN trained with noisy signal data and validated in noisy condition with SNR = 0dB.



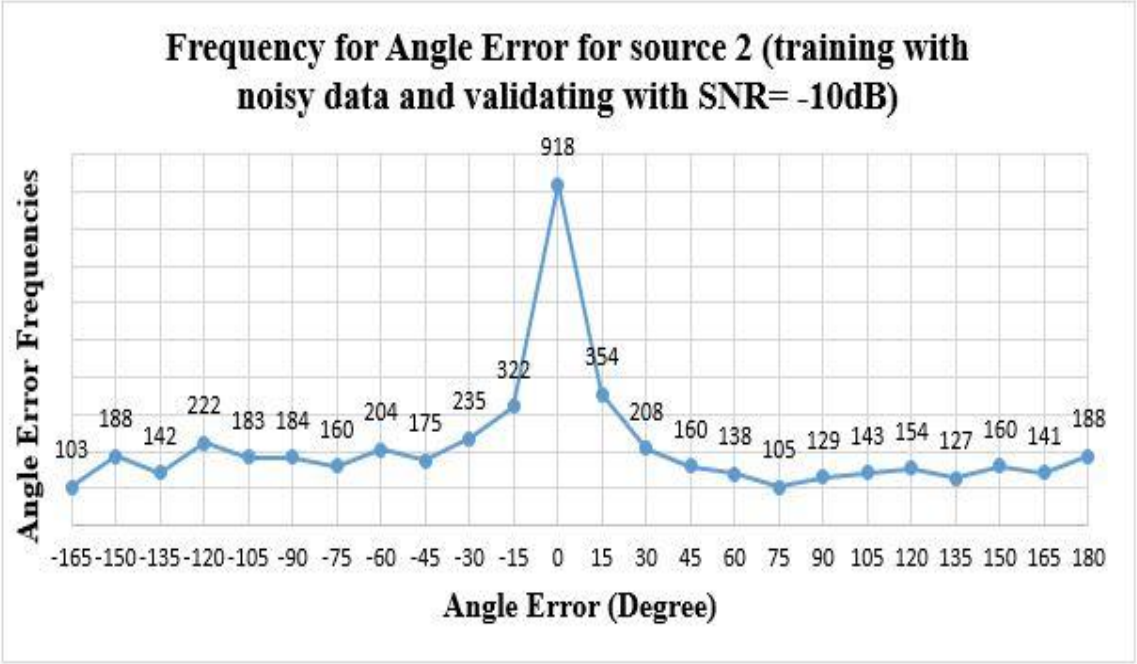


Figure 5.35: Angle error frequencies for source one and two predicted by DNN trained

In this experiment, the model was trained and tested with noisy data at different SNRs. The results explain how the localization performance has been enhanced due to there was a

previous knowledge about these ratios which caused a better localization performance compared with previous experiment when the model has no prior knowledge about noisy data patterns. Figures 5.33, 5.34 and 35 showed the frequency of angle error points of the localization model at each SNR. And, the better localization performance was at 10dB as shown in figure 5.33. This proved that the localization performance to predict the both sources has been improved with a higher SNR.

Experiment 3: Added directional noise

In this test, the white noise signals were added to each channel of HRTFs (left and right) independently before the HRTF convolving stage. In this case, the sound and noise are emitted in the same direction because they are coming from the same sound sources. Directional noise is simulated the electronic and electrical noise that resulted from sound waves transmission devices and equipment. This type of noise is common in the electronics and communication systems, its defined as unwanted disturbance in an electrical signal or an error affects an important information of the communication signals.

The experimental findings demonstrate some good results; sometimes better than the model performance in under the clean conditions. The more reasonable explanation for this state is, with the directional noise, the resolution of signals that are emitted from the same locations is increased which impact positively on the localization model performance In spite of the directional noise is being simulated as coming from the same location as the sound source so it is not a realistic condition, but could it be interesting to consider as it actually improves performance of the localization model.

Table 5-11 explains the experimental results of applying the multisource localization model based on DNN with data that generated from adding a directional noise. Where white noise of 500ms added to the binaural signal of each sound from different sources to generate training and validating data. The model is trained with training data of three noise levels at SNRs of 10dB, 0db, and -10dB and validated with controlled SNRs.

Table 5-11: Training the multisource localization model with directional noise of all SNRs and validating the model with noisy data over various SNRs separately.

SNR dB	Source one estimation accuracy ±15˚	Source two estimation accuracy ±15˚
10	**0.88**	**0.85**
0	**0.88**	**0.86**
-10	**0.92**	**0.91**

Figures 5.36, 5.37 and 5.38 show the multisource localization performance to predict different level of noise at three different SNRs. In each plot, the x-axis represents the angle error degrees that resulted from computing the signed differences between the actual angle and predicted ones. The y- axis refers to the frequency angle error for each angle. This result is from IRCAM of angles in three elevation ranges (-15˚, 0˚, 15˚) with validation speakers which resulted 5441 output points.

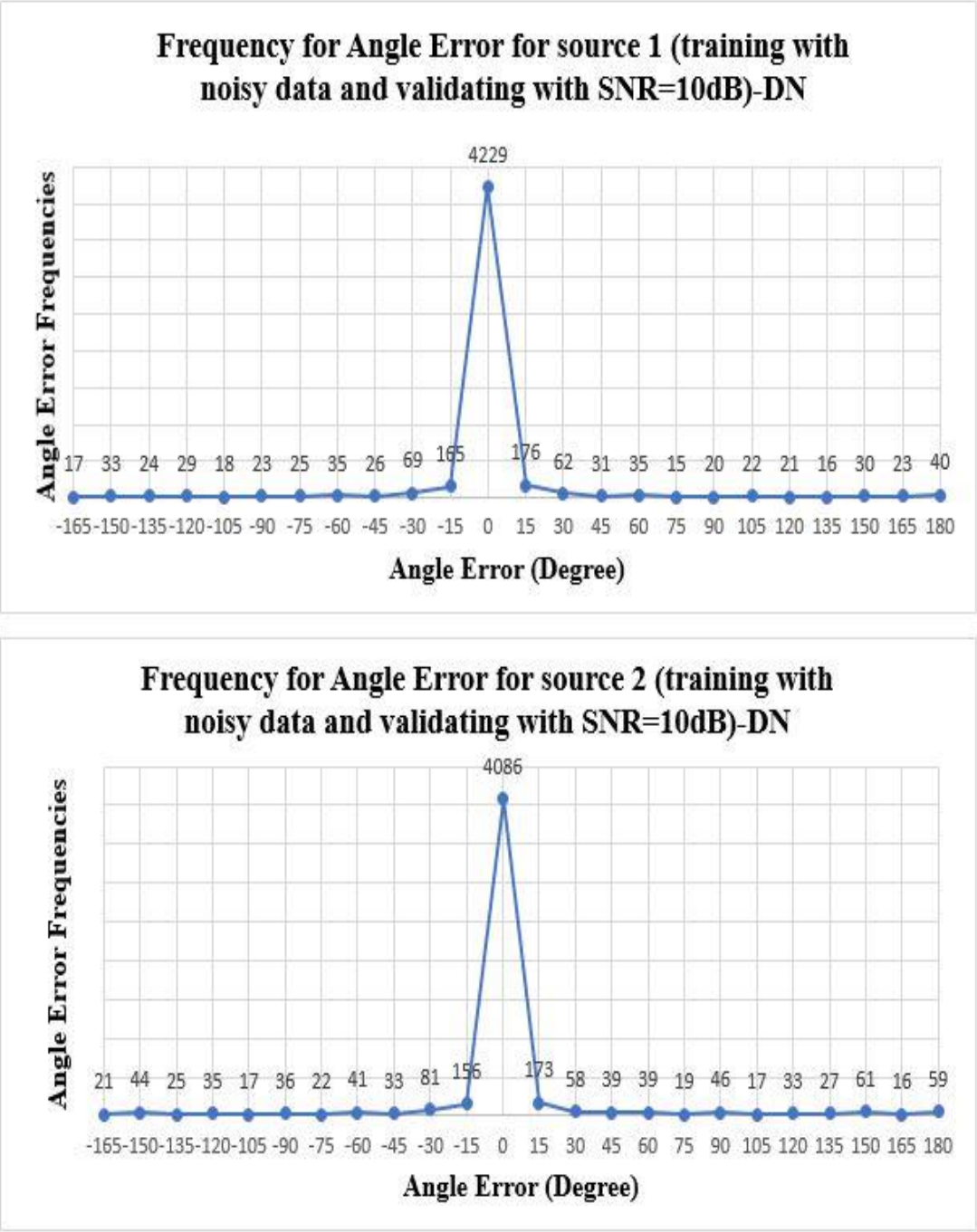


Figure 5.36: Angle error frequencies for source one and two predicted by DNN trained with noisy signals (directional noise) and validated in noisy condition with SNR = 10dB.

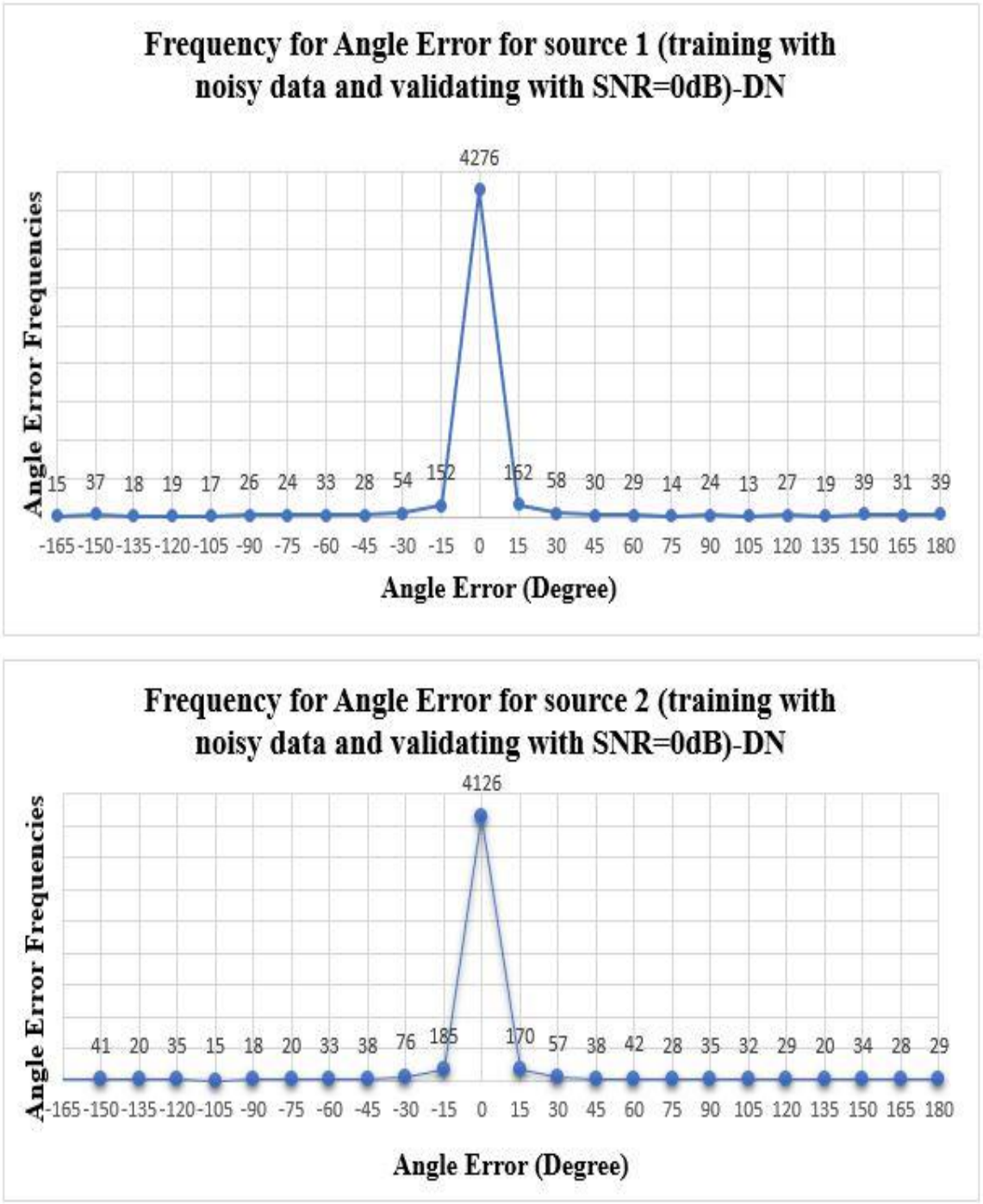


Figure 5.37: Angle error frequencies for source one and two predicted by DNN trained with noisy signals (directional noise) and validated in noisy condition with SNR = 0dB.

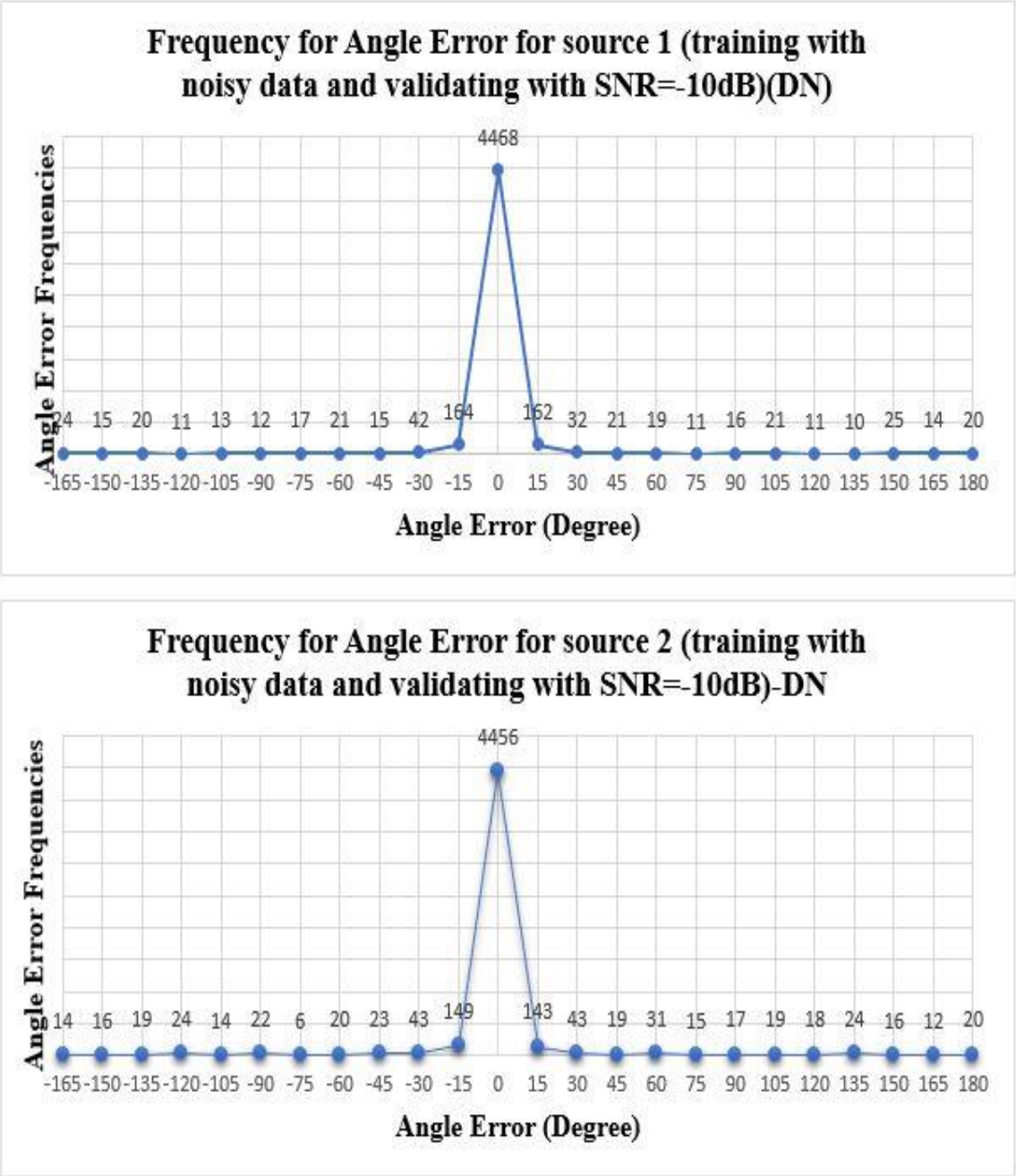


Figure 5.38: Angle error frequencies for source one and two predicted by DNN trained with noisy signals (directional noise) and validated in noisy condition with SNR = -10dB.

5.9 Discussion

In this chapter, a novel idea for multisource sound localization by using only two ears has been presented. The method is inspired by the way humans estimate the location by using the binaural features such as interaural time difference ITD and interaural level difference ILD. HRTFs have been employed to acquire the binaural information. The SNN presents more realistic representation of human hearing by mimicking the binaural time delays in its simulations. The frequency features of input responses were analysed by using set of gamma-tone filter bank. The spiking neural networks (SNN) works as a binaural feature extraction algorithm to extract the timing information from the binaural responses. A DNN is then trained to process the firing rates from numerous coincident spiking neurons to predict the locations of multiple simultaneous sources.

The localization process has two steps: The first step is to predict the number of sources in the incoming signal by analysing the SNN firing rates. Once the number of sources is known the appropriate localization model (single or multisource localisation) can be selected in order detect the source directions. Logistic regression was applied to create a best fit logistic curve to separate between the two sources and one source signals. The model showed a better performance in predicting the number of sources from different speech signals and even under noisy conditions compared to the localization model based SNN.

In a second step the SNN firing rate features were used to train a DNN to perform a classification task. In this case, the DNN learned from examples, where each example is associated with two predefined labels (the location of source one and source two).

The localization model is first tested in a task to localise single sound sources emitted from a unique location. Different speech samples belonging 100 speakers contributed to train and test the single sound source localization model. The localization model was then extended to two simultaneous sources generated from all possible combination for 20 speakers (17 speakers for training and 3 speakers for validation).

Two types of machine learning methods were applied to process the spiking neural networks firing rate features for multisource sound localization. Firstly, the deep neural network was examined for multisource localization which returned a high accuracy of 91% and 89% for

source one and two. Moreover, the angle errors between the actual and predicted locations have been analysed. Two types of angle errors have been determined; front-back confusion and left-right angle error. These forms of angle errors have relative impact on the source one and source two localization judgments, comparatively modest error on the range from ±5˚ to ±15˚ and the characteristic form of errors recognised as back-front confusions. There are no significant left-right error probabilities observed in the multisource localization model experiments. Whereas the source prediction accuracy of the multisource localization model was frequently affected by a front-back confusion error type. In this case it is important to mention and take in an account that these experiments used a static head which brings more complexity to deal with sound signals that are issued from the back.

The experiment results demonstrate that the localization accuracy enhancement highly depended on increasing the number of training samples that were used to train the deep neural network. The experimental outcomes demonstrate that the localization performance of multisource localization model have been improved by increasing the number of speakers that contribute in generating the deep neural network training data sets. And, this is reasonable due to the increase teaching examples of the machine learning models. To test this practically, first the model was trained with data that was generated by using only 10 speakers where only 45 possible combinations between these speakers participated in constructing the training data. The position estimation accuracy for both locations with ±15 angle degree did not exceed 55%. To improve the multisource sound localization performance, the number of speakers is raised to be 17, producing 136 possible combinations between participated speakers. Thus, the multisource estimation has been boosted by achieving localization accuracy in ±15˚ reach to 90% and 89% for source one and source two respectively as shown in table 5.7. When the number of speakers is raised to be 17 which caused an increasing in the teaching examples for each location combination resulted a significant enhancing in the multisource localization performance.

For evaluation and comparison purposes, the DNN localization performance was compared with other machine learning methods. Multisource localization models based on SVM and SNN were investigated to study their performance in localizing multi sound sources. The results from these two methods were analysed and compared with the DNN localization performance. SNN with a ‘two-winner-takes-all’ concept was implemented to detect the two

locations. SNN based multisource localization model showed poor performance in estimating source one with slightly better performance in predicting source two. SVM classifier performed better than SNN but still its performance limited and less than DNN in predicting both sources.

The spiking neural based localization model output firing rates was processed using various machine learning methods including DNN and SVM. This novel idea has been tested and the outcomes with different machine learning algorithms have been demonstrated. The DNN showed a better localization performance compared with SVM and SNN. The DNN can learn important patterns in the data to enable successful localisation. Also, the non-linearly separable data, needs non-linear learner for the best performance. So that, the SVM with linear kernel showed a poor localization performance.

Moreover, the multi-condition noisy environments impact on the multisource localization model performance have been examined in three experiments. Firstly, the impact of background noise has been investigated when the localization model was trained with clean data and tested with noisy data at different SNRs. Secondly, the multisource localization model was trained with multi-condition background noise at SNRs of 10dB, 0dB, and -10dB and tested at controlled SNR. The findings demonstrate an enhancement in the model performance in predicting source one and source two when the model trained using noisy data. The final experiment examined the impact of the directional noise on the multisource localization model performance.

It is necessary to calculate the signal-to-noise-ratio (SNR) in order to define the strength of a signal. It is easy to extract the useful information or detect a true signal from the raw signal at the higher SNRs due to the power of a signal is higher than the power of the background noise. Experimentally, the localization model has been tested with poor sound signals at low SNRs at 10dB, 0dB and -10dB. While, the better human hearing is 30 dB and above. However, the findings have been demonstrating an enhancing in the localization performance by increasing the signal to noise ratio. The knowledge of this ratio has many important applications that related with enhance the hearing experience. For example, people who use the hearing aids.

Finally, most of the chapter experiments have been done using two types of HRTF databases; IRCAM and KEMAR dummy head. Each one of these data has special impact on the multi-source localization model performance due to the differences in the anatomical

parameters (head size, ear shape and torso). Also, using two different HRTFs to test the multisource advocates the model generalisation.

5.10 Summary

1. The spiking neural based localization model output firing rates was processed using various machine learning methods including DNN and SVM. This novel idea has been tested and the outcomes with different machine learning algorithms have been demonstrated. The DNN showed a better localization performance compared with SVM and SNN. The DNN can learn important patterns in the data to enable successful localisation. Also, the non-linearly separable data, needs non-linear learner for the best performance. So that, the SVM with linear kernel showed a poor localization performance.
2. Moreover, the multi-condition noisy environments impact on the multisource localization model performance have been examined in three experiments. Firstly, the impact of background noise has been investigated when the localization model was trained with clean data and tested with noisy data at different SNRs. Secondly, the multisource localization model was trained with multi-condition background noise at SNRs of 10dB, 0dB, and -10dB and tested at controlled SNR. The findings demonstrate an enhancement in the model performance in predicting source one and source two when the model trained using noisy data. The final experiment examined the impact of the directional noise on the multisource localization model performance.
3. It is necessary to calculate the signal-to-noise-ratio (SNR) in order to define the strength of a signal. It is easy to extract the useful information or detect a true signal from the raw signal at the higher SNRs due to the power of a signal is higher than the power of the background noise. Experimentally, the localization model has been tested with poor sound signals at low SNRs at 10dB, 0dB and -10dB. While, the better human hearing is 30 dB and above. However, the findings have been demonstrating an enhancing in the localization performance by increasing the signal to noise ratio. The knowledge of this ratio has many important applications that related with enhance the hearing experience. For example, people who use the hearing aids.

4. Finally, most of the chapter experiments have been done using two types of HRTF databases; IRCAM and KEMAR dummy head. Each one of these data has special impact on the multi-source localization model performance due to the differences in the anatomical parameters (head size, ear shape and torso). Also, using two different HRTFs to test the multisource advocates the model generalisation.

CHAPTER 6

LOCALIZATION WITH NON-INDIVIDUALIZED HRTFS

Chapter Overview

All the previous experiments were done by using matched HRTFs to train and test the machine learning models. In this chapter, the localisation models for single andmulti-sources were tested using mismatched HRTFs to investigate the localisation model performance with the non-individual HRTF. In this chapter, the problem of non-individual HRTFs is reviewed in section 6.1. The performance of the single sound source based SNN with mismatched HRTF is shown in section 6.2. Followed by the performance of different machine learning techniques for single sound source localisation model with mismatched HRTFs. The multisource localisation models with non-individualised HRTFs are produced in section 6.3. Furthermore, some suggested solution for generic localisation model is shown in section 6.4.

6.1 The non-individual HRTFs

Recently, there is growing in the importance of the usage of head-related transfer functions (HRTFs) to expect the direction of any sound signal. Moreover, it's become more interested by improving and transferring information to the listeners by estimation the accurate sound signal direction (Cunningham and Streeter 2001). Binaural sound localisation is referred to human's capability to use binaural cues to predict the direction of sound (Cheng' and Wakefield 1999). HRTFs show an outstanding localisation performance if individualised HRTF are used (Mendonca et al. 2014). Individuality refers to the properties of a HRTF which are functions of the unique anatomical parameters of a person (Panna, torso and head). It is no possible to measure evey individual's HRTF because it is a costly and time-consuming process. The non-individual Head-Related Transfer Functions is necessary for the most of binaural applications when it represents an admitted substitution for the individual HRTF to be a generic HRTF for these applications. There is a notable dispute related with non-individualized HRTF ability to make possible results in the Auralization implementations compared to the individual HRTF (Mendonca et al. 2014, Andreopoulou and Katz 2015).

The problem of mismatched HRTFs can be summarised as follows: When listening through HRTFs measured from one's own ears a listener reports auditory events that appear 'externalised', i.e. that seem to arise from sources outside of the listener's head. When listening through HRTFs measured from another subject, i.e., 'non-individualised' HRTFs the listener often complains that auditory events are spatially diffuse, and listeners often make incorrect judgements of the source locations (Wenzel et al. 1993, Moller et al. 1995).

Non-individuality is one of the most significant issues in binaural audio. It isessential to find a generic model able to work with various types of HRTFs which represent different subjects and solve the non-individual HRTF problem.

In this chapter, mismatched HRTFs were used to test single source and multisource localisation. Two HRTF datasets have been used to train different localisation models. Data capture using different, mismatched, HRTFs were applied to test the localisation models. The primary goal of this chapter was to explore problem of mismatched HRTF and to quantify the degradation in localisation performance so that a generic model for use by any listener may be possible.

6.2 The HRTFs dimensionality adjustment

The IRCAM and KEMAR HRTFs data sets have different sizes. Also, they have unequal angles measurements. So it is necessary to modify one of them to match the other from where of the database size and its locations measurements to make the resolutions are comparable .IRCAM has 187 measurements while KEMAR has 710 measurements, the size adjustment process included decimating the KEMAR database so that is has a resolution with only 187 locations. In some cases, exact matches were not possible, so measurements position closest in angle were selected. Table 6-1 shows the IRCAM and KEMAR HRTF data set after the size adjustment process. On the left hand, the table explains the elevation angles for IRCAM and KEMAR. On the right hand, it shows the azimuth angles for IRCAM with adjusted azimuth angles from KEMAR for selected elevation (-40 for KEMAR and -45 for IRCAM). As mentioned in chapter 3, KEMAR has irregular increments at all elevation levels so that each elevation level has almost different azimuth measurements. The process of KEMAR size adjustment has been done manually by keeping the equal or nearest angles in the same index of IRCAM HRTF data set and remove the other unwanted azimuth angles.

Table 6-1: The IRCAM and adjusted KEMAR HRTF datasets.

IRCAM Elevation	Adjusted KEMAR Elevation
-40	-45
-30	-30
-10	-15
0	0
10	15
30	30
40	45
60	60
70	75
90	90

IRCAM Azimuth	Adjusted KEMAR Azimuth
0	0
15	13
30	32
45	45
60	58
75	77
90	90
105	103
120	122
135	135
150	148
165	167
180	180
195	193
205	206
220	219
235	238
250	251
265	264
280	283
295	296
310	309
330	328
345	347

6.3 Evaluate the single source models with mismatched HRTFs

The process starts by updating the spiking neural model for single sound source localisation to work with mismatched HRTFs. In the following experiments, the SNN has been trained with one HRTF and tested with locations measured using another HRTF. The experiments involved two aspects; firstly, the IRCAM HRTF data set was used to train the SNN and KEMAR was used for testing it. In the second aspect, the SNN was trained with a KEMAR dummy head and tested with IRCAM HRTF set. The work involved testing the both HRTFs to investigate which one has a better localisation performance as a generic HRTF.

Experiment 1: SNN was trained with white noise signal convolved with IRCAM and tested with KEMAR

At first, the SNN has been trained using different instances of 500 ms white noise signal convolved with different locations from IRCAM data sets. Then, the model is tested with different instants of white noise that convolved with the KEMAR data set. Figure 6.1 shows the single sound source localisation using the winner takes all SNN firing rate approach performance in estimating azimuth angles from the KEMAR data set when the model trained by IRCAM data.

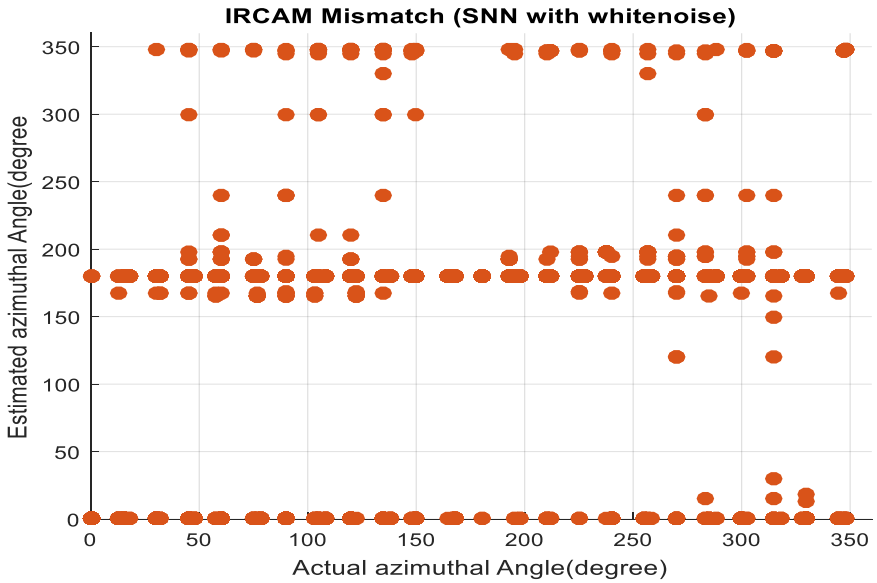


Figure 6.1: SNN performance in estimating azimuth angles with mismatched HRTFs when IRCAM in training and testing with KEMAR.

The x-axis refers to the actual azimuth angles while the y-axis represents the predicted azimuth angles. Figure 6.2 explains the azimuth angle error for the SNN with mismatched HRTFs. The x-axis indicates the actual azimuth angles, and the y-axis indicates the azimuth predicted angle error.

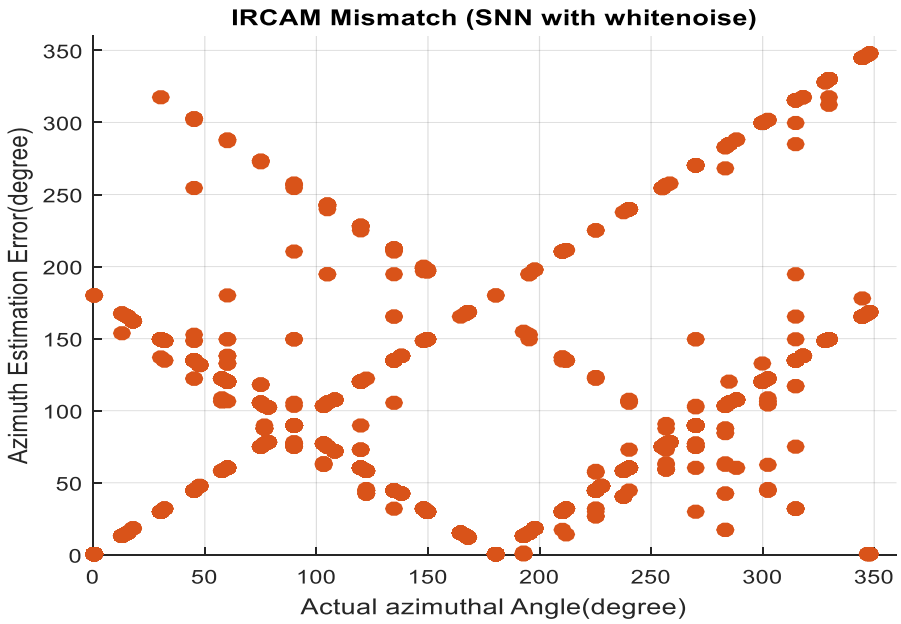


Figure 6.2: Estimation angle error of azimuth angles by applying SNN with mismatched HRTFs when IRCAM in training and testing with KEMAR.

The figures 6.1 and 6.2 are demonstrated the perceptual distortions in predicting the azimuth angles when using non-individual HRTFs. The results showedvery high front/back confusion where the model prediction was flipped entirely between angles 0˚and 180˚.

Figure 6.3 shows the SNN performance in predicting the elevation angles. The x-axis indicates the actual elevation angles, and the y-axis refers to the predicted elevation angle that resulted from applying SNN with mismatched HRTFs. Figure 6.4 demonstrated the elevation angle errors when SNN trained with IRCAM and tested with KEMAR. In this figure, the x-axis refers to actual elevation, and the y-axis indicates to elevation angle errors.

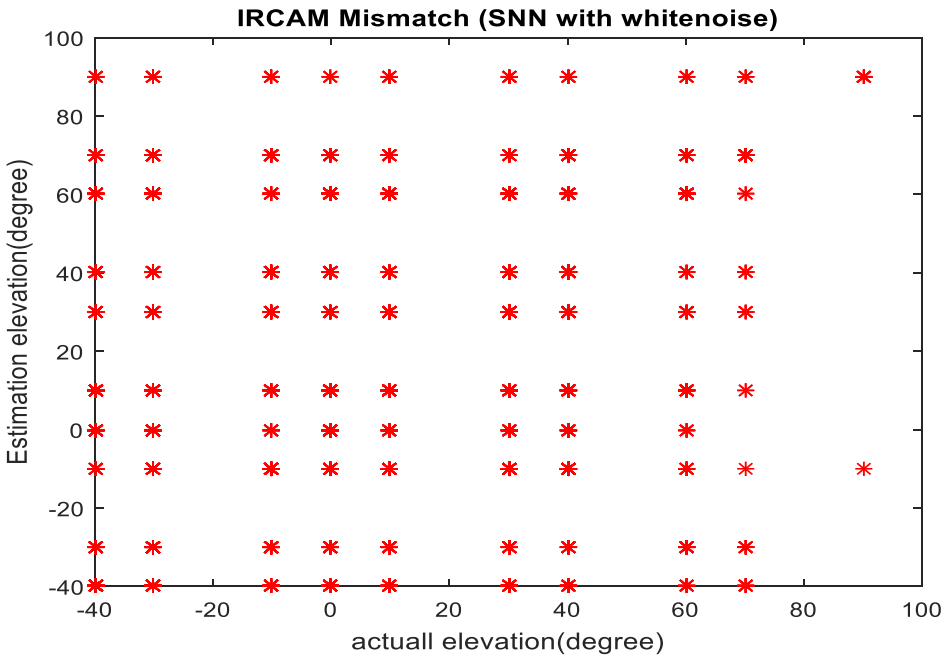


Figure 6.3: SNN performance in estimating elevation angles with mismatched HRTFs when IRCAM in training and testing with KEMAR.

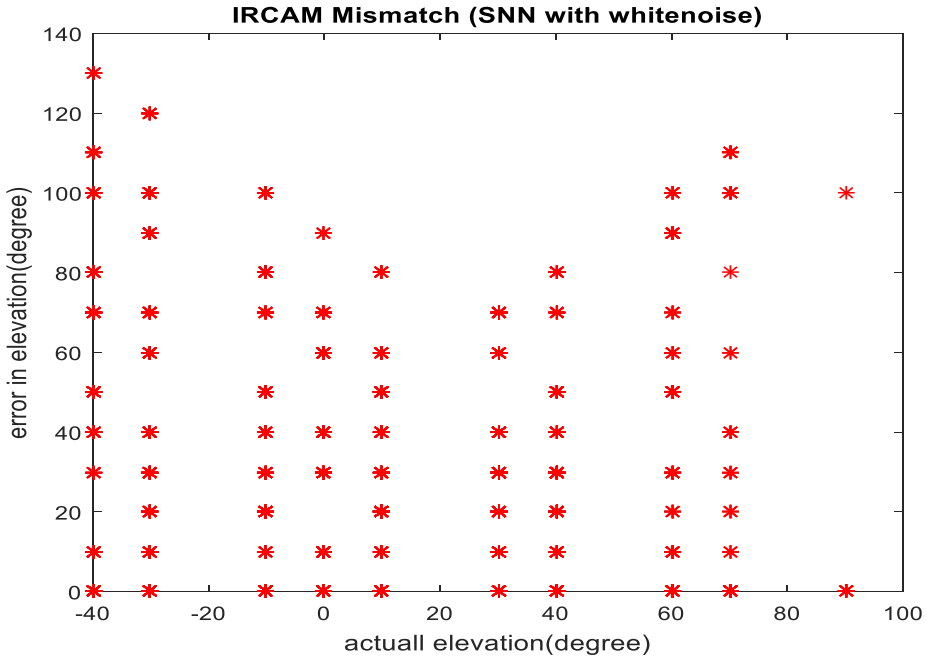


Figure 6.4: Estimation angle error of elevation angles by applying SNN with mismatched HRTFs when IRCAM in training and testing with KEMAR.

The figures demonstrate the angular distortions in the vertical plane when applying SNN with non-individual HRTF. As the experimental results demonstrate, there is an extremely high in the angular error related to estimate both azimuth and elevation angles due to the mismatching between the ITD cues belong both HRTF databases that was used to train and test the SNN. This mismatching caused an increasing in the front-back and up-down ambiguities that led to low localization performance compared with the individual HRTFs which can significantly enhance the sound localization performance. The outcomes present that an unmatched listener's head size is one of the fundamental rises of side image direction distortion in virtual sound reproduction.

Experiment 2: SNN was trained with speech samples convolved with IRCAM and tested with KEMAR.

The single source localisation model was trained by using a variety of the speechsamples to investigate the localisation performance with mismatched HRTFs and speech samples(Al-Noori 2017). The model was tested with speech signals convolved with KEMAR HRTF to test its performance with mismatched HRTFs. Figure 6.5 demonstrates the SNN performance in predicting azimuth angles using non-individual HRTF with speech signal.

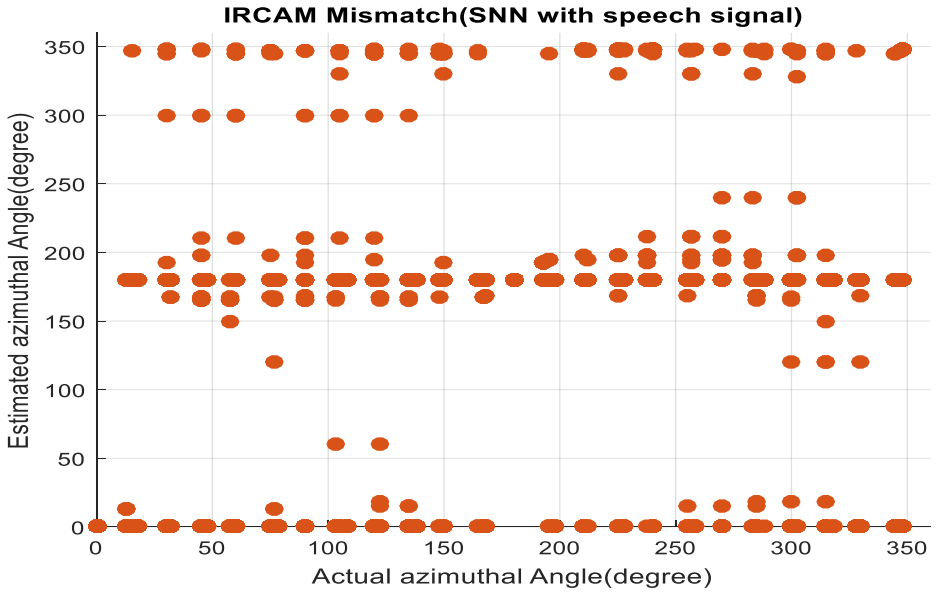


Figure 6.5: SNN performance in estimating with azimuth angle from speech signal convolved with IRCAM in training and testing with KEMAR.

In figure 6.5, the x-axis refers to the actual azimuth angles while the y-axis represents the estimated azimuth angle. Figure 6.6 shows the estimation angle error of azimuth from training SNN with speech sample convolved with IRCAM and tested with different speech samples convolved with KEMAR. The x-axis represents the actual azimuth angles, and the y-axis refers to the angle error of azimuth angle resulted from computing the absolute difference between the actual angle and predicted angle.

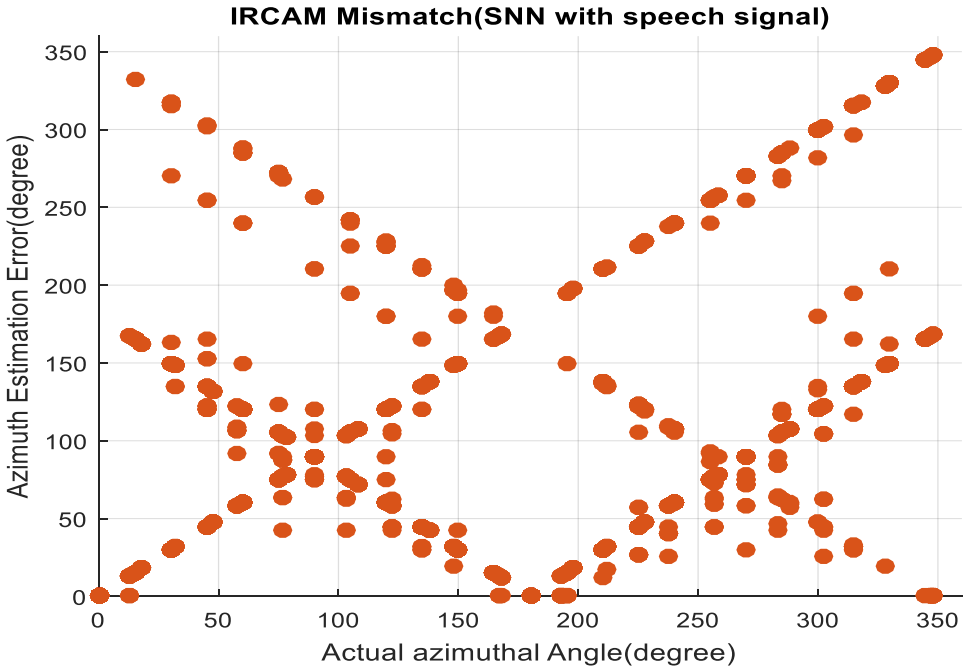


Figure 6.6: Estimation angle error of azimuth by SNN trained with speech sample convolved with IRCAM and tested with different speech samples convolved with KEMAR.

Again, both figure 6.5 and 6.6 demonstrate that sound source localisation based SNN with non-individual HRTFs encounter difficulty in recognising the locations due to the front back ambiguity in the vertical plane.

Figure 6.7 explains the SNN performance in estimating elevation angles with mismatched HRTFs when the localisation model trained using speech sample convolved with IRCAM and tested with KEMAR. The x-axis refers to the actual elevation angles while the y-axis refers to the predicted elevation angles

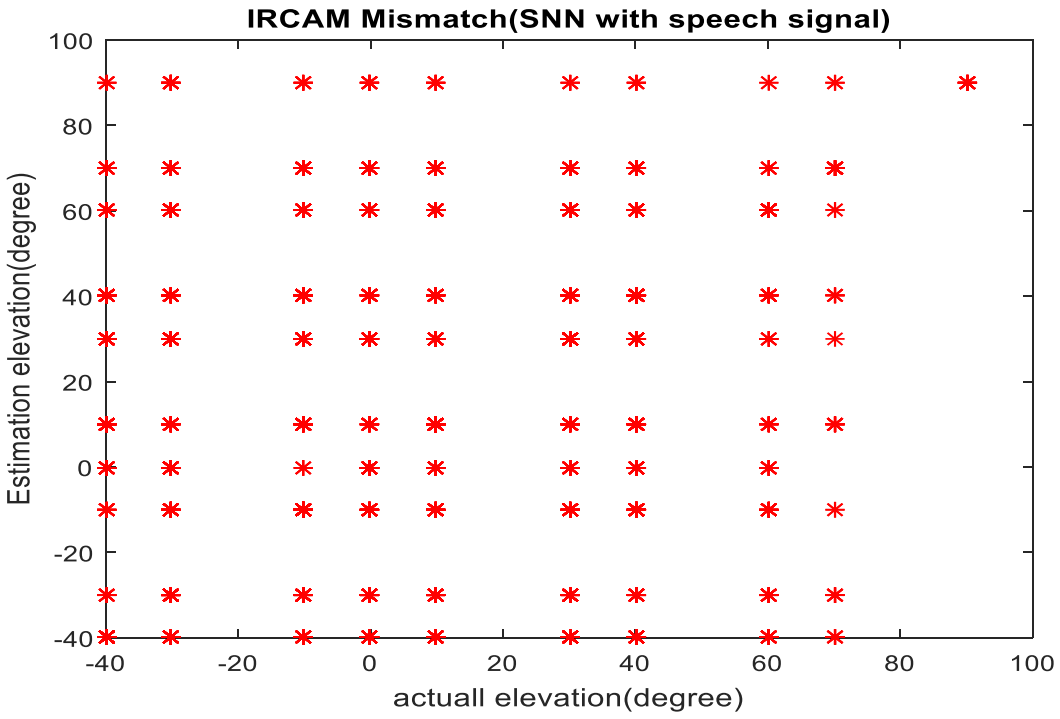


Figure 6.7: SNN performance in estimating elevation angles with mismatched HRTFs when it trained with speech sample convolved with IRCAM and tested with KEMAR.

The elevation prediction angle error is shown in figure 6.8 when the single sound source localisation model has been trained with different speech samples convolved with IRCAM HRTF ant tested with a new speech sample convolved with KEMAR HRTF. The x-axis refers to the actual elevation angles, and the y-axis refers to the elevation angle errors.

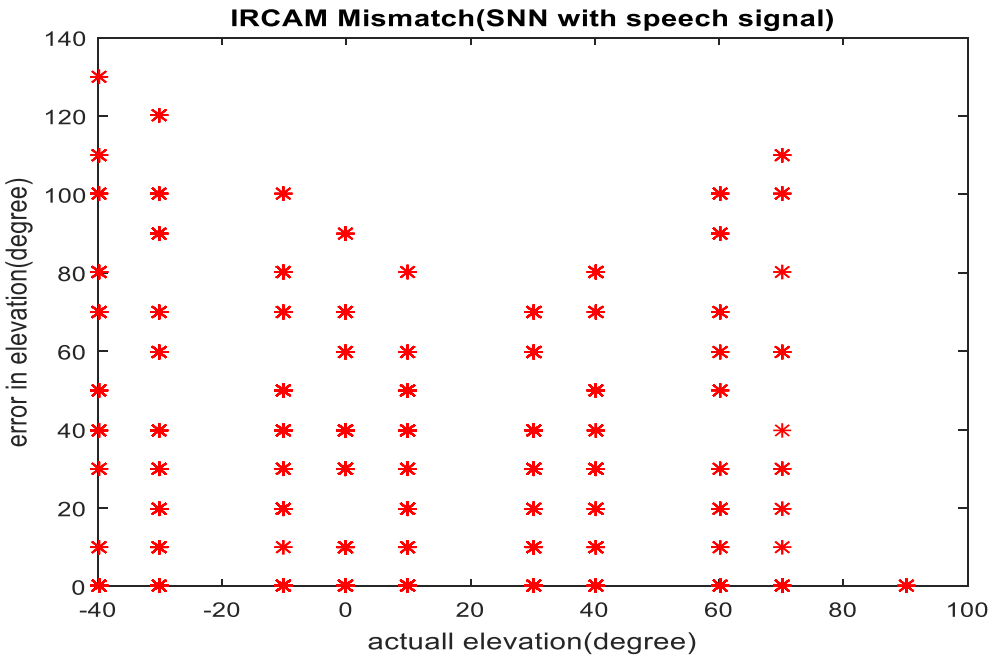


Figure 6.8: Estimation angle error of elevation by applying SNN with speech sample convolved with IRCAM in training and testing with KEMAR.

Experiment 3: SNN was trained with speech samples convolved with KEMAR and tested with IRCAM

To test the SNN localisation performance with mismatched HRTF when KEMAR is the training head, and the testing sound signals come through IRCAM HRTF, SNN has been trained with speech samples convolved with KEMAR and tested with IRCAM. The azimuth angle prediction performance is shown in figures 6.9 and 6.10. Similarly, to the previous experiment, the results demonstrated that the SNN based localisation model exhibits a high level in front back confusion and the prediction is wholly flipped between three angles 0˚, 180˚ and 360˚

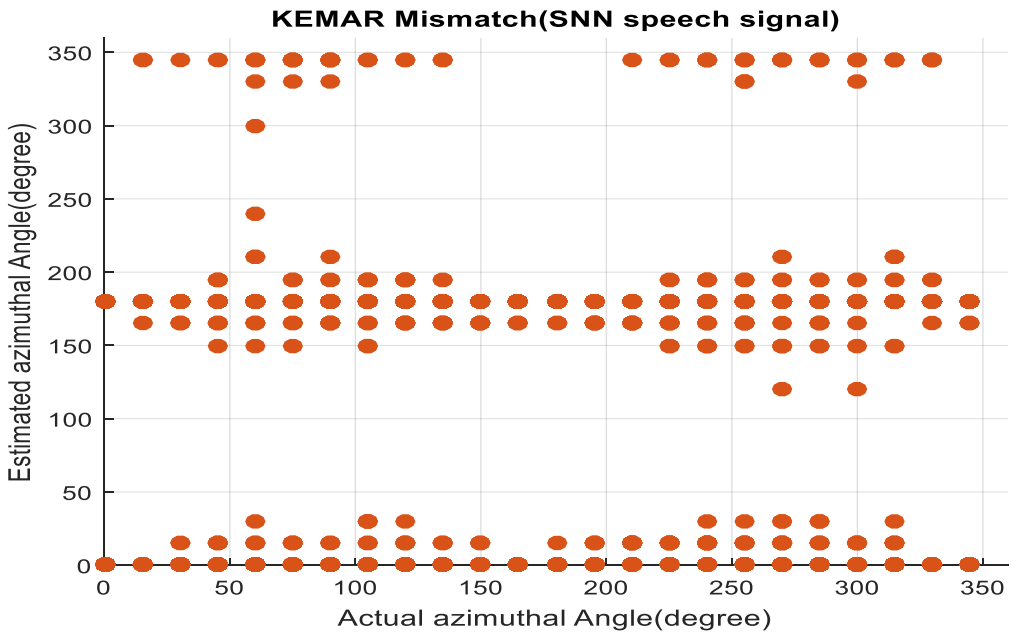


Figure 6.9: SNN performance in predicting azimuth angle when speech samples and KEMAR in training and tested with IRCAM.

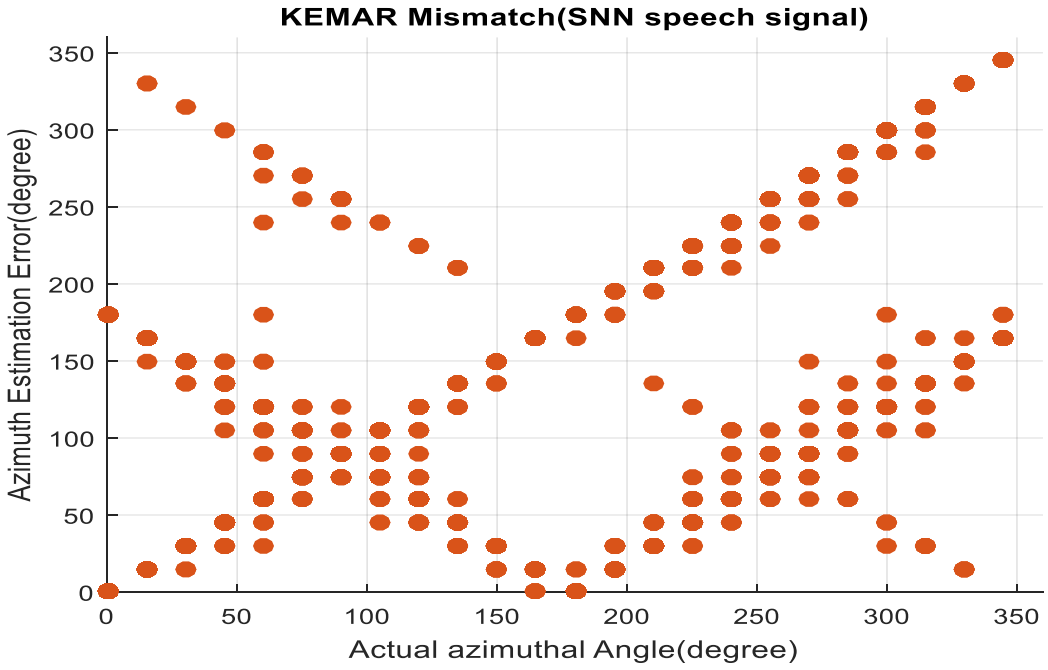


Figure 6.10: Estimation angle error of azimuth resulted from SNN with mismatched HRTFs when KEMAR in training and tested with IRCAM.

Figure 6.11 shows to the actual and predicted elevation angles from applying SNN localisation model with KEMAR in training and IRCAM in testing. The elevation angle errors that resulted from the computed the absolute angle error between the original and predicted elevation angles are shown in figure 6.12.

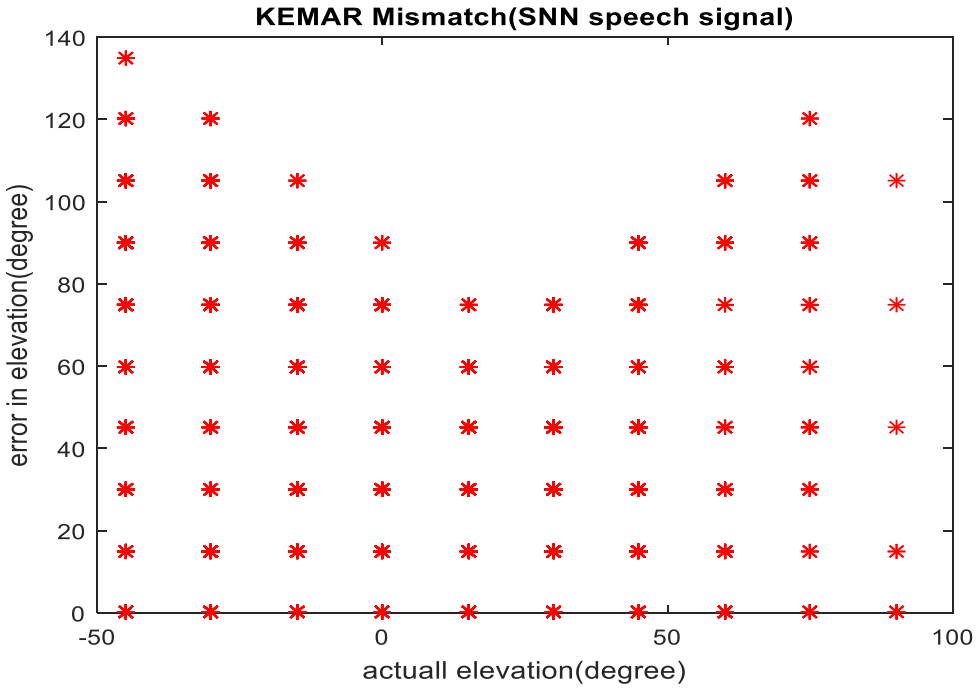


Figure 6.11: The SNN performance in predicting the elevation angles with speech samples and KEMAR in training and tested with IRCAM.

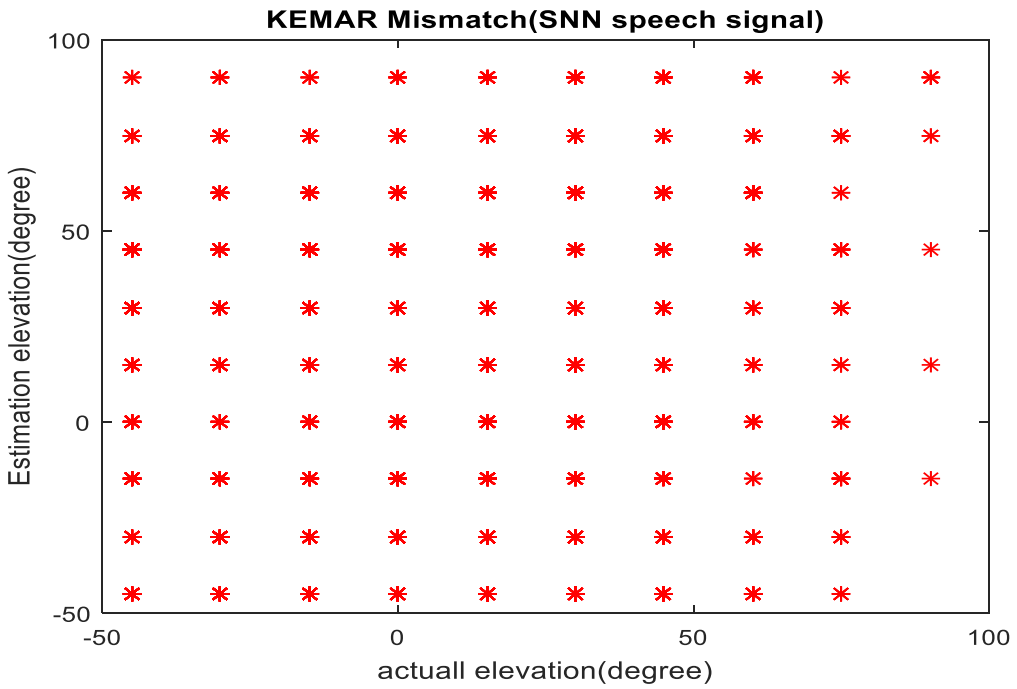


Figure 6.12: Estimation angle error of elevation angles by applying SNN with speech samples and KEMAR in training and tested with IRCAM.

6.4 Single sound source localisation based on different machine learning methods with mismatched HRTFs

In this section, the SNN applied as a pre-processing method to extract the binaural features from input signals. The resulted features (firing rates from the SNN) were used to train and test the SVM with linear kernel. The localisation problem is formulated as a classification problem where each class refers to a single source location. Here, the same models were tested to study their localisation performance with non-individual HRTFs.

These experiments included generated two datasets; the first consisted of different speech samples generated using two speakers (Male and Female) from the SALU-AC speech database convolved with the KEMAR HRTF. These samples were used to train a SVM. The machine learning method was tested by using data generated from convolved speech samples with the IRCAM HRTF. Figure 6.13 shows the SVM performance in predicting azimuth angle when speech samples and IRCAM in training and tested with KEMAR.

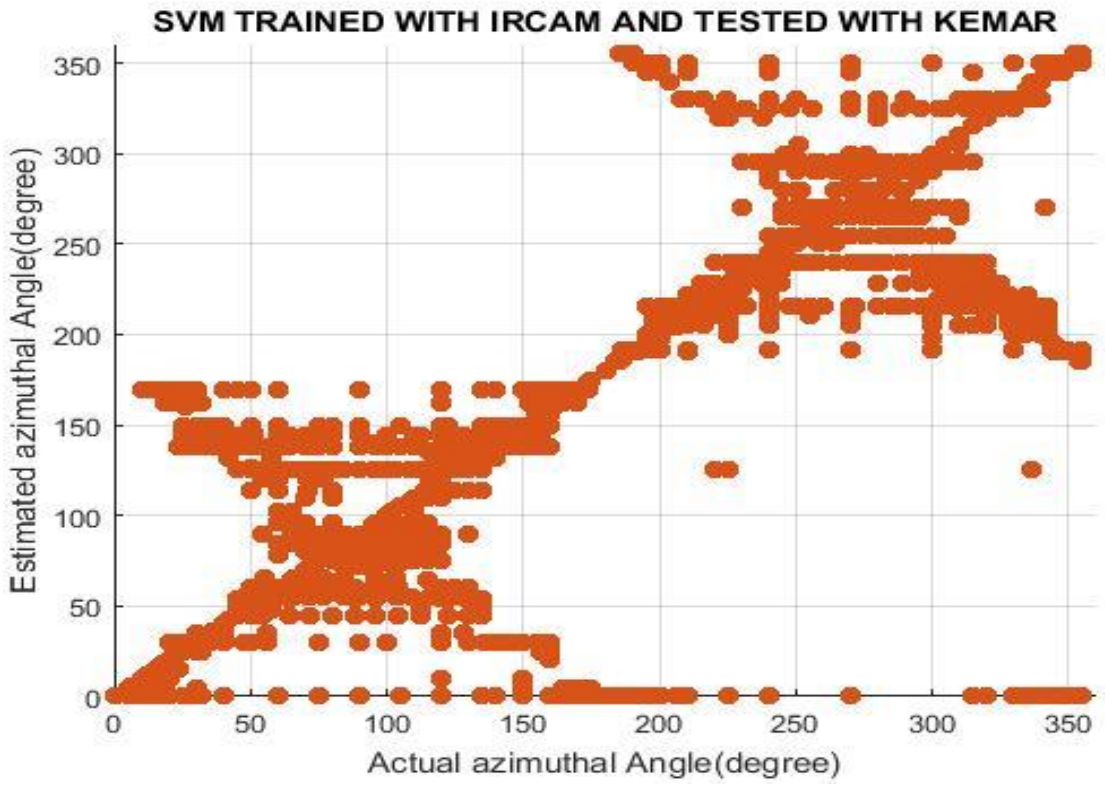


Figure 6.13: SVM performance in predicting azimuth angles when IRCAM in training and tested with KEMAR.

In figure 6.13, the x-axis represents the actual azimuth angles, and the y-axis refers to the estimation angle error that computed from finding the differences between the actual and predicted azimuth angles. Figure 6.14 shows the SVM performance in estimating elevation angles. The x-axis refers to the actual elevations while the y-axis refers to the elevation angle error resulted from computing the absolute difference between the actual and predicted elevation angles.

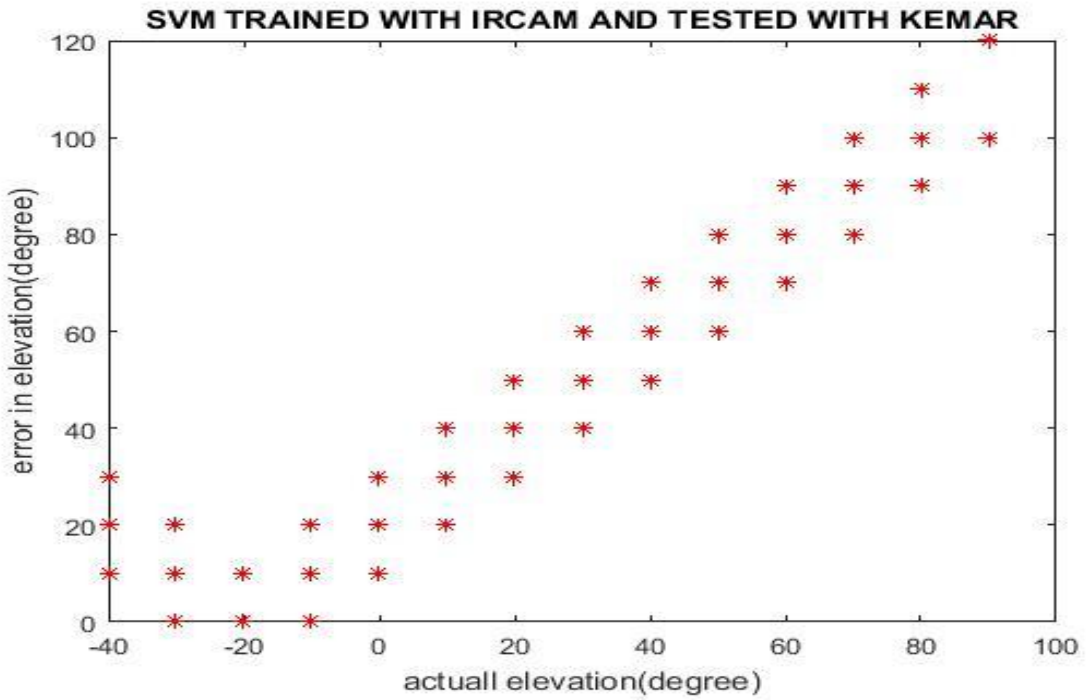


Figure 6.14: The angle error of elevation from SVM trained with IRCAM and tested with KEMAR.

Table 6-2: The azimuth and elevation estimation accuracy by applying SNN, SVM and random forest with non-individual HRTFs.

The localisation model with non-individual HRTFs	Azimuth estimation accuracy ± 15˚	Elevation estimation accuracy ± 15˚
SNN	0.16	0.28
SVM	0.52	0.44
Random forest	0.48	0.41

Table 6.2 shows the azimuth and elevation estimation accuracy by applying three machine learning models (SNN, SVM and random forest). The accuracy has been computed from the signed angle error for both azimuth and elevation.

The accuracy increases when SVMs or random forests were used to compensate for the HRTF mismatch. However, the front-back confusions remain the main source of error. A

systematic or bias error is visible figure 6.14. This type of error clarifies the divergences are not due to chance alone. The proposed localisation model is unable to compensate for the non-individual HRTF problem because of the main issue is due to the difference in size between IRCAM and KEMAR. The size differences between the KEMAR dummy head and IRCAM subject affects the time of arrival at both ears that caused an ambiguity in the ITD. This ambiguity caused an increase in the front-back confusion error.

6.5 The multisource localisation models with non-individual HRTFs

In this section, the multisource localisation model from chapter 5 based DNN is tested with mismatched HRTFs. The multisource localisation model was being trained with IRCAM HRTFs and tested with speech samples convolved with KEMAR HRTFs. The experimental outcomes of multisource localisation model with non-individual HRTFs are shown in the following figures.

Figure 6.15 shows the confusion matrix plot for source one predicted by multisource localisation model with mismatched HRTFs when the model has been trained with IRCAM and tested with KEMAR. The training set included all the azimuth angle range of 0 to 345 at 0 elevation level of IRCAM. In contrast, the testing data contains the azimuth angle range from 0 to 345 at 0 elevation level of KEMAR.

In the first experiments that showed the localization performance for the single source model based on SNN, the model has been trained with data from IRCAM i.e. generate training data from IRCAM, then test the model with data from KEMAR. Generated data from IRCAM with IRCAM embedded in SNN and the test data generated from KEMAR HRTF but with IRCAM embedded SNN. And, this simulated the mismatched HRTF and the experimental outcome demonstrated high angle error due to the front-back and up-down confusions. The multisource localization model that based on processing the SNN output firing rates using DNN was used to test the mismatched HRTFs. The following experimental results also show high angle error and front-back confusion in spite of the model has been trained with data from IRCAM with SNN embedded IRCAM and tested with data generated from KEMAR with SNN embedded with KEMAR.

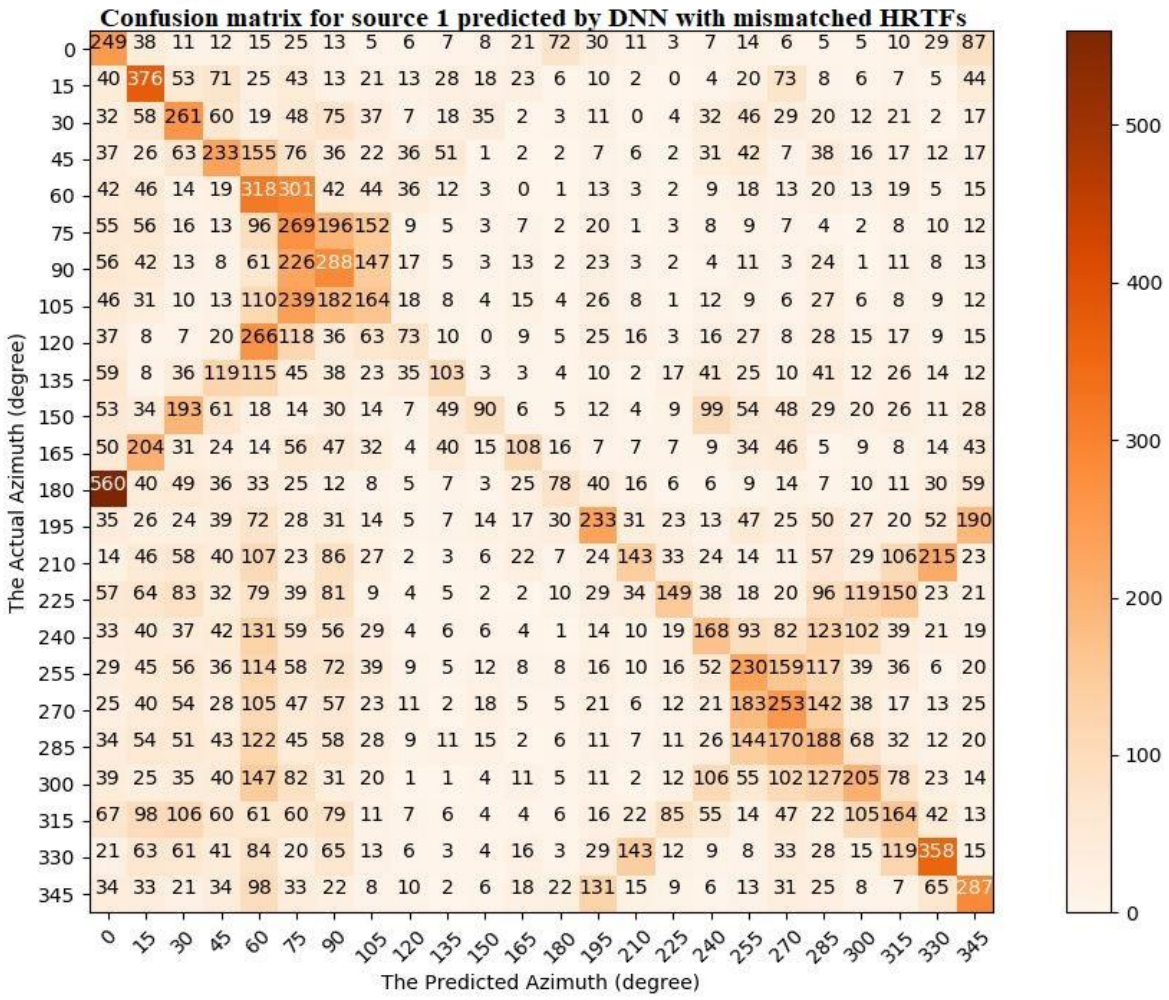


Figure 6.15: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation model with mismatched HRTFs (IRCAM in training and KEMAR in testing).

Figure 6.16 shows the confusion matrix plot for the source two predicted by multisource localisation model with mismatched HRTFs when the model has been trained with IRCAM and tested with KEMAR. In the figures 6.17 and 6.18, the x-axis refers to the predicted azimuth while the y-axis refers to actual azimuth value.

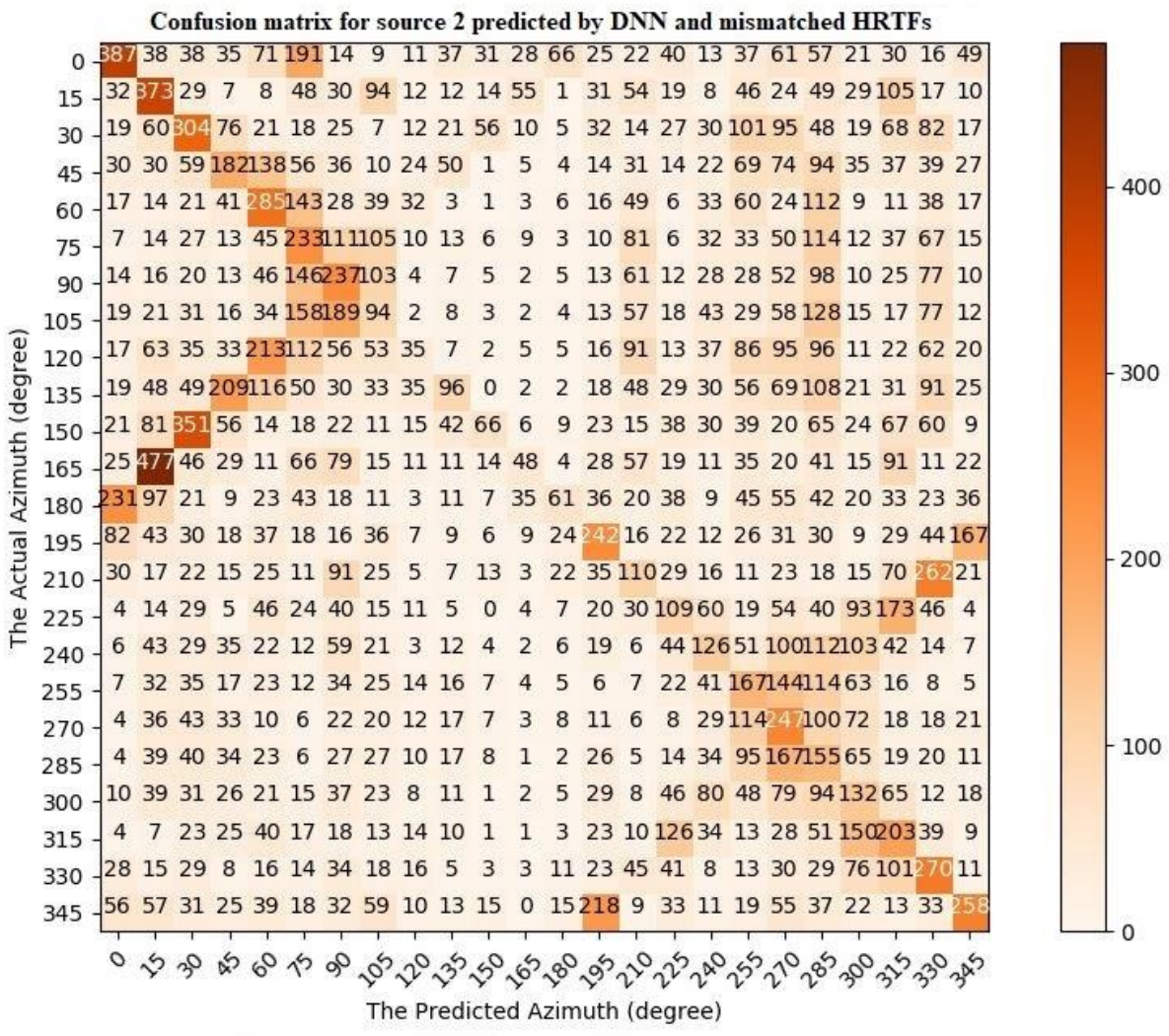


Figure 6.16: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation model with mismatched HRTFs (IRCAM in training and KEMAR in testing).

Figure 6.17 and 6.18 explain the confusion matrix plots of estimation angle errors for source one and source two. The x-axis refers to the predicted azimuth angle error, and the y-axis refers to the actual azimuth values.

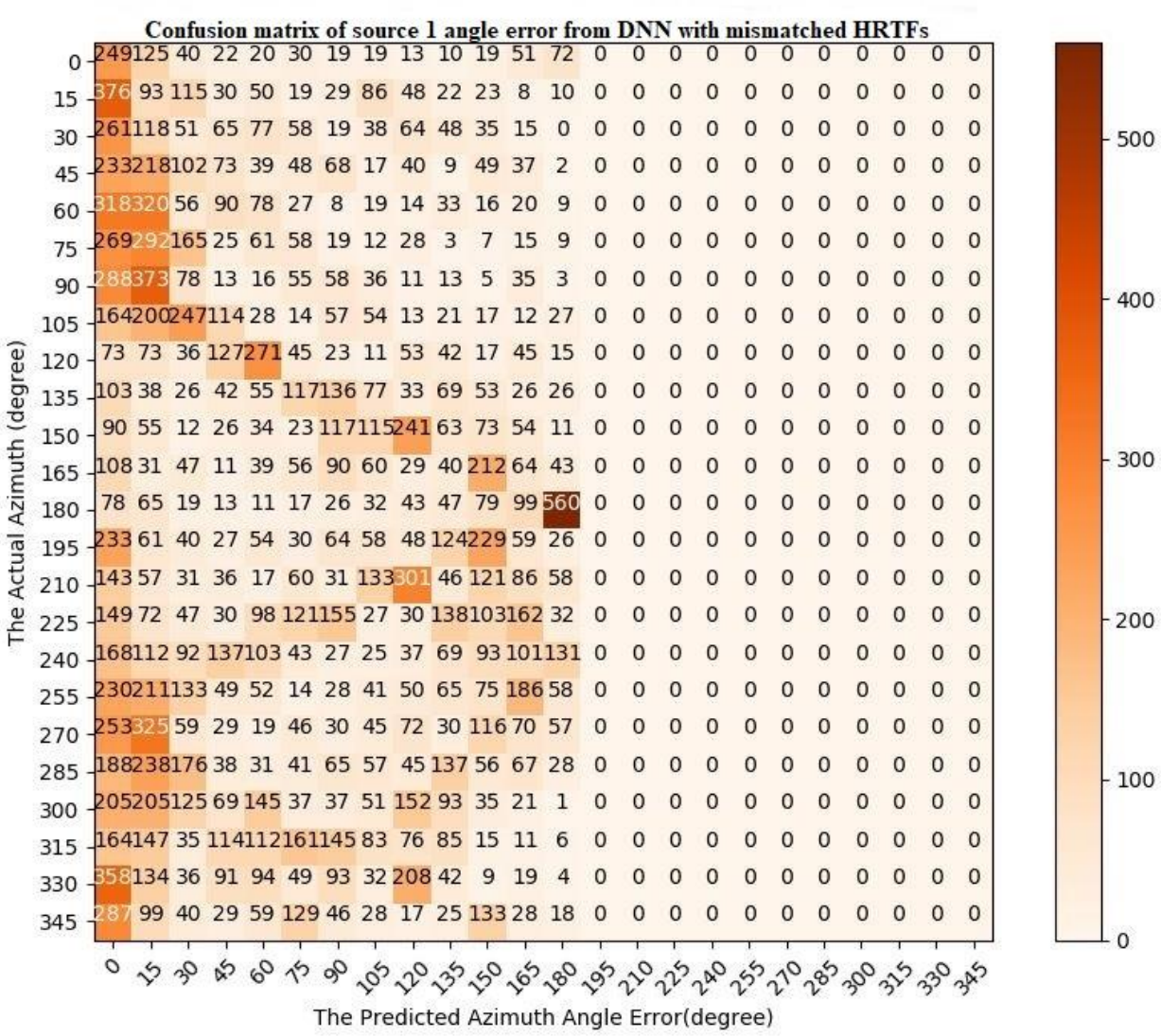


Figure 6.17: The sources one azimuth angle errors from applying multisource localisation model with mismatched HRTFs (IRCAM in training and KEMAR in testing).

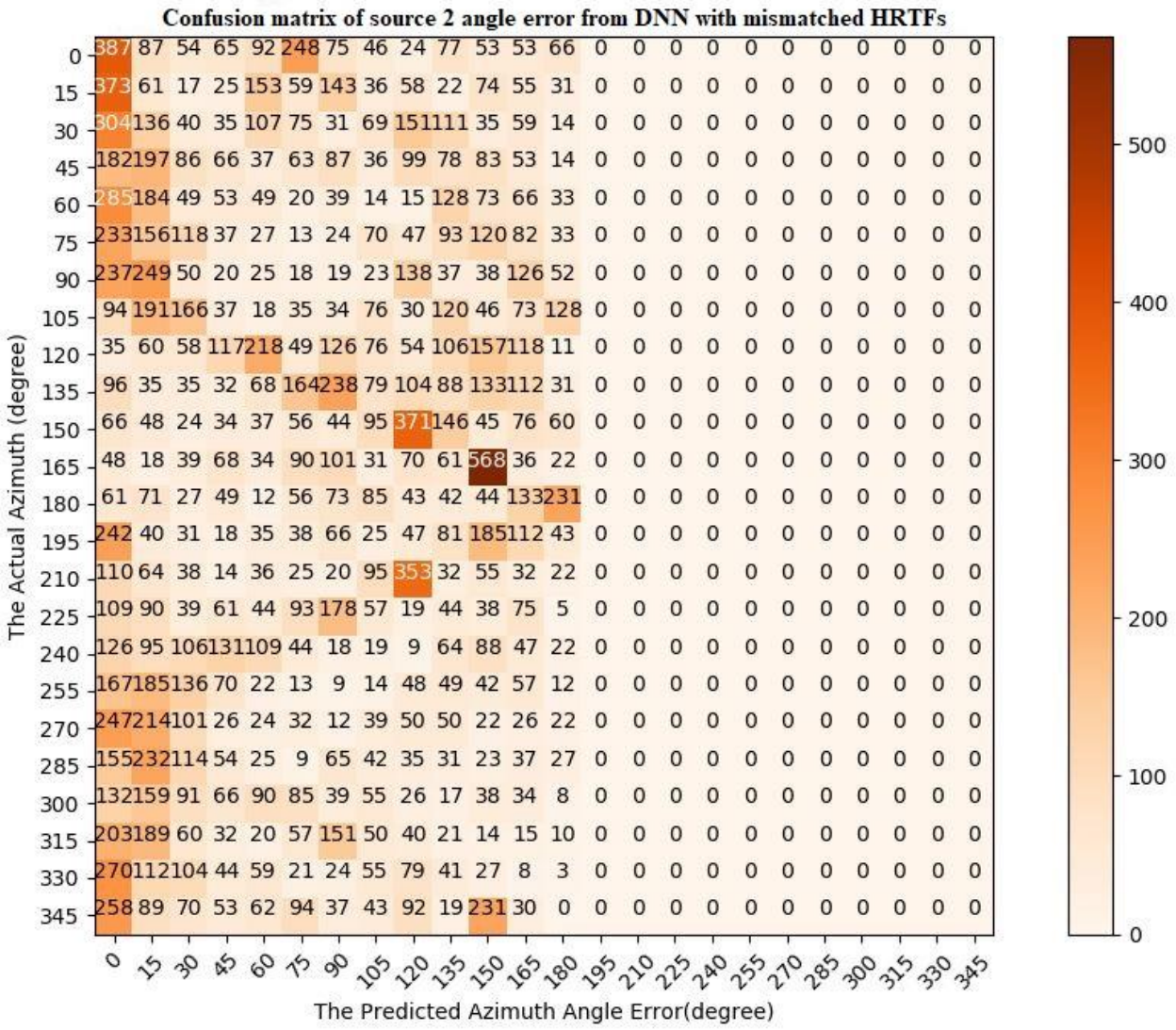
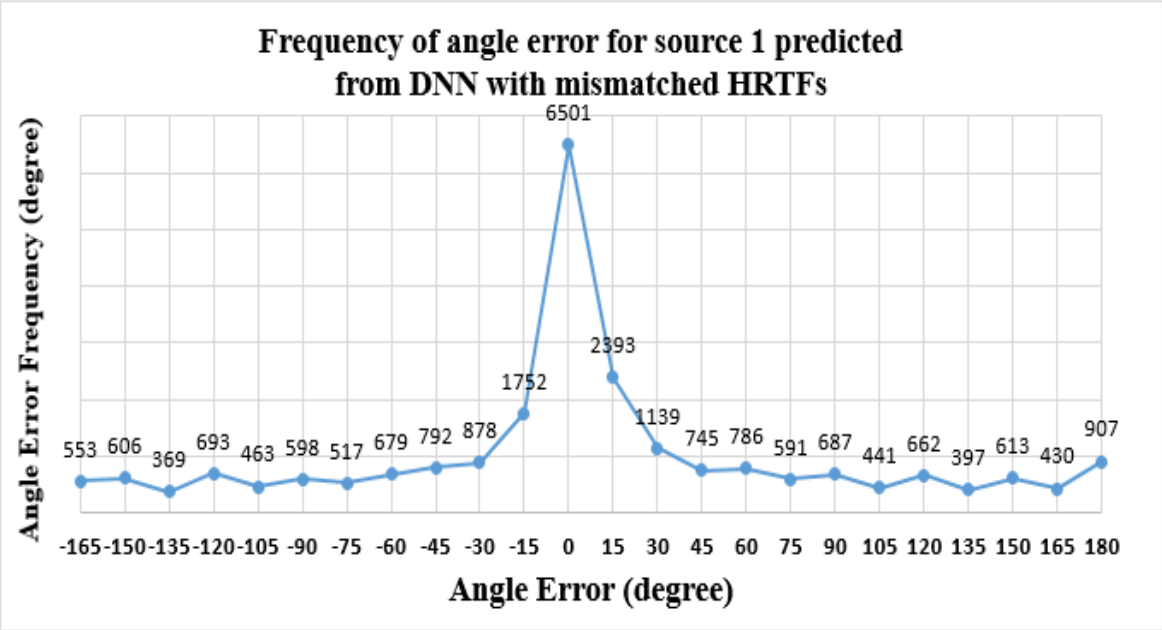


Figure 6.18: The sources two azimuth angle errors from applying multisource localisation model with mismatched HRTFs (IRCAM in training and KEMAR in testing).

Figure 6.19 represents the frequency of angle errors for source one, and source two predicted DNN based multisource localisation model. The total size of testing data that shown in this figure is 24129 output points. The figure demonstrates the higher levels of front-back confusion to predict source one and much higher for source two.



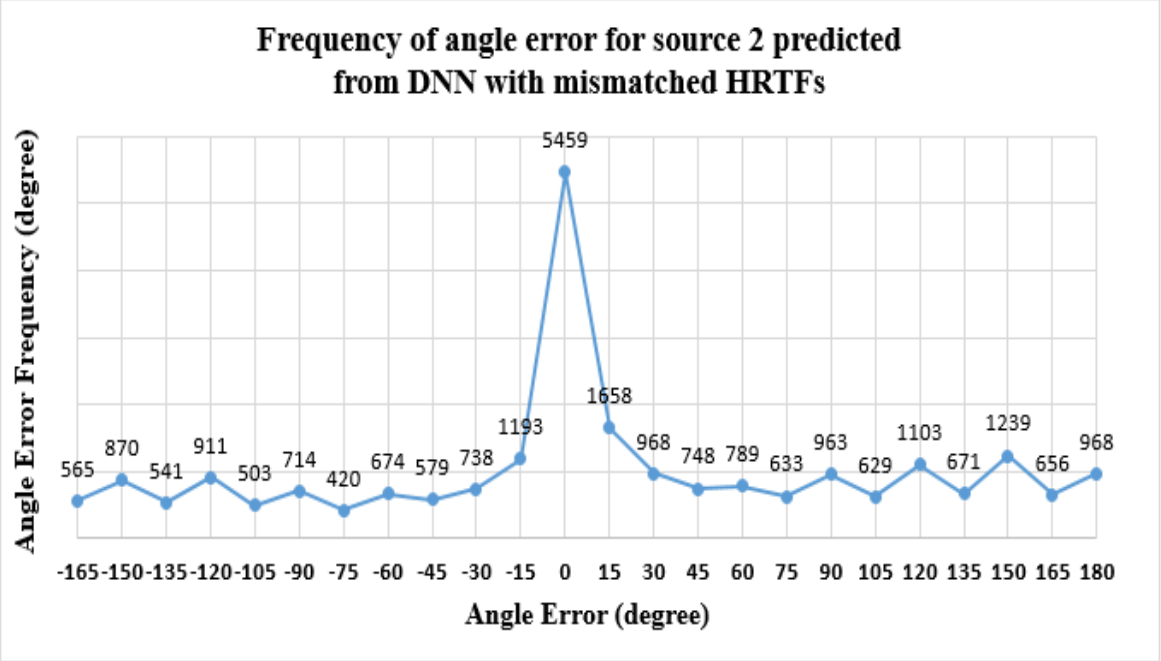


Figure 6.19: source one and source two angles errors frequency from applying multisource localisation model based with nonindividual HRTFs

The model localisation accuracy has been calculated from computing from signed angle error for source one and source two. The DNN based multisource localisation model achieved 44% accuracy for source one and 34% for source two. It is notable that the machine learning alone was unable to solve the non-individual HRTF problem. That because it is needing to generate a huge labelled data by using many HRTF databases that represent different head and panna structures to train the machine learner. Current research has suggested method to compensate for nonindividual HRTFs, for example; scaling the HRTF to the individual using morphological criteria tuning of spectral cues; using numerical computations and subjective selection; and adjusting ITDs to the individual (Lindau 2010). The ITD cues for each angle of these two HRTF datasets have been computed to clarify the difference between the two HRTFs, as shown in figures 6.20 and 6.21. The ITD computed from estimating the interaural time delay Δt by looking for peaks in the cross correlation between the left and right channels.

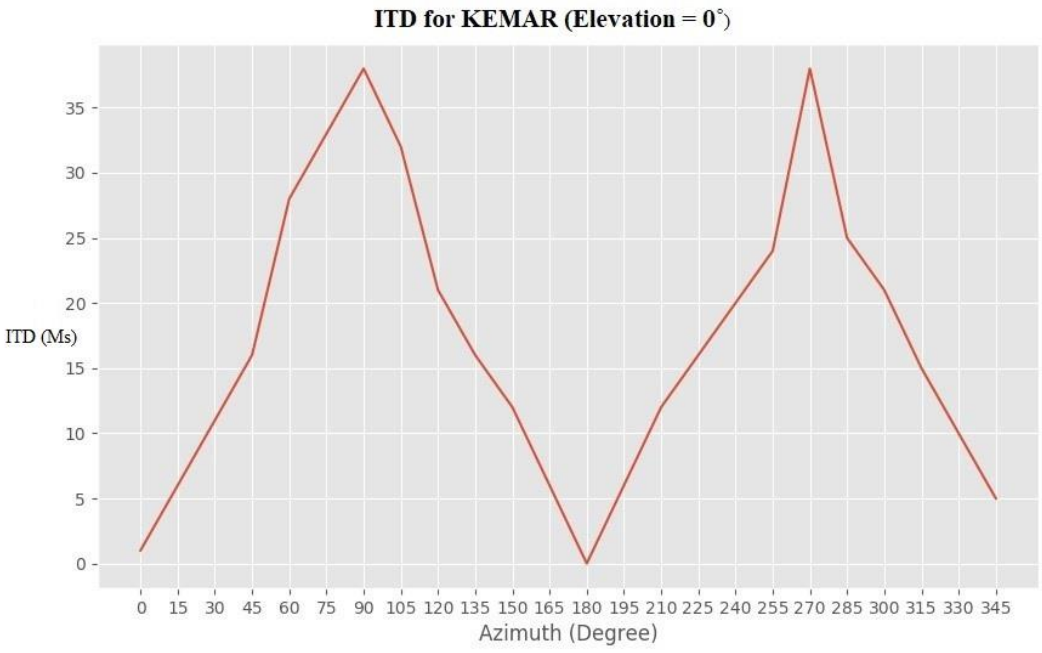


Figure 6.20: The ITD for KEMAR dummy head.

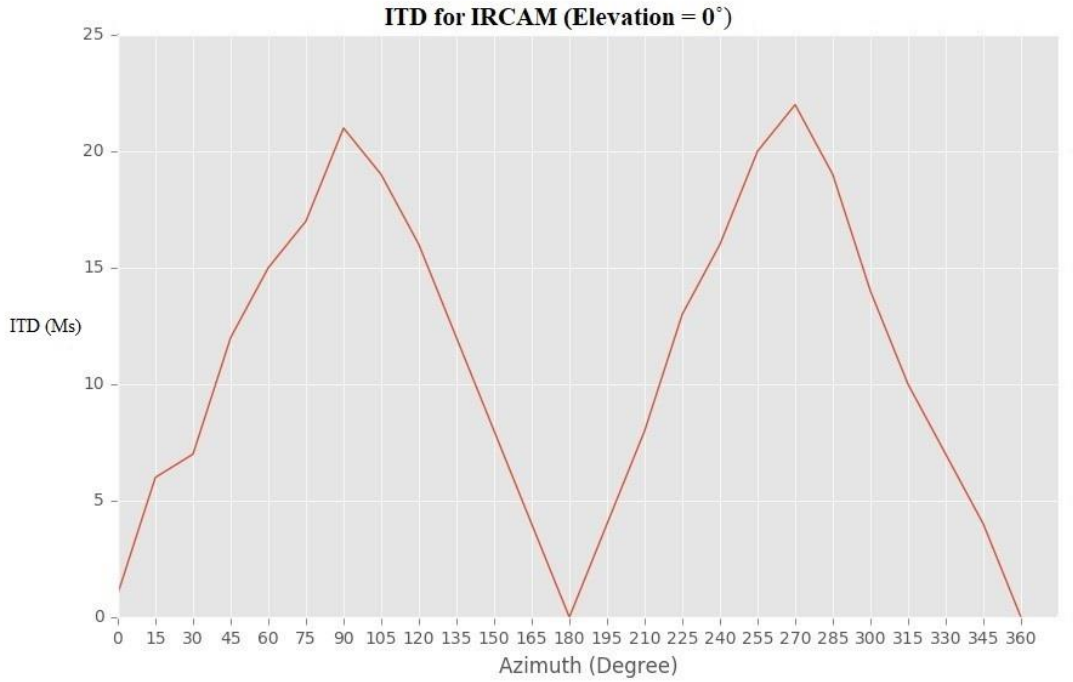


Figure 6.21: The ITD for IRCAM subject.

Figure 6.20 shows the time differences for KEMAR dummy head for each angle at elevation 0. Figure 6.21 shows the time differences for IRCAM at elevation 0. The mismatched HRTFs caused by the difference in the ITD of these two HRTFs. The ITD of IRCAM is less than the ITD of KEMAR dummy head.

As future work, the ITD may be adjusted to an individual as a reasonable solution for mismatched HRTFs. The time differences cue of IRCAM has been adjusted to match the KEMAR time differences. Figure 6.22 shows the IRCAM time differences cues after an adjustment to match the KEMAR time differences cues. This however requires new data to test the machine learning models for single source and multisource localisation.

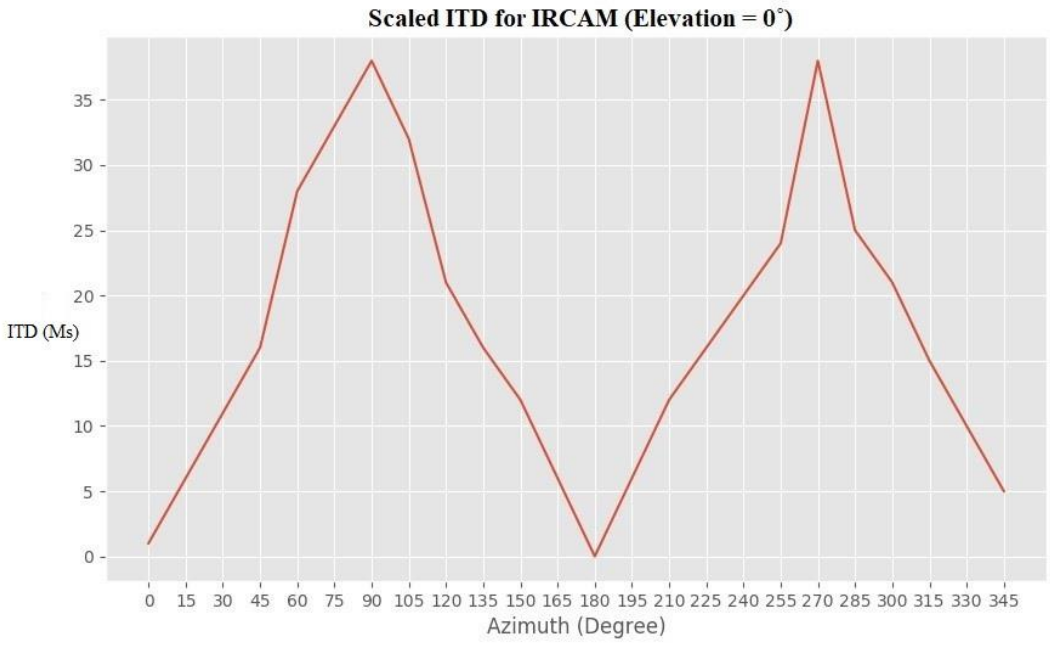


Figure 6.22: Scaled ITD for IRCAM to match the ITD of KEMAR.

6.6 Chapter Discussion

In this chapter, the localisation models have been tested with mismatched HRTFs. Two types of models have been presented, the first one where the SNN coincident neuron with the adopted the SNN as a binaural feature pre-processor and different machine learning methods were applied to predict the incoming sound signal angles. These two localisation frameworks were also implemented for single and multi-sources with mismatched HRTFs. Significantlocalisation angle errors have beenindicated through the experiments due to the front-back confusion for both single source and multisource localisation models. Also, the chapter reviewed some suggested solutions to tackle the HRTFs mismatched problem. One of these solutions is adjustment the time differences cues to the individual. The interaural time differences for KEMAR and IRCAM have been computed. The ITD of IRCAM subject was adjusted to match the ITD of KEMAR dummy head. The adjusted HRTF could contribute in a future experiment by generatingnew data and applying them to the localisation models with non-individual HRTFs.

CHAPTER 7

CONCLUSIONS AND FUTURE WORKS

The main goal of this research study is concerned with using the maximum firing rate over an assemblage of spiking neurons to locate sound sources in three dimensions using only two sensors. The spiking neural network utilised as a feature extractor where output firing rates were processed and various machine learning methods including DNN and SVM trained to predict the locations of multiple simultaneous sources. This novel idea has been tested with different machine learning algorithms and has shown excellent performance for single and multisource localization. The DNN learns important patterns over the assemblage of firing rates to enable successful localisation; non-linearly separable data needs non-linear learner.

This research has developed a new model to solve the multisource localization problem which is robust localization model and applicable for real time applications. Different machine learning approaches have been compared and their effectiveness sound source localization in the presence of environmental noise examined. The speech data used in the project was from a database (SALU-AC) which contains audio speech samples recorded in different languages in addition to English. These samples were recorded without being limited to particular text messages. The algorithm was tested data, held-out from training, from a number of talkers to ensure generalisability.

Furthermore, two types of noise were investigated; diffuse and directional noise. The research summary and conclusions, in addition to some of future work suggestions, are giving in the following sections.

7.1 Summary and conclusion

1. The localization model based on a spiking neural network presented by Goodman is reviewed and replicated with two HRTF data sets, KEMAR dummy head the IRCAM data set. The method is inspired by the way humans estimate the location by using the binaural features such as interaural time difference ITD and interaural level difference ILD. HRTFs have been employed to acquire the binaural information. The SNN presents more realistic representation of human hearing by

mimicking the binaural time delays in its simulations. The frequency features of input responses were analysed by using set of gammatone filter bank. The localization model has been tested with diverse types of input signals including gaussian white noise, uniform white noise, pure tone modulated white noise and different speech samples that were collected in an anechoic environment. In addition, the localization model performance was investigated with single frequency and octave frequency to demonstrate the effectiveness of localization cues on the localization model performance. Two localization related performance factors were examined; the input signal duration and the number of gamma-tone frequency channels and their impact on localization model robustness are explained. The results explain the enhancement of the localization performance by increasing the input signal duration as well as the number of gamma-tone frequency bands. Furthermore, signal to noise ratio is shown to play a significant role in the robustness of the localization model. The model has been examined with different SNRs to identify the effect on performance in various levels of background noise. The outcomes show the variation of the effect of different SNRs on the performance of single sound source localization.

2. The spiking neural localization model was expanded and tested to localize two simultaneous sound signals emitted from two separated locations. The experimental results demonstrated that the SNN based localization model was unable to process the spiking neural firing rate to accurately locate the two sources due to the ambiguity in the input signal results from mixing two sound signals. The spiking neural based localization model output firing rates was processed using various machine learning methods including DNN and SVM. A novel idea for sound localization by using only two ears has been presented. The spiking neural networks (SNN) model is utilised as a binaural feature extraction algorithm to extract the timing information from the binaural responses. Various machine learning algorithms were then trained and compared predict source locations from the firing rates. Its performance has been compared with SNN based localization model for a single source. Varied sizes of labelled data have been generated to train and validate the machine learning models. The localisation problem is formulated as a

classification problem where each class represents a single source location. The results show differences in the performance of the various machine learning approaches in localizing single sound source. Also, they demonstrate some differences between IRCAM and KEMAR impact on the localization performance. These differences were handled by increasing the size of training data. For evaluation and comparison purposes, the DNN localization performance was compared with other machine learning methods. Multisource localization models based on SVM and SNN were investigated to study their performance in localizing multi sound sources. The results from these two methods were analysed and compared with the DNN localization performance. SNN with a 'two-winner-takes-all' concept was implemented to detect the two locations. SNN based multisource localization model showed poor performance in estimating source one with slightly better performance in predicting source two. SVM classifier performed better than SNN but still its performance limited and less than DNN in predicting both sources.

3. The SNN firing rate features were used to train a DNN to perform a classification task. In this case, the DNN learned from examples, where each example is associated with two predefined labels (the location of source one and source two). This novel idea has been tested and the outcomes with different machine learning algorithms have been demonstrated. The DNN showed a better localization performance compared with SVM and SNN. The DNN can learn important patterns in the data to enable successful localisation. Also, the non-linearly separable data, needs non-linear learner for the best performance. So that, the SVM with linear kernel showed a poor localization performance.
4. A novel idea for improving the method for multi-source sound localization by using only two ears has been presented. The localization process has two steps: The first step is to predict the number of sources in the incoming signal by analysing the SNN firing rates. Once the number of sources is known the appropriate localization model (single or multisource localisation) can be selected in order detect the source directions. Logistic regression was applied to create a best fit logistic curve to sperate between the two sources and one source signals. The model showed a better

performance in predicting the number of sources from different speech signals and even under noisy conditions.

5. The localization model was first tested in a task to localise single sound sources emitted from a unique location. Different speech samples belonging 100 speakers contributed to train and test the single sound source localization model. The localization model was then extended to two simultaneous sources generated from all possible combination of 17 speakers and different 3 speakers for validation.
6. Two types of machine learning methods, SVM and DNN, were applied to process the spiking neural networks firing rate features for multisource sound localization. Firstly, the deep neural network was examined for multisource localization which returned a high accuracy of 91%for the one of input sources and 89% for another source. Moreover, the angle errors between the actual and predicted locations have been analysed. Two types of angle errors have been determined; front-back confusion and left-right angle error, comparatively modest error on the range from $\pm 5^{\circ}$ to $\pm 15^{\circ}$ and the characteristic form of errors recognised as back-front confusions. There are no significant left-right error probabilities observed in the multisource localization model experiments. Whereas the source prediction accuracy of the multisource localization model was frequently affected by a front-back confusion error type. In this case it is important to mention and take in an account that these experiments used a static head which brings more complexity to deal with sound signals that are issued from the back.
7. The experiment results demonstrate that the localization accuracy enhancement highly depended on the number of training samples that were used to train the deep neural network. The experimental outcomes demonstrate that the localization performance of multisource localization model is improved by increasing the number of speakers tin the training data sets. And from the machine learning perspective, this is reasonable due to the increase teaching examples of the machine learning models. The machine learning model depend on find the function to map the input data to the output data. This mapping function required enough data to capture the relationships between the input features from a side and between input features and output features from another side. To test this practically, first the model

was trained with data that was generated by using only 10 speakers where only 45 possible combinations between these speakers participated in constructing the training data. The position estimation accuracy for both locations with ±15 angle degree did not exceed 55%. To improve the multisource sound localization performance, the number of speakers is raised to be 17, producing 136 possible combinations between speakers. Thus, the multisource estimation has been boosted by achieving localization accuracy to within 90% and 89% within ±15˚ for source one and source two respectively as shown in table 5.7.

8. The impact of background noise on the on the multisource localization model performance have been examined in three experiments. Firstly, the localization model was trained with clean data and tested with noisy data at different SNRs. Secondly, the multisource localization model was trained with multi-condition background noise at SNRs of 10dB, 0dB, and -10dB and tested at controlled SNR. The findings demonstrate an enhancement in the model performance in predicting source one and source two when the model trained using noisy data. The final experiment examined the impact of the directional noise on the multisource localization model performance. It is easy to extract the useful information or detect a true signal from the raw signal at the higher SNRs due to the power of a signal is higher than the power of the background noise. Experimentally, the localization model has been tested with poor sound signals at low SNRs at 10dB, 0dB and -10dB. While, the better human hearing is at SNR 30 dB and above. The findings have been demonstrating an enhancing in the localization performance by increasing the signal to noise ratio while the minimum signal to noise ratio for this system was -10dB. Knowledge of this ratio has many important applications that related with enhance the hearing experience. For example, people who use the hearing aids.
9. The experiments have been done using two types of HRTF databases; IRCAM and KEMAR dummy head. Each one of these data has special impact on the multi-source localization model performance due to the differences in the anatomical parameters (head size, ear shape and torso). Also, using two different HRTFs to test the multisource advocates the model generalisation.

10. The localisation models have been tested with mismatched HRTFs. Single-source and multisource localization frameworks were implemented with mismatched HRTFs. Significant localisation angle errors have been indicated through the experiments due to the front-back confusion for both single source and multisource localisation models. Also, the chapter reviewed some suggested solutions to tackle the HRTFs mismatched problem. One of these solutions is adjustment the time differences cues to the individual. The interaural time differences for KEMAR and IRCAM have been computed. The ITD of IRCAM subject was adjusted to match the ITD of KEMAR dummy head. The adjusted HRTF could contribute in a future experiment by generating new data and applying them to the localisation models with non-individual HRTFs.

7.2 contribution to knowledge

In this work, one of the main contributions to knowledge is the proposal of a smart localization model to solve the multi-source localization challenge. The model has been tested with various sound signals and under different noise conditions. The contribution is summarized by deep examination of two levels of neural networks (SNN and DNN) and linking between them to present an ideal solution for binaural hearing issues and sound signal processing. Different hearing-related transfer function data sets have been checked to explore the influence on localization performance. Several machine learning models have been examined to test their strength in performing single source and multi-source localization by using only binaural signals. The non-individual HRTF problem was examined and the experimental results showed that applying machine learning to solve the mismatch HRTFs was subject to availability of enough training data. This data refers to different subjects (different anatomical structures). Also, this work reviews some suggested solutions to tackle the non-individual HRTF issue that rely on adjustment of the time differences cues received at the two ears of the individual.

7.3 Suggestions for Future Works

This section briefly gives some suggestions for future work that may be adopted to expand the work given in this research:

1. Examine the performance of multisource localization model for more than two sources, three or four sources are mixed together. The process needs generating a new data has combination between the three input sources or four input sources to train and test the machine learning model. It solves the multi-source localisation when number of sensors (two sensors) is less than sources.
2. Examine the performance of multisource localization model under reverberation condition. To investigate the localization performance in enclosed environments (e.g. room) when the sound produced in a space is reflected off surfaces, like walls, the floor or the ceiling. The reflected sound will lead to generate many sound images for the original sound that may have bad impact on the localization performance.
3. Investigate the localization performance with others deep learning models as like convolusion neural networks CNN to examine it performance to solve the multisource localization challenges under different conditions; more than two sources, background noise and reverberation environments.
4. This research is focused mainly on solving the multisource localization challenge by successfully localizing two simultaneous sound sources. Future work may focus on using the model to improve speaker recognition task, knowledge of the location of a source can improve the performance of source separation algorithms.
5. Applying the Long Short-Term Memory (LSTM) network as a special type of recurrent neural network (RNN) to process the time and frequency representations in the firing rate input features. The RNN is well suited for the analysis of time series data and may be more successful than applying static neural networks to time averaged data.
6. Explore the spiking neural models for localizing multisource sound signals. And, compared their performance with the current spiking neural model (leaky integrated and firing model). Additionally, including learning into the spiking neural network will significantly improve performance and negate the need for the DNN.

APPENDICES

Appendix I

Additional Plots from Chapter 4

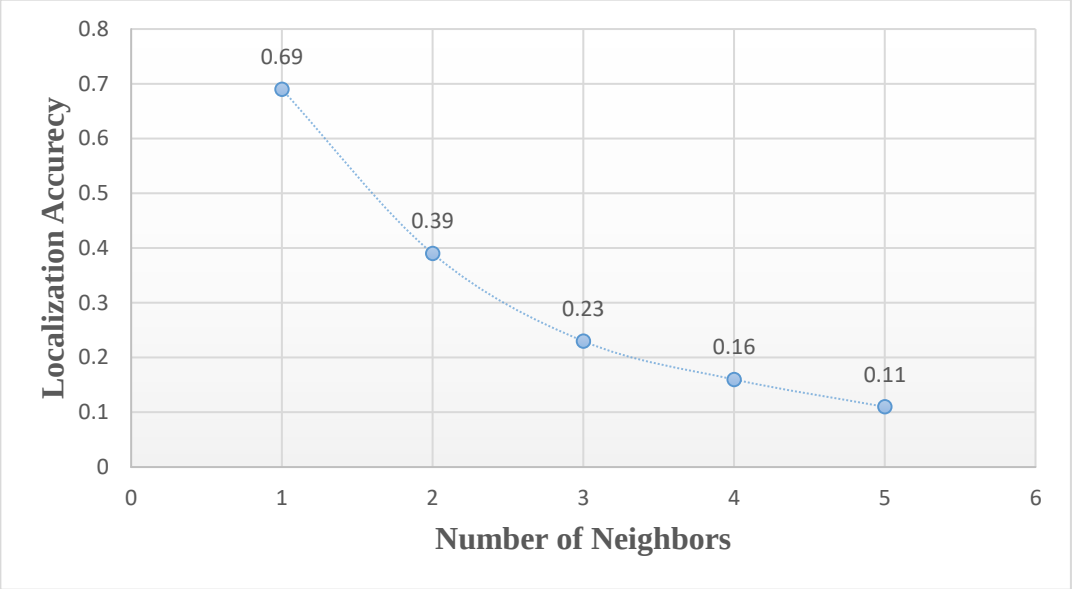


Figure I.1: KNN machine learning number of neighbours and its effect on localization accuracy using 187 different instances of white noise (500 ms duration).

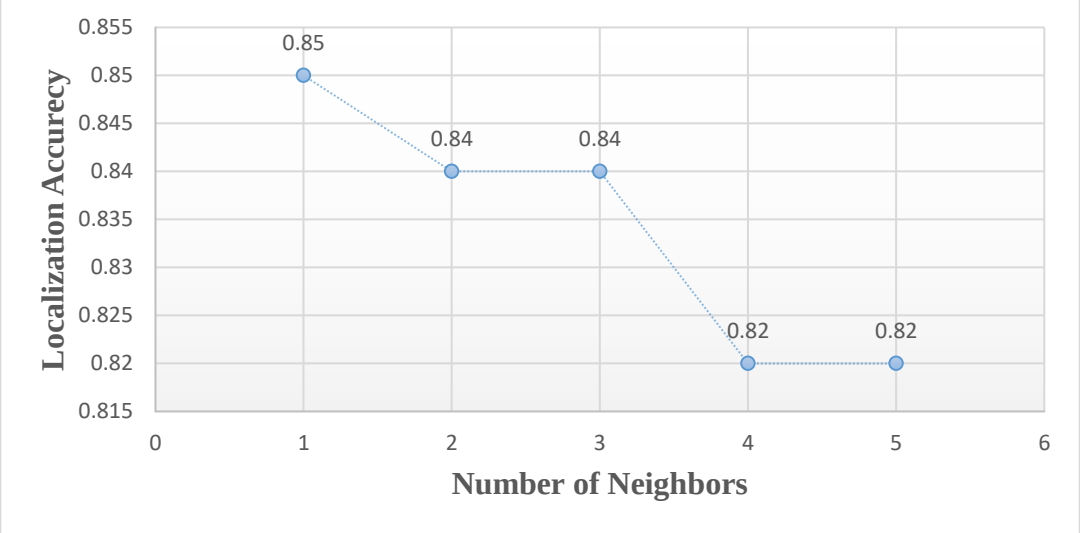


Figure I.2: KNN machine learning number of neighbours and its effect on localization accuracy using 187* 20 different instances of white noise (500ms duration).

Figure I.2:

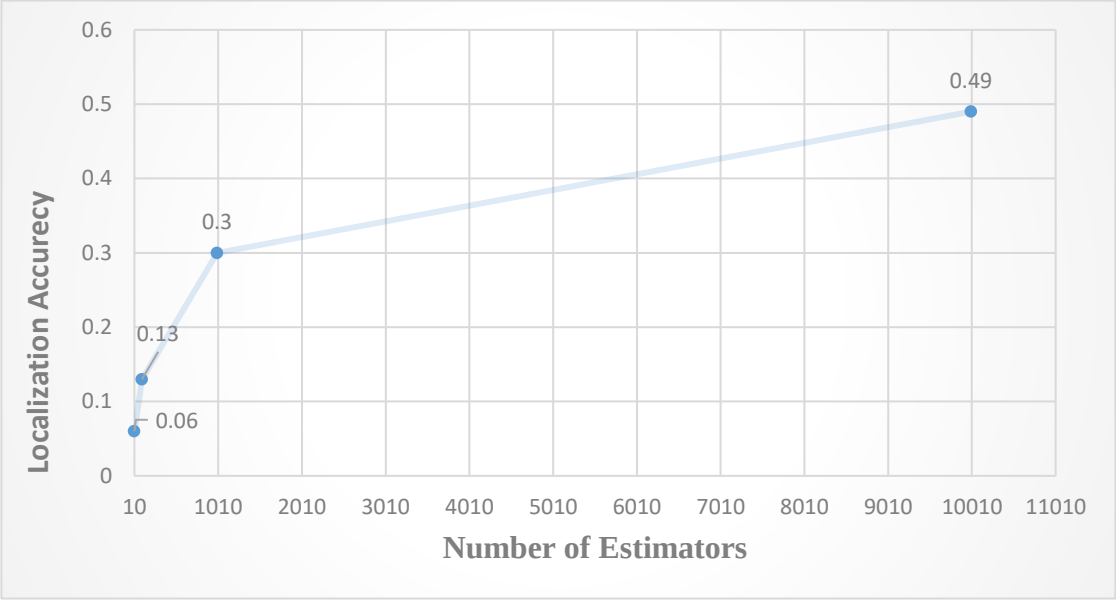


Figure I.3: Random Forest ML number of estimators and its effect on localization accuracy using data generated from 187 different instances of white noise (500ms duration).

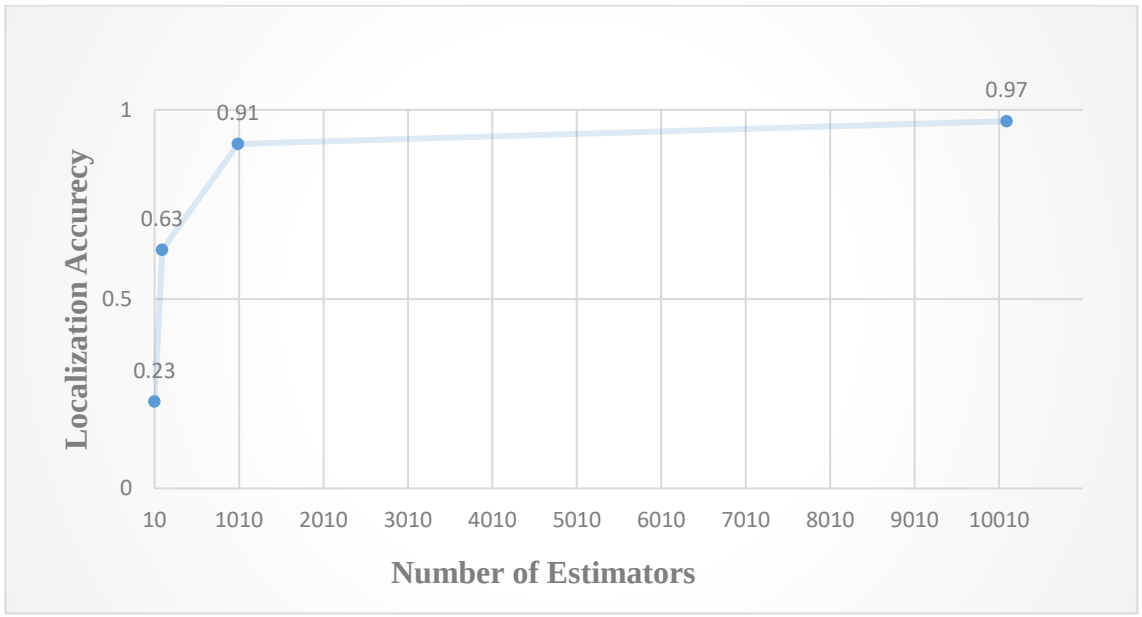


Figure I.4: Random Forest ML number of estimators and its effect on localization accuracy using 187* 20 different instances of white noise (500ms duration).

Appendix II

Additional results from chapter 4 and 5.

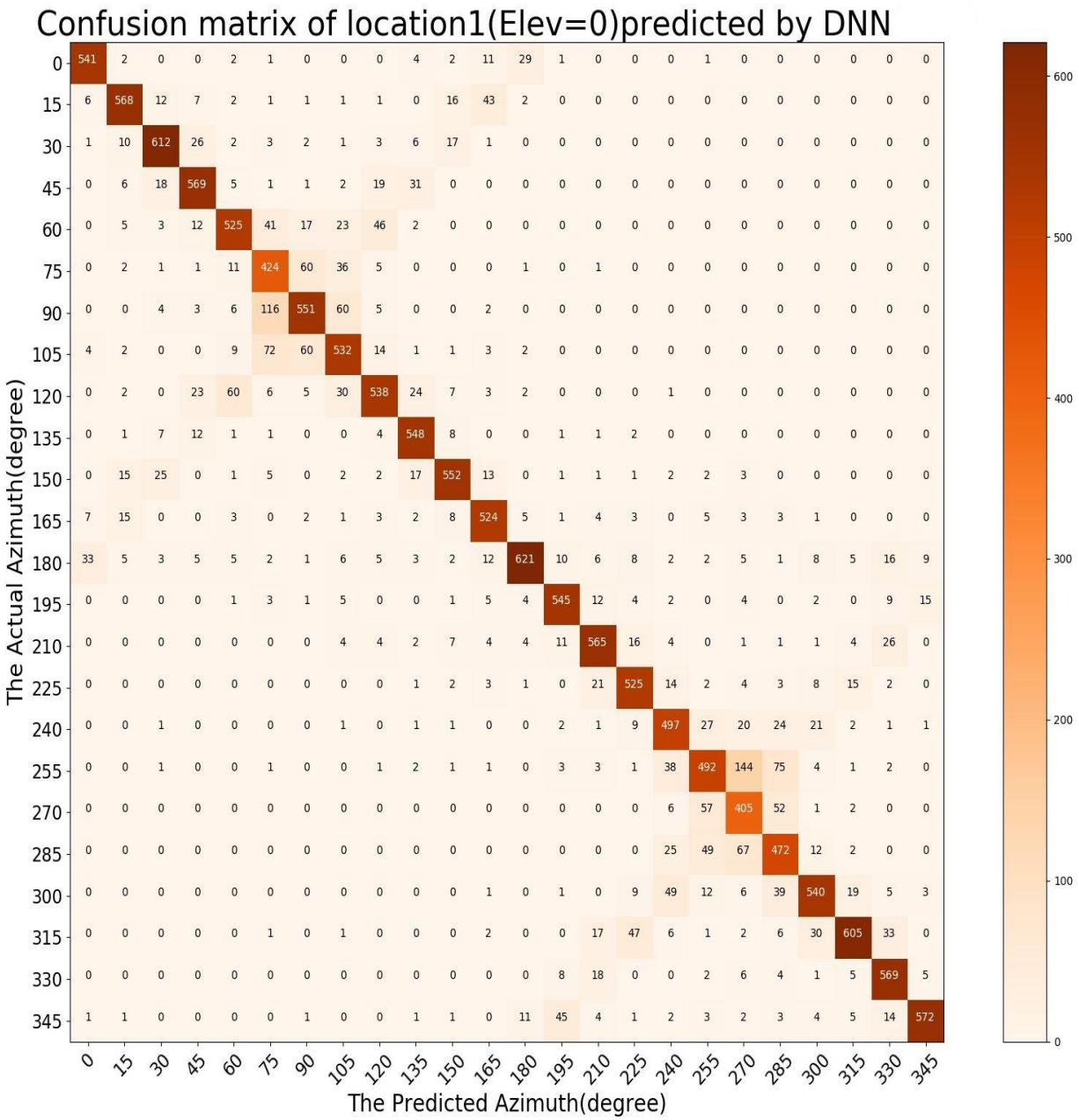


Figure II.1: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation based DNN model with data generated at elevation 0˚ of IRCAM HRTFs with validation speakers.

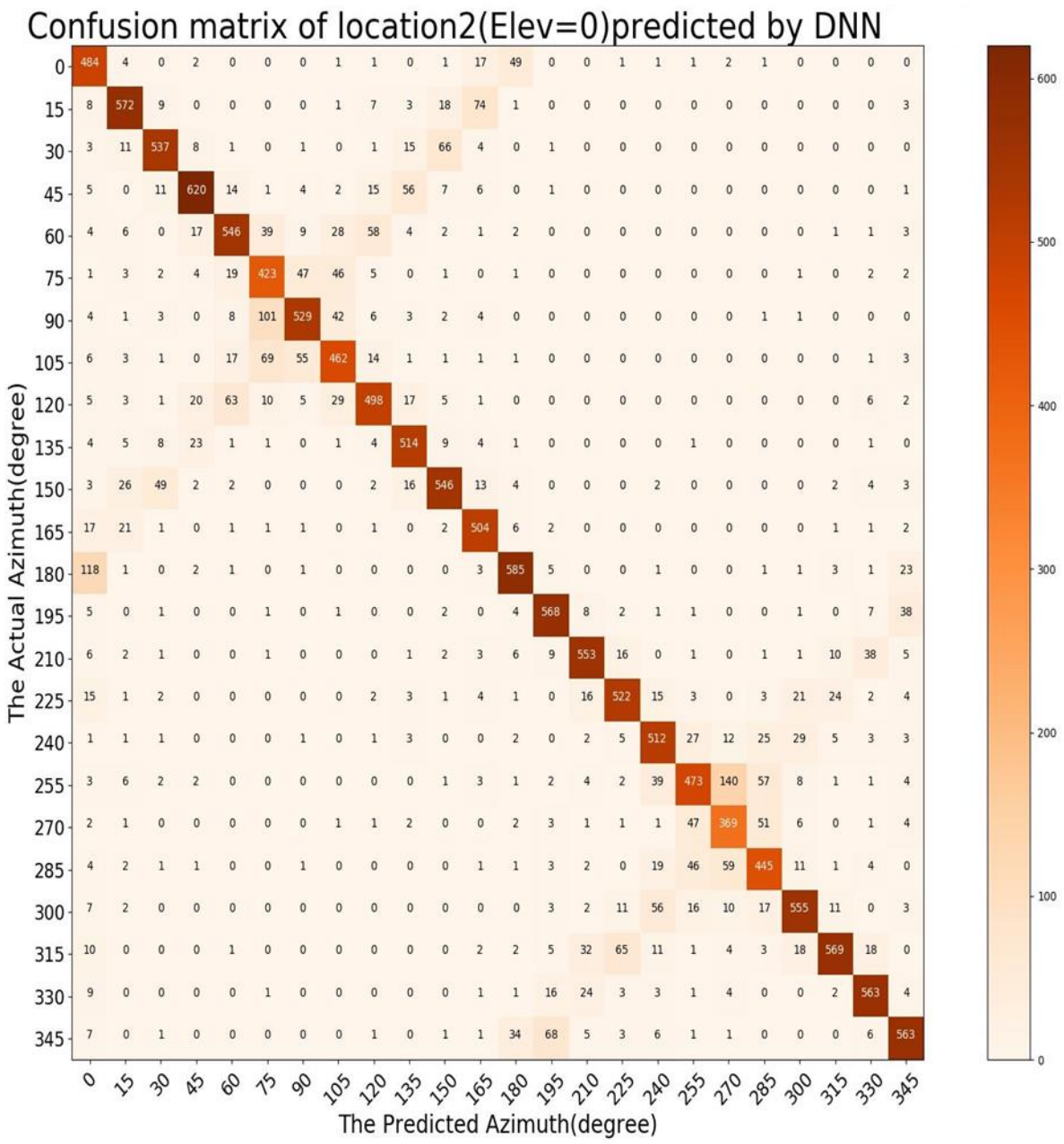


Figure II.2: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation based DNN model with data generated at elevation 0˚ of IRCAM HRTFs with validation speakers.

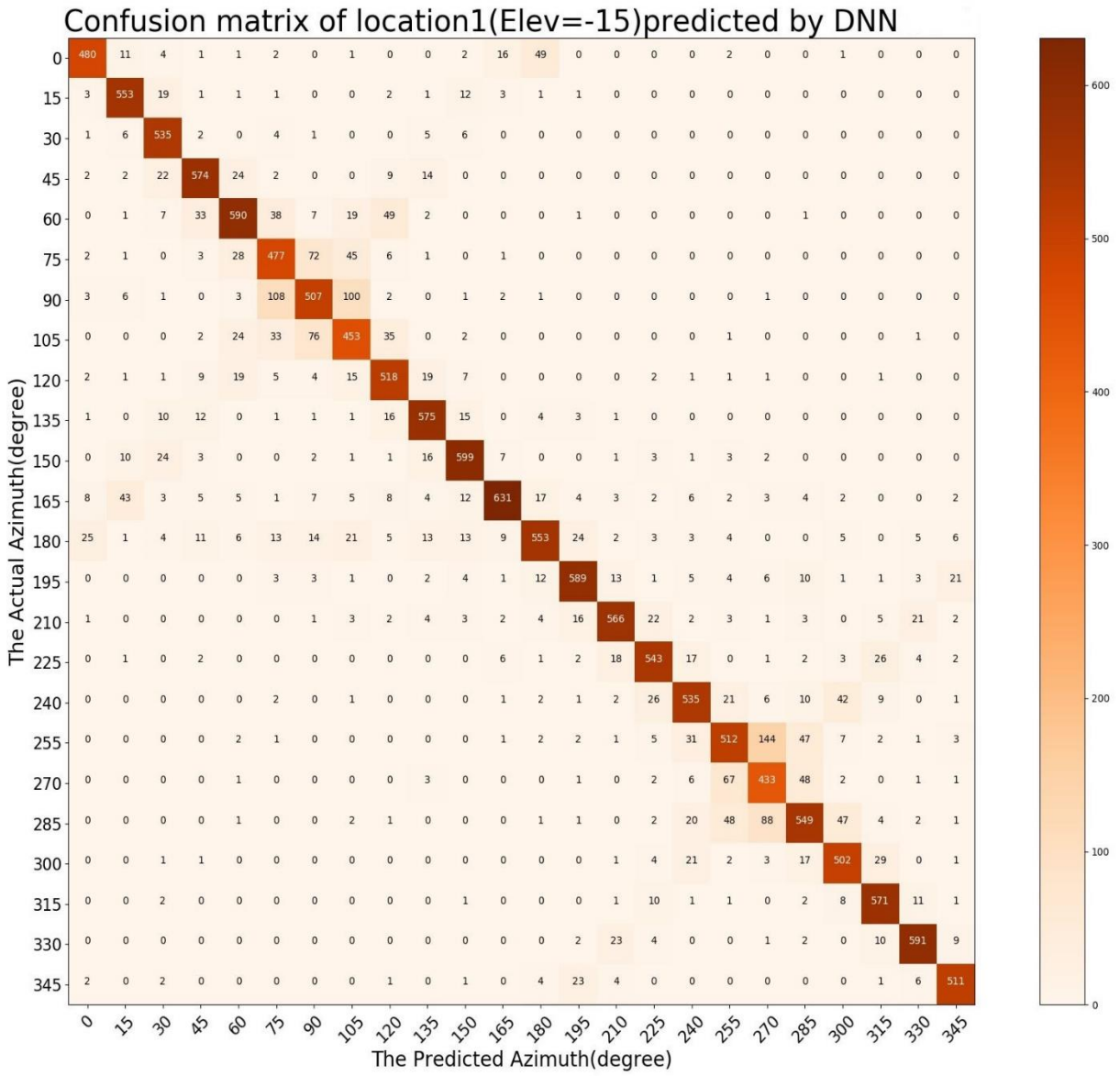


Figure II.3: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -15˚ of IRCAM HRTFs with validation speakers.

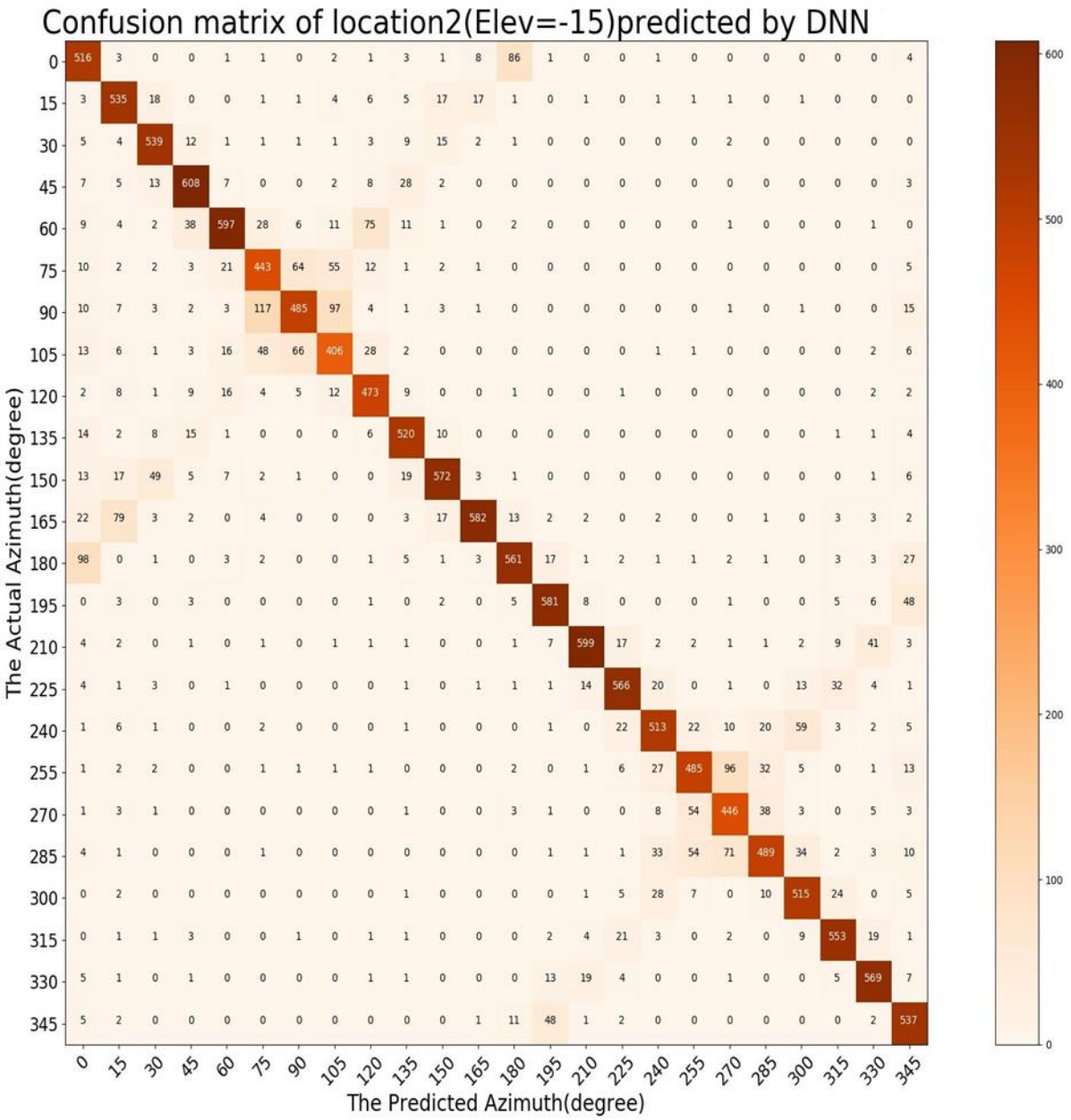


Figure II.4: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -15˚ of IRCAM HRTFs with validation speakers.

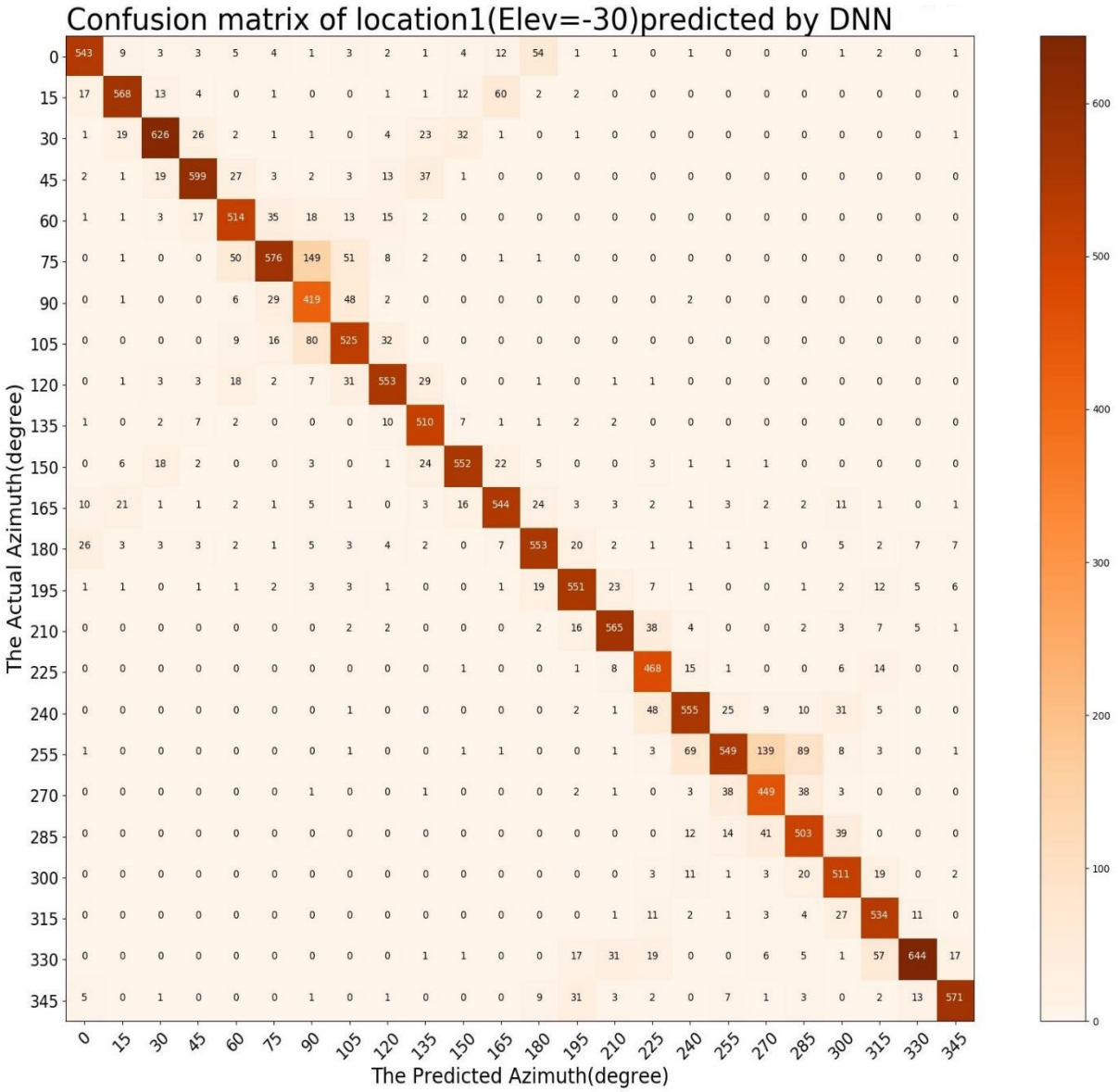


Figure II.5: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -30˚ of IRCAM HRTFs with validation speakers.

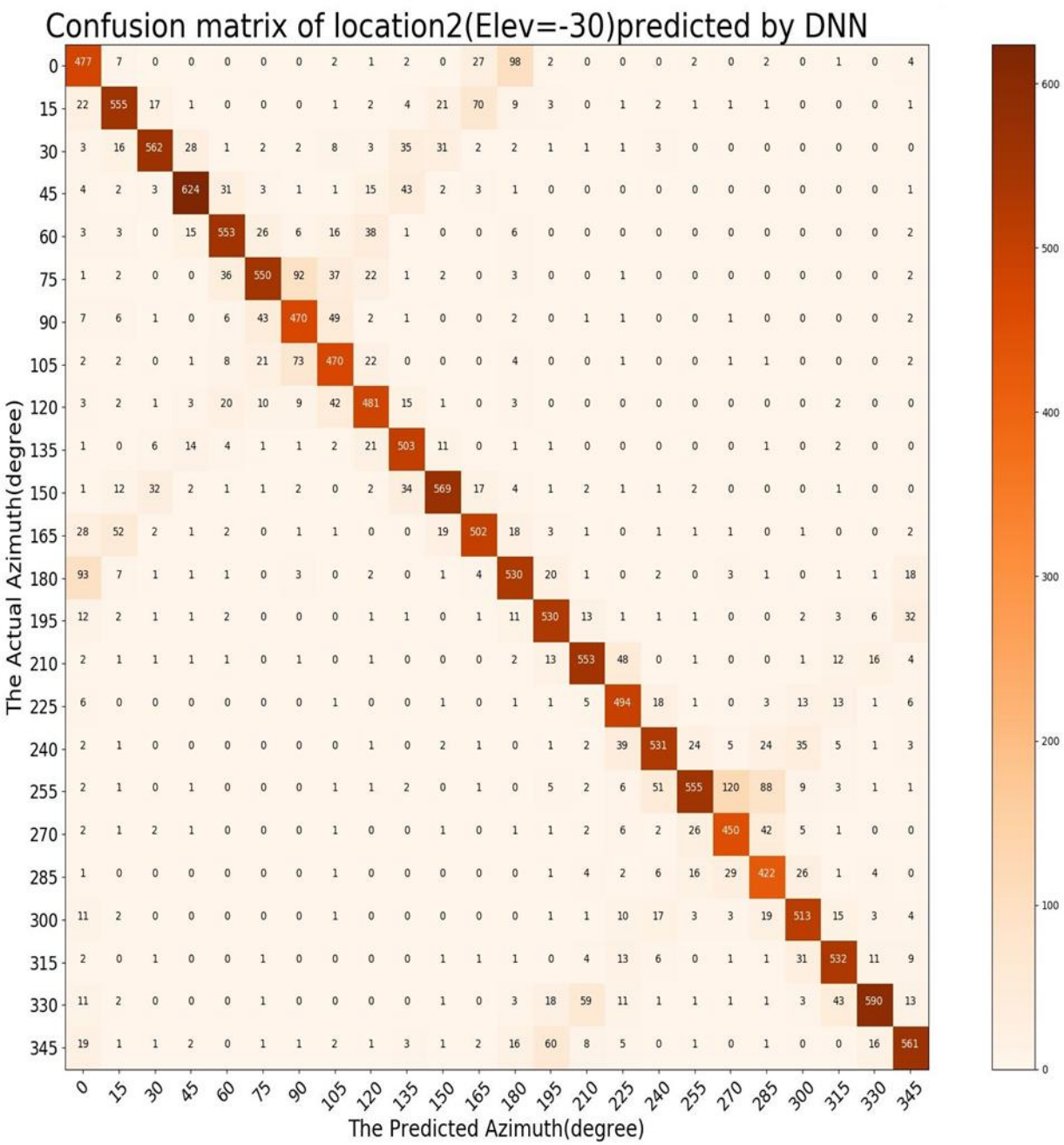


Figure II.6: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -30˚ of IRCAM HRTFs with validation speakers.

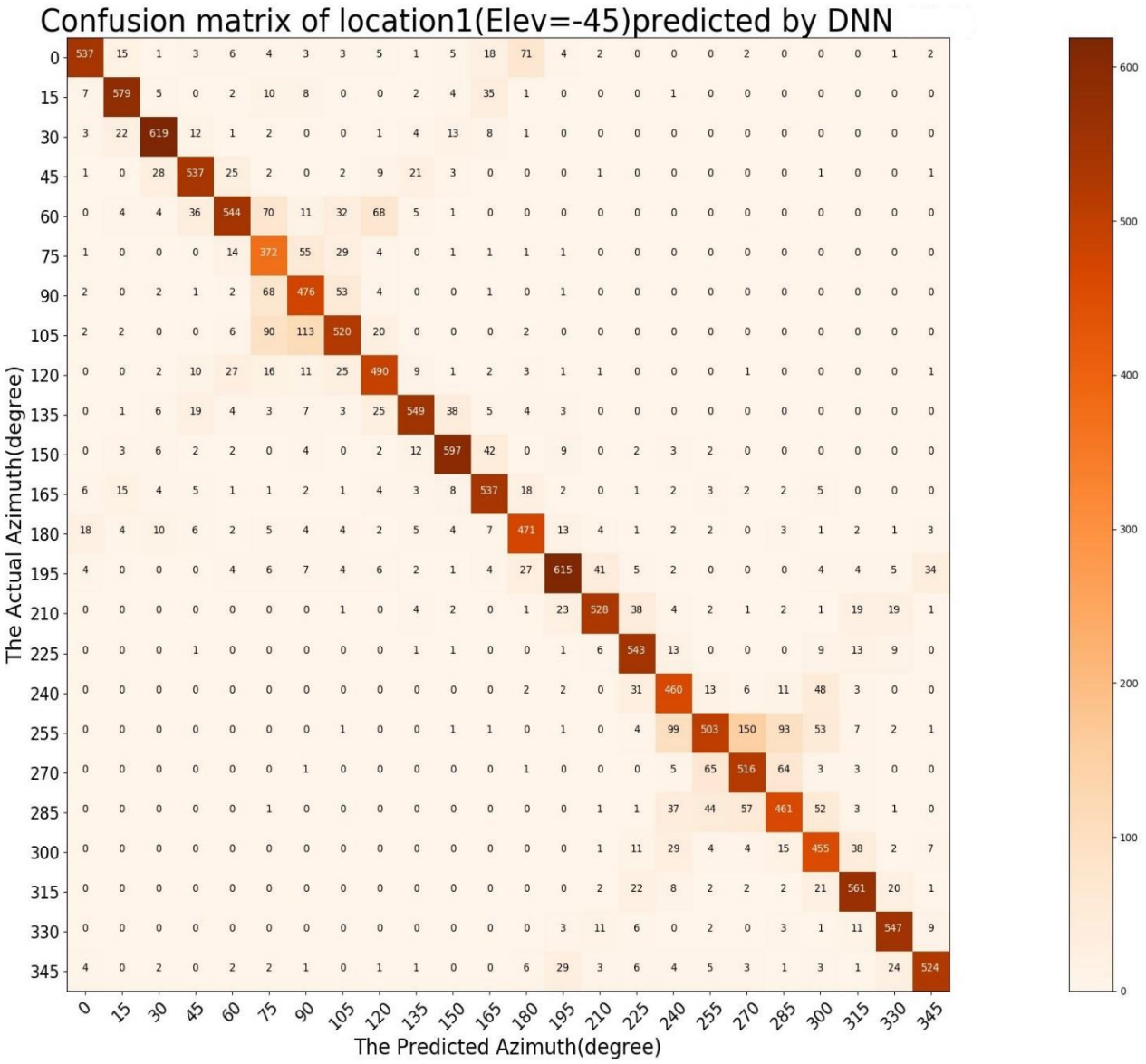


Figure II.7: The confusion matrix plot for the source one azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -45˚ of IRCAM HRTFs with validation speakers.

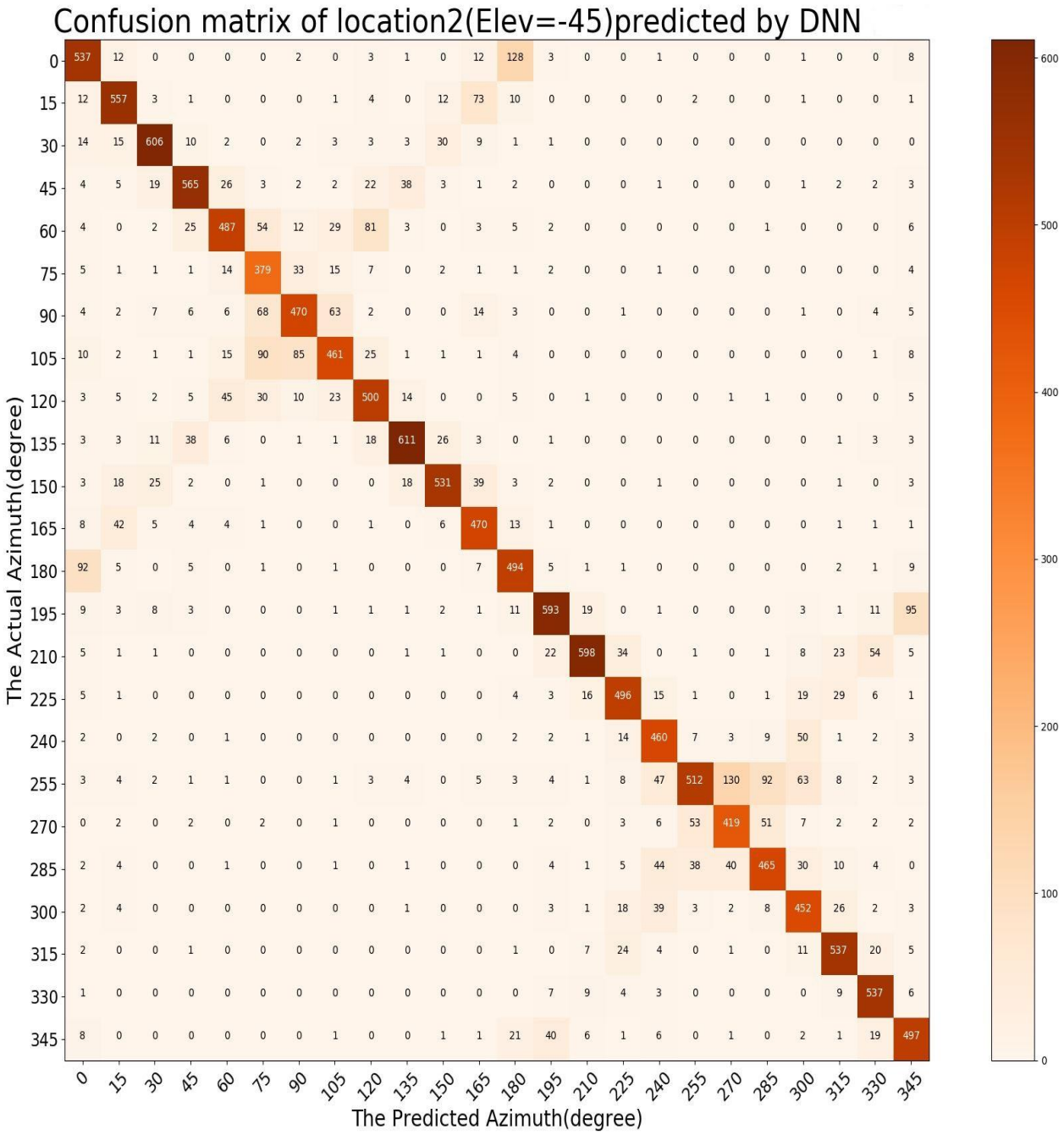


Figure II.8: The confusion matrix plot for the source two azimuth angles predicted by multisource localisation based DNN model with data generated at elevation -45˚ of IRCAM HRTFs with validation speakers.

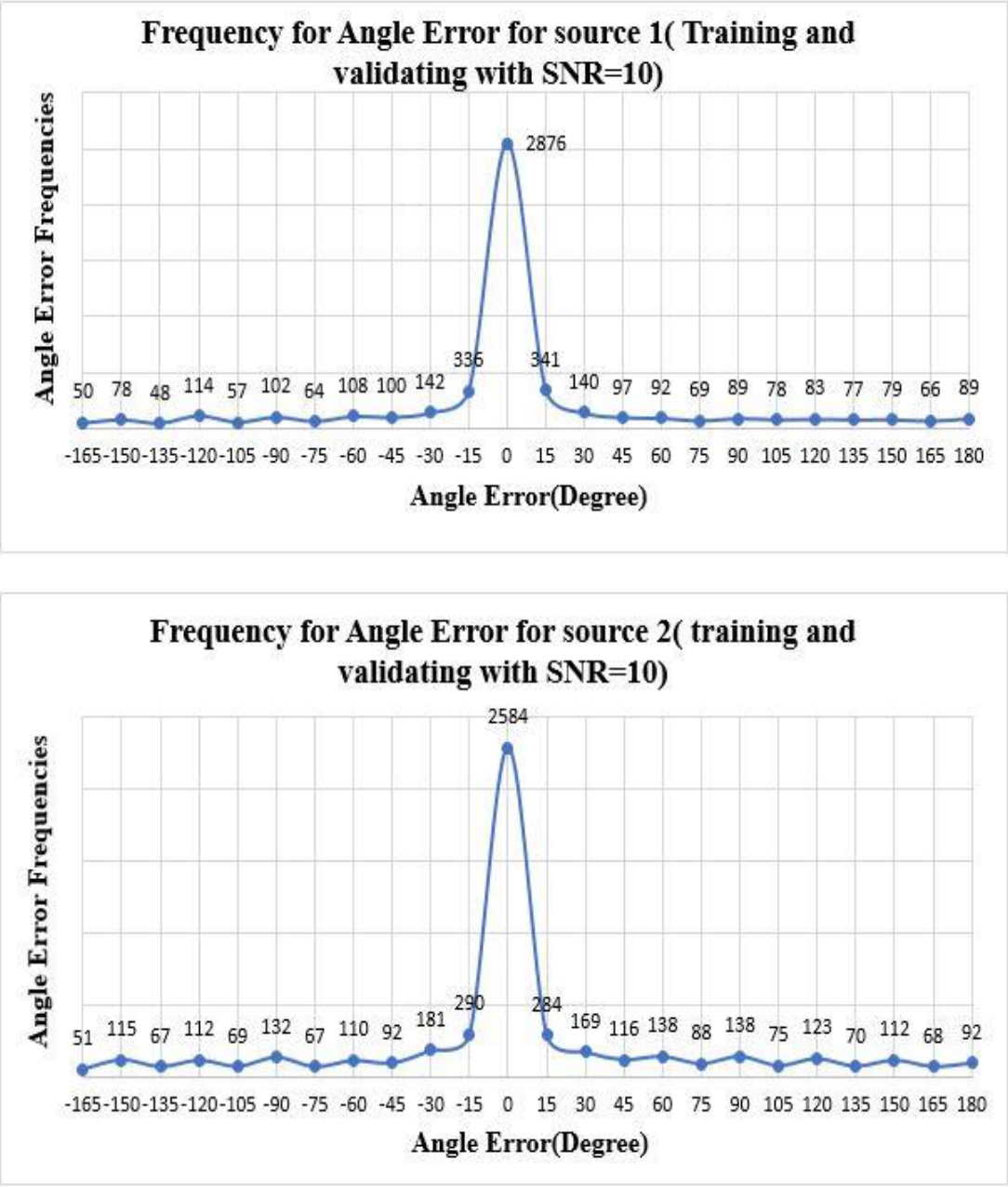


Figure II.9: Angle error frequencies for source one and two predicted by DNN trained and validate in noisy condition at SNR = 10dB.

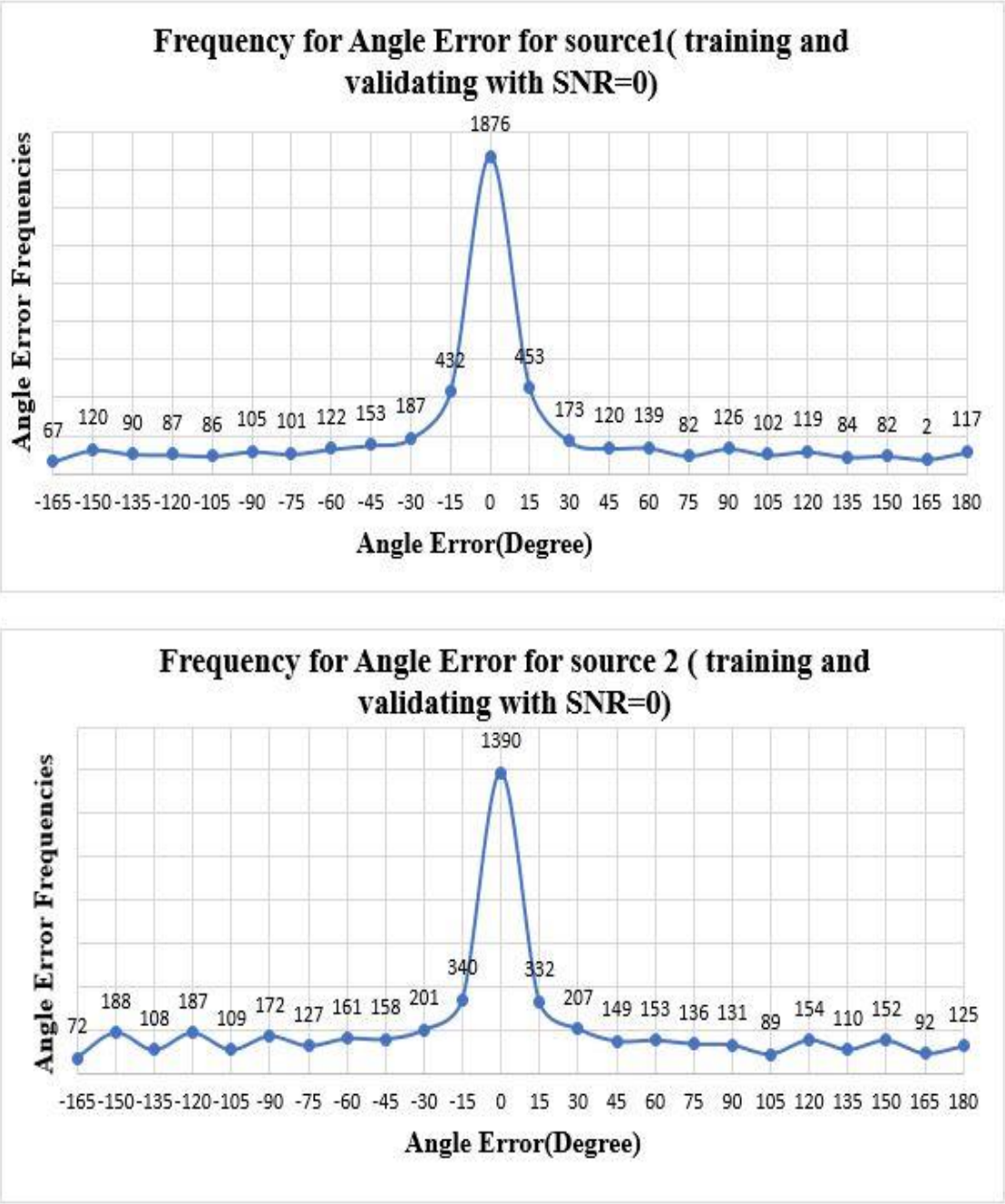
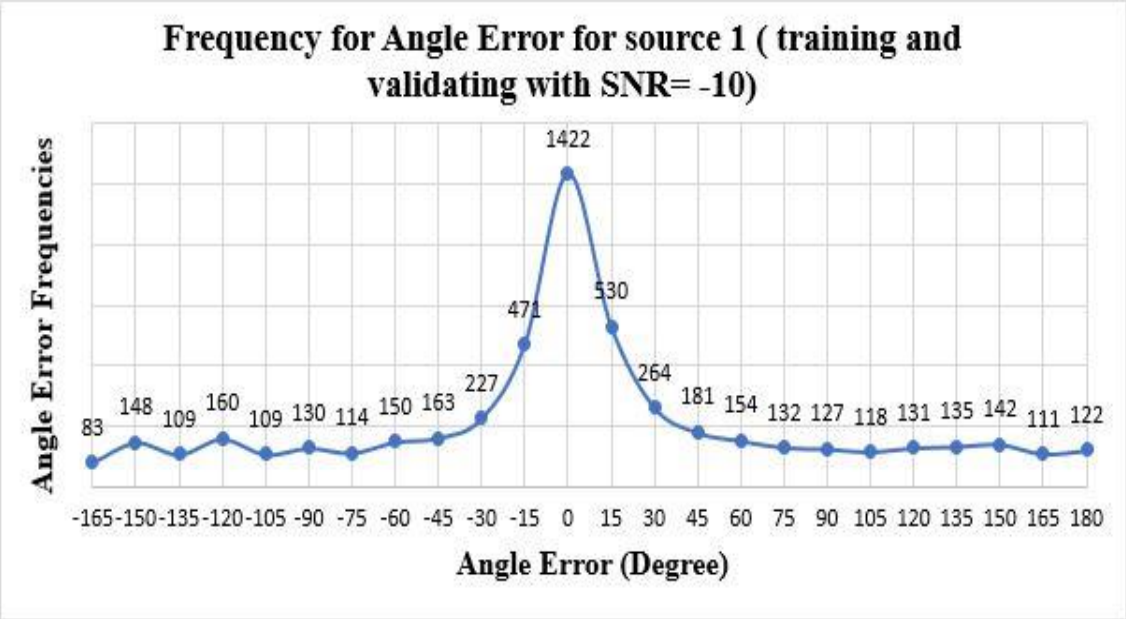


Figure II.10: Angle error frequencies for source one and two predicted by DNN trained and validate in noisy condition at SNR = 0dB.



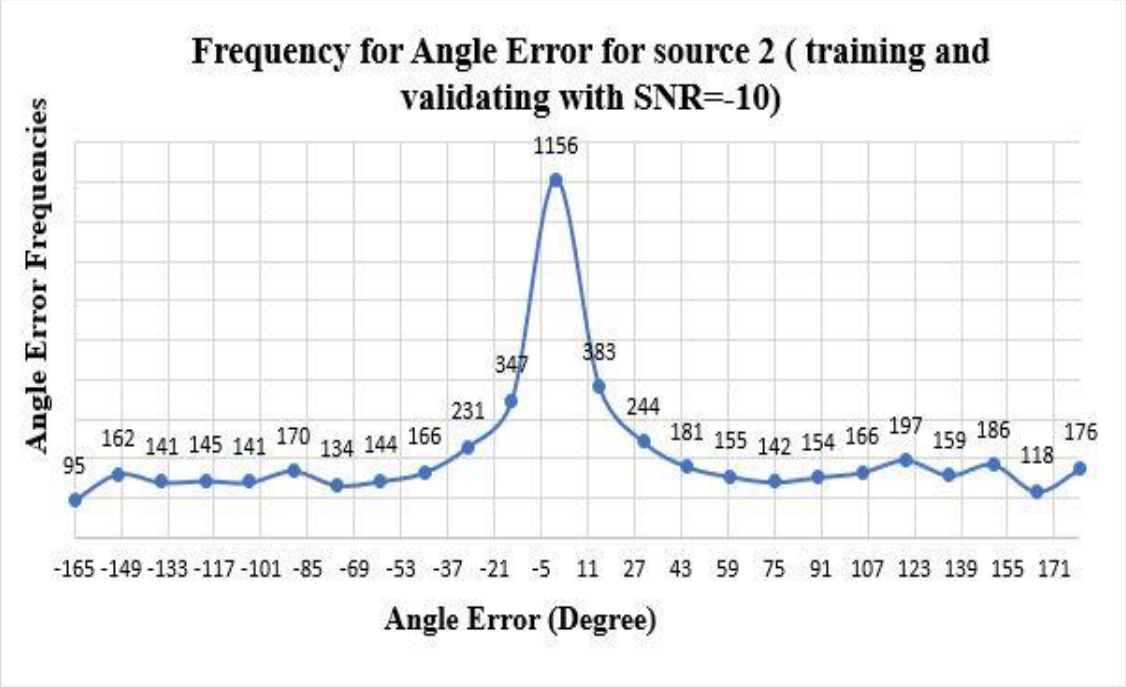


Figure II.11: Angle error frequencies for source one and two predicted by DNN trained and validate in noisy condition at SNR = -10dB.

REFERENCES

REFERENCES

Alberti, P. W. 2001 The anatomy and physiology of the ear and hearing. Occupational exposure to noise: Evaluation, prevention, and control: 53-62.

Agterberg, M. J., et al. 2012 Contribution of monaural and binaural cues to sound localization in listeners with acquired unilateral conductive hearing loss: improved directional hearing with a bone-conduction device. Hearing Research 286(1): 9-18.

Algazi, V. R., Avendano, C., and Duda, R. O. 2001 Elevation localization and head-related transfer function analysis at low frequencies, J. Acoust. Soc. Am. 109, 1110–1122.

Al-Noori, A. 2017 Robust speaker recognition in presence of non-trivial environmental noise (toward greater biometric security), University of Salford.

Al-Noori, Ahmed H; Al-Karawi, Khamis A; Li, Frances. 2015 Improving Robustness of Speaker Recognition in Noisy and Reverberant Conditions Via Training, European Intelligence and Security Informatics Conference, pp.180-180

Alex Graves, Navdeep Jaitly, 2014 Towards End-To-End Speech Recognition with Recurrent Neural Networks. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1764-1772, 2014.

AHVENINEN, J., KOPČO, N. & JÄÄSKELÄINEN, I. P. 2014 Psychophysics and neuronal bases of sound localisation in humans. Hearing research, 307, 86-97.

ANDÉOL, G., MACPHERSON, E. A. & SABIN, A. T. 2013 Sound localization in noise and sensitivity to spectral shape. Hearing Research, 304, 20-27.

ANDERSON, J. A. A. 1988Simple neural network generating an interactive memory. Neurocomputing: foundations of research, MIT Press, 181-192.

Andreopoulou, A. and B. F. Katz, 2015 ON THE USE OF SUBJECTIVE HRTF EVALUATIONS FOR CREATING GLOBAL PERCEPTUAL SIMILARITY METRICS OF ASSESSORS AND ASSESSEES. The 21st International Conference on Auditory Display (ICAD–2015) July 8-10, 2015, Graz, Austria: Areti Andreopoulou.

BALADHANDAPANI, A. & NACHIMUTHU, D. 2015 Evolutionary learning of spiking neural networks towards quantification of 3D MRI brain tumor tissues. Soft Computing - A Fusion of Foundations, Methodologies & Applications, 19, 1803-1816.

Baldi, P. 2012 Autoencoders, unsupervised learning, and deep architectures. Proceedings of ICML workshop on unsupervised and transfer learningJMLR: Workshop and Conference Proceedings 27:37–50, 2012.

Ballard, D. H. 1987 Modular learning in neural networks. Proc. AAAI, pp. 279–284.

BASHEER, I. & HAJMEER, M. 2000 Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods, 43, 3-31.

BEST, V., CARLILE, S., JIN, C. & VAN SCHAIK, A. 2005 The role of high frequencies in speech localization. J. Acoust. Soc. Am, 118, 353-363.
Bechler D, Kroschel K. 2003 Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array. In: Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03), pp 315–318.

Bengio, Y. 2012 Practical recommendations for gradient-based training of deep architectures. Neural networks: Tricks of the trade, Springer: 437-478.

Bengio, Y., et al .2014 Deep generative stochastic networks trainable by backprop .International Conference on Machine Learning.

Berkly, S. 1993Neural Network Model for Sound Localization in Binaural Fusion. In PROCEEDINGS OF THE INTERNATIONAL COMPUTER MUSIC CONFERENCE (pp. 256-256). INTERNATIONAL COMPUTER MUSIC ACCOCIATION.

Blauert J. 1997Spatial Hearing (Revised edition), MIT Press, Cambridge, MA, England

Booij, O. 2004 Temporal pattern classification using spiking neural networks. Unpublished master's thesis, University of Amsterdam

BRAASCH, J. 2005 Modelling of binaural hearing. Communication acoustics. Springer.

Bronkhorst, A. W. 1995 Localization of real and virtual sound sources, J. Acoust. Soc. Am. 98, 2542–2553.

BULANOVA, A., TEMAM, O. & HELIOT, R.2012 Spiking neural networks application to signal processing: observation of dynamical systems.

Buscema, M. 1998 Back propagation neural networks. Substance use & misuse 33(2): 233-270.

CALLISTER, W. D. & RETHWISCH, D. G. 2007Materials science and engineering: an introduction, Wiley New York.

CALMES, L. 2009Biologically inspired binaural sound source localization and tracking for mobile robots. RWTH Aachen University.

CARTY, B. 2010Movements in Binaural Space: Issues in HRTF Interpolation and Reverberation, with applications to Computer Music. National University of Ireland Maynooth.

Chamasemani, F. F. and Y. P. Singh. 2011 Multi-class support vector machine (SVM) classifiers--an application in hypothyroid detection and classification. Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on, IEEE.

Chen, H. and W. Ser. 2009 Acoustic source localization using LS-SVMs without calibration of microphone arrays. Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on, IEEE.

Cheng', C. I. and G. H. Wakefield 1999 Introduction to Head.Related Transfer Functions (HRTFs): Representations of HRTFs in Time, 5026 (I- 1)

Chung, H., et al. 2016 Deep neural network using trainable activation functions. Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE.

Comon P, Jutten C. 2010 Handbook of blind source separation: independent component analysis and applications. Academic Press, Elsevier, Burlington

CUNNINGHAM, B. G. S. and T. STREETER , 2001 Spatial Auditory Display: Comments on Shinn-Cunningham et al., ICAD 2001. ACM Transactions on Applied Perception, Vol. 2, No. 4, October 2005, Pages 426–429. Vol. 2(4, October 2005): Pages 426–429.

Goodman,D.F.M and Brette, R. 2011 Spike-timing-based computation in sound localization. PLoS computational biology 6 (11), e1000993

Goodman, D.F.M and Brette, R. 2010 learning-to-localise-sounds-with-spiking-neural-networks.Advances in Neural Information Processing Systems 23

DAN, Y. & POO, M.-M. 2004 Spike timing-dependent plasticity of neural circuits. Neuron, 44, 23-30.

Datum, M. S., et al. 1996 An artificial neural network for sound localization using binaural cues. The Journal of the Acoustical Society of America 100(1): 372-383.

DASGUPTA, B. & SCHNITGER, G. 1994 The power of approximating: a comparison of activation functions. MATHEMATICAL RESEARCH, 79, 641-641.

Daucé, E. 2014 Toward STDP-based population action in large networks of spiking neurons. ESANN, Citeseer.

DAVIES, S. 2013Learning in spiking neural networks. Citeseer.

Demidova, L., et al. 2016. Big data classification using the SVM classifiers with the modified particle swarm optimization and the SVM ensembles. International Journal of Advanced Computer Science and Applications (IJACSA) 7(5): 294-312.

Deng, L. and Platt,J. C. 2014 Ensemble deep learning for speech recognition. Fifteenth Annual Conference of the International Speech Communication Association.

DHULL, R. A. S. K. 2015 Review on Acoustic Source Localization Techniques Department. European Journal of Advances in Engineering and Technology, 2015, 2(9): 72-77.

DIAZ, C., SANCHEZ, G., DUCHEN, G., NAKANO, M. & PEREZ, H. 2016 An efficient hardware implementation of a novel unary Spiking Neural Network multiplier with variable dendritic delays. Neurocomputing, 189, 130-134.

DONGARE, A., KHARDE, R. & KACHARE, A. D. 2012 Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT), 2, 189-193.

Drullman, R. and A. W. Bronkhorst 2000 Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. The Journal of the Acoustical Society of America 107(4): 2224-2235.

DUDA, R. O. & MARTENS, W. L. 1998 Range dependence of the response of a spherical head model. The Journal of the Acoustical Society of America, 104, 3048-3058.

Evgeniou, T. and M. Pontil. 1999 Support vector machines: Theory and applications. Advanced Course on Artificial Intelligence, Springer.

Ferster, D. and N. Spruston. 1995 Cracking the neuronal code. Science 270(5237): 756.

Gai, Y., et al. 2013 Behavioural and modelling studies of sound localization in cats: effects of stimulus level and duration. Journal of Neurophysiology 110(3): 607-620.

GARDNER, W. G. & MARTIN, K. D. 1995 HRTF measurements of a KEMAR. The Journal of the Acoustical Society of America, 97, 3907-3908.

Gerstner, W., et al. 1998 14 Hebbian Learning of Pulse Timing in the Barn Owl Auditory System.

GERSTNER, W. & KISTLER, W. M. 2002 Spiking neuron models: Single neurons, populations, plasticity, Cambridge university press.

GLACKIN, B., WALL, J. A, MCGINNITY, T. M., MAGUIRE, L. P. & MCDAID, L. J. 2010 A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization. Frontiers in computational neuroscience, 4.

Glasberg, B.R. and Moore, B.C., 1990 Derivation of auditory filter shapes from notched-noise data. Hearing research, 47(1-2), pp.103-138.

Goodfellow, I., et al. 2016 Deep learning, MIT press Cambridge.

Graves, A., et al. 2013 Speech recognition with deep recurrent neural networks. Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, IEEE.

Graves, Alex Navdeep Jaitly. 2014Towards End-To-End Speech Recognition with Recurrent Neural Networks. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1764-1772, 2014.

GROSSBERG, S. 1976 Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological cybernetics, 23, 121-134.

Groethe, M. Pecka, D. McAlpine 2010 Mechanisms of Sound Localization in Mammals, Physiol Rev 90: 983–1012, doi:10.1152/physrev.00026.2009

HAGAN, M. T., DEMUTH, H. B., BEALE, M. H. & DE JESÚS, O. 1996Neural network design, PWS publishing company Boston.

HAO, M., LIN, Z., HONGMEI, H. & ZHENYANG, W. 2007 A novel sound localization method based on head related transfer function. Electronic Measurement and Instruments, 2007. ICEMI'07. 8th International Conference on. IEEE, 4-428-4-432.

Hastie, T. and R. Tibshirani . 1998 Classification by pairwise coupling. Advances in neural information processing systems.

He, K., et al. 2016 Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

Hebrank, J., and Wright, D. 2005 Spectral cues used in the localization of sound sources on the median plane, J. Acoust. Soc. Am. 56, 1829–1834.

HEBB, D. 1949 The Organization of Behavior. New York: John Weley & Sons. Inc.

HEILMANN, D. W. I. G., DOEBLER, D. & BOECK, M. 2014 Exploring the limitations and expectations of sound source localization and visualization techniques. INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Institute of Noise Control Engineering, 4003-4011.

HENDERSON, J. A., GIBSON, T. A. & WILES, J. 2015 Spike Event Based Learning in Neural Networks. arXiv preprint arXiv:1502.05777.

HOPFIELD, J. J. 1982 Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79, 2554-2558.

HORNSTEIN, J., LOPES, M., SANTOS-VICTOR, J. & LACERDA, F. 2006 Sound localization for humanoid robots-building audio-motor maps based on the HRTF. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 1170-1176.

Hsu, C.-W. and C.-J. Lin .2002 A comparison of methods for multiclass support vector machines. IEEE transactions on neural networks 13(2): 415-425.

Ian Goodfellow, et al. 2016Deep Learning. eBook, http://www.deeplearningbook.org/
Inoue J. 2001. Effects of stimulus intensity on sound localization in the horizontal and upper-hemispheric median plane, J UOEH 23: 127–138.

Ishi CT, Chatot O, Ishiguro H, Hagita N 2009 Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. pp 2027–2032

Izhikevich, E. M. 2001 Resonate-and-fire neurons. Neural networks, 14(6), 883-894.

Izhikevich, E. M. 2003 Simple model of spiking neurons. IEEE Transactions on neural

Izhikevich, E. M. 2004 Which model to use for cortical spiking neurons?. IEEE transactions on neural networks, 15(5), 1063-1070.

IZHIKEVICH, E. M. 2006 Polychronization: computation with spikes. Neural computation, 18, 245-282.

JEFFRESS, L. A. 1948 A place theory of sound localization. Journal of comparative and physiological psychology, 41, 35.

Jia, M., et al. 2017Real-time multiple sound source localization and counting using a soundfield microphone." Journal of Ambient Intelligence and Humanized Computing 8(6): 829-844.

JIN, Z. & WANG, D. 2009 A supervised learning approach to monaural segregation of reverberant speech. IEEE Transactions on Audio, Speech, and Language Processing, 17, 625-638.

JINDONG, L., ERWIN, H. & WERMTER, S. 2008 Mobile robot broadband sound localisation using a biologically inspired spiking neural network. 2191-2196.

Joubaud, T., et al. 2017Sound localization models as evaluation tools for tactical communication and protective systems. J Acoust Soc Am 141(4): 2637.

Jürgen Schmidhuber 2015 Deep Learning. Scholarpedia, 10(11):32832. Online, http://www.scholarpedia.org/article/Deep_Learning

KAPRALOS, B., JENKIN, M. & MILIOS, E. 2008. Virtual audio systems. Presence, 17**,** 527-549.

Karhunen, J., et al. 2015 Unsupervised deep learning: A short review. Advances in Independent Component Analysis and Learning Machines, Elsevier: 125-142.

KASINSKI, A. and F. PONULAK 2006 COMPARISON OF SUPERVISED LEARNING METHODS FOR SPIKE TIME coding in spiking neural network .

KERBER, I. S. & SEEBER, I. B. U. 2012 Sound localization in noise by normal-hearing listeners and cochlear implant users. Ear and hearing, 33**,** 445.

Kingma, D. P. and M. Welling 2013 Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kim PJ, Young ED 1994 Comparative analysis of spectro-temporal receptive fields, reverse correlation functions, and frequency tuning curves of auditory nerve fibres. J Acoust Soc Am 95: 410.

Kirk, E. C. and A. D. Gosselin-Ildari 2009 Cochlear labyrinth volume and hearing abilities in primates. The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology 292(6): 765-776.

Knapp C, Carter G. 1976 The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing. 24(4):320-7.

KOHONEN, T. 1972 Correlation matrix memories. IEEE transactions on computers, 100, 353-359.

KRIENER, L. & PFEIL, T. 2014 Binaural Sound Localization in Spiking Neural Networks.
Krizhevsky, A., et al. 2012 Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems.

KRÖSE, B., KROSE, B., VAN DER SMAGT, P. & SMAGT, P. 1993 An introduction to neural networks.

KUEBLER, E. S. & THIVIERGE, J.-P. 2014 Spiking variability: Theory, measures and implementation in MATLAB. Quant. Methods Psychol, 7, 131-142.

Kuhn, G. F. 1977 Model for the interaural time differences in the azimuthal plane, J. Acoust. Soc. Am. 62, 157–167.

KULKARNI, A., ISABELLE, S. & COLBURN, H.1995 On the minimum-phase approximation of head-related transfer functions. Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on, IEEE, 84-87.

Laufer-Goldshtein, B., et al. 2016 Semi-Supervised Sound Source Localization Based on Manifold Regularization. IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP) 24(8): 1393-1407.

Lee, H., et al. 2009 Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th annual international conference on machine learning, ACM.

LeCun, Y., et al. 2015 Deep learning. Nature 521(7553): 436.

Li, X. and H. Liu 2013 Sound source localization for HRI using FOC-based time difference feature and spatial grid matching. IEEE transactions on cybernetics 43(4): 1199-1212.

Lindau, M. A. 2010 ON THE EXTRACTION OF INTERAURAL TIME DIFFERENCES FROM BINAURAL ROOM IMPULSE RESPONSES.
LIPPMANN, R. 1987. An introduction to computing with neural nets. IEEE Assp magazine, 4, 4-22.

Loesch, B, Uhlich S, Yang B .2009 Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform. In: IEEE/SP 15th Workshop on Statistical Signal Processing, 2009. SSP '09. pp 677–680

Lombard A, Zheng Y, Buchner H, Kellermann W. 2011. TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis.

MA, C., QI, J., LI, D. & LIU, R. 2015 Improving bottleneck features for automatic speech recognition using gammatone-based cochleagram and sparsity regularization. Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, 2015. IEEE, 63-67.

Macpherson E.A., Middlebrooks J. C. 2000 Localization of brief sounds: effects of level and background noise, J Acoust Soc Am 108: 1834–1849.

Macpherson E.A., Middlebrooks J. C. 2002Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited.The Journal of the Acoustical Society of America 111(5 Pt 1):2219-36

Macpherson, E. A., and Middlebrooks, J. C. 2002 Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited, J. Acoust. Soc. Am. 111, 2219–2236.

MARKOWSKA-KACZMAR, U. & KOLDOWSKI, M. 2015 Spiking neural network vs multilayer perceptron: who is the winner in the racing car computer game. Soft Computing - A Fusion of Foundations, Methodologies & Applications, 19, 3465-3478.

MARIAN, I., REILLY, R. & MACKEY, D, 2002 Efficient event-driven simulation of spiking neural networks. Proceedings of the 3rd WSEAS international conference on neural networks and applications. MIT Press Cambridge, MA.

Maroonroge, S., et al. 2000 Basic anatomy of the hearing system. Helmet-Mounted Displays: Sensation, Perception and Cognition Issues. Fort Rucker, Alabama: US Army Aeromedical Research Laboratory.306-279 :

Masquelier, T. and S. J. Thorpe 2005 Unsupervised learning of visual features through Spike Timing Dependent Plasticity. PLoS Computational Biology preprint (2007): e31.

Ma, C., et al. 2015 Improving bottleneck features for automatic speech recognition using gammatone-based cochleagram and sparsity regularization. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific, IEEE.

MAY, T., VAN DE PAR, S. & KOHLRAUSCH, A. 2011 A probabilistic model for robust localization based on a binaural auditory front-end. Ieee transactions on audio, speech, and language processing, 19, 1-13.

MCCULLOCH, W. S. & PITTS, W. 1943 A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5, 115-133.

MENDONÇA, C., et al. 2014 ADAPTATION TO NON-INDIVIDUALIZED SPATIAL SOUND THROUGHAUDIOVISUAL EXPERIENCE." AES 55th International Conference, Helsinki, Finland, 2014 August 27–29.

Middlebrooks, J. C. 1999 Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. The Journal of the Acoustical Society of America 106(3): 1493-1510.

MILLER, D. A. 2013 Modeling HRTF For Sound Localization in Normal Listeners and Bilateral Cochlear Implant Users.

Michael Nielsen. 2017 Neural Networks and Deep Learning., eBook, Chapter2, http://neuralnetworksanddeeplearning.com/index.html

Mokri, Y., et al. 2015 Effect of background noise on neuronal coding of interaural level difference cues in rat inferior colliculus. European Journal of Neuroscience 42(1): 1685-1704.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. 1995Head-related transfer functions of human subjects,J. Audio Eng. Soc.43, 300–321.

MURRAY, J. C., ERWIN, H. & WERMTER, S.2004 Robotics sound-source localization and tracking using interaural time difference and cross-correlation. AI Workshop on NeuroBotics.

MULANSKY, M., BOZANIC, N., SBURLEA, A. & KREUZ, T.2015 A guide to time-resolved and parameter-free measures of spike train synchrony. Event-based Control, Communication, and Signal Processing (EBCCSP), International Conference on, 2015. IEEE, 1-8.

NAM, J., KOLAR, M. A. & ABEL, J. S.2008 On the minimum-phase nature of head-related transfer functions. Audio Engineering Society Convention 125. Audio Engineering Society.

Nesta F, Omologo M. 2012 Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. IEEE Trans Audio Speech Lang Process 20(1):246–260

NILSSON, N. J. 1996 Introduction to machine learning. An early draft of a proposed textbook.

Nikoskinen, T. 2015 From neural networks to deep neural networks. Aalto University School of Science.

NIMITYONGSKUL, S. & KAMMER, D. 2009Frequency response based sensor placement for the mid-frequency range. Mechanical Systems and Signal Processing, 23, 1169-1179.

O'CONNOR, P. 2012 A real-time sensory-fusion model using a Deep beilive network with spiking neurons.

O'CONNOR, P., NEIL, D., LIU, S.-C., DELBRUCK, T. & PFEIFFER, M. 2015 Real-time classification and sensor fusion with a spiking deep belief network. Neuromorphic Eng. Syst. Appl, 61, 1-10.

Oppenheim, A. V. and R. W. Schafer 2014 Discrete-time signal processing, Pearson Education.

Parseihian, G. and B. F. Katz. 2012 Rapid head-related transfer function adaptation using a virtual auditory environment. The Journal of the Acoustical Society of America 131(4): 2948-2957.

PAUGAM-MOISY, H. & BOHTE, S. 2012 Computing with spiking neuron networks. Handbook of natural computing. Springer.

PAVLIDI, D., PUIGT, M., GRIFFIN, A. & MOUCHTARIS,2012A Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. ICASSP, 2625-2628.

Pavlidi D, Griffin A, Puigt M, Mouchtaris A. 2013 Real-time multiple sound source localization and counting using a circular microphone array. IEEE Trans Audio Speech Lang Process 21(10):2193–2206

Peterson, B., et al. 1996 Spectrum analysis-amplitude and frequency modulation. Hewlett Packard.

PIETILA, G. & LIM, T. C. 2012 Intelligent systems approaches to product sound quality evaluations–A review. Applied Acoustics, 73, 987-1002.

Pillay, N. and P. Govender. 2017 Multi-Class SVMs for Automatic Performance Classification of Closed Loop Controllers. CONTROL ENGINEERING AND APPLIED INFORMATIC 19(3):3-12.

POURMOHAMMAD, A. & AHADI, S. M. 2013 N-dimensional N-microphone sound source localization. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 1-19.

QIAN, J. & EDDINS, D. A. 2008 The role of spectral modulation cues in virtual sound localizationa). The Journal of the Acoustical Society of America, 123, 302-314.

RAKERD, B., HARTMANN, W. M. & MCCASKEY, T. L. 1999 Identification and localization of sound sources in the median sagittal plane. The Journal of the Acoustical Society of America, 106, 2812-2820.

RANGANATHAN, A. & KIRA, Z. 2003 Self-organization in artificial intelligence and the brain. College of Computing, Georgia Institute of Technology.

Recio-Spinoso, A. and N. P. Cooper, 2013 Masking of sounds by a background noise - cochlear mechanical correlates. Journal of Physiology 591(10): 2705-2721.

ROMAN, N. & WANG, D. 2008 Binaural tracking of multiple moving sources. IEEE Transactions on Audio, Speech, and Language Processing, 16, 728-739.

ROSENBLATT, F. 1958 The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65, 386.

Sadowski, P. 2016 Notes on backpropagation. homepage: https://www. ics. uci. edu/~ pjsadows/notes. pdf (online).

Salvati, D. 2012 Acoustic source localization using microphone arrays.

Schmidhuber, J. 2015 Deep learning in neural networks: An overview. Neural Networks 61: 85-117.

Schmidt R. 1986 Multiple emitter location and signal parameter estimation. IEEE Trans Antennas Propag 34(3):276–280

Scott, N. M. 2015 Evolving Spiking Neural Networks for Spatio-and Spectro-Temporal Data Analysis: Models, Implementations, Applications, Auckland University of Technology.

SHEAFFER, J. 2013From source to brain: modelling sound propagation and localisation in rooms.Citeseer.

SHIMOYAMA, R.2012 Bio-inspired sound source localization compensated for sound diffraction by binaural head and torso. Computational Intelligence and Cybernetics (CyberneticsCom), IEEE International Conference on, 2012. IEEE, 79-82.

Shiiki Y, Suyama K. 2015 Omnidirectional sound source tracking based on sequential updating histogram. In: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp 1249–1256

Shoko, A., et al. 2007 Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal processing 87(8): 1833-1847.

Singh, L., Chetty, G. and Singh, S., 2012 A novel algorithm using MFCC and ERB gammatone filters in speech recognition. Journal of Information Systems and Communication, 3(1), p.358.

Slaney, M., 1993 An efficient implementation of the Patterson-Holdsworth auditory filter bank. Apple Computer, Perception Group, Tech. Rep, 35, p.8.

Smolensky, P. 1986 Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281. MIT Press.

Smola, A. J. and B. Schölkopf .2004 A tutorial on support vector regression. Statistics and computing 14(3): 199-222.

SO, R., LEUNG, N., BRAASCH, J. & LEUNG, K. 2006a A low cost, non-individualized surround sound system based upon head related transfer functions: An ergonomics study and prototype development. Applied ergonomics, 37, 695-707.

SONG, T., QU, T., WU, X. & CHEN, J. 2016 An artificial neural network model for predicting sound direction in different acoustic environments.

Song, P., et al. 2017 Acoustic source localization using 10-microphone array based on wireless sensor network. Sensors and Actuators A: Physical 267: 376-384.

STROMATIAS, E. 2011 Developing a supervised training algorithm for limited precision feed-forward spiking neural networks. arXiv preprint arXiv:1109.2788.

STROMATIAS, E. & MARSLAND, J. S. 2015 Supervised learning in Spiking Neural Networks with limited precision: SNN/LP. Neural Networks (IJCNN), International Joint Conference on, 2015. IEEE, 1-7.

Sun, Y., et al. 2018 Indoor Sound Source Localization with Probabilistic Neural Network. IEEE Transactions on Industrial Electronics 65(8): 6403-6413.

Swartling M, Sllberg B, Grbi N, 2011. Source localization for multiple speech sources using low complexity non-parametric source separation and clustering. Signal Process 91(8):1781–

TADDESE, B. T. 2006 Sound Source Localization and Separation.

Takeda, R. and K. Komatani 2016a Sound source localization based on deep neural networks with directional activate function exploiting phase information. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE.

Takeda, R. and K. Komatani, 2016b Discriminative multiple sound source localization based on deep neural networks using independent location model. Spoken Language Technology Workshop (SLT), 2016 IEEE, IEEE.

TALAGALA, D. S., ZHANG, W., ABHAYAPALA, T. D. & KAMINENI, A. 2014 Binaural sound source localization using the frequency diversity of the head-related transfer function. The Journal of the Acoustical Society of America, 135, 1207-1217.

VALIN, J.-M., MICHAUD, F., ROUAT, J. & LÉTOURNEAU, D.2003 Robust sound source localization using a microphone array on a mobile robot. Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference , IEEE, 1228-1233.

VAN OPSTAL, J. 2016The Auditory System and Human Sound-localization Behavior, Academic Press.

Vapnik, V. 1995 The natural of statistical theory, New York: Springer-Verlag.

Vapnik, V. 1998 Statistic learning theory. Willey, New York.

VREEKEN, J. 2002 Spiking neural networks, an introduction. Institute for Information and Computing Sciences, Utrecht University Technical Report UU-CS-2003-008.

WALL, J. A., MCDAID, L. J., MAGUIRE, L. P. & MCGINNITY, T. M. 2012 Spiking neural network model of sound localization using the interaural intensity difference. IEEE transactions on neural networks and learning systems, 23, 574-586.

WALLACH, H. 1940 The role of head movements and vestibular and visual cues in sound localization. Journal of Experimental Psychology, 27, 339.

WANG, D. & BROWN, G. J. 2006Computational auditory scene analysis: Principles, algorithms, and applications, Wiley-IEEE Press.

WANG, H. & KAVEH, M. 1985 Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33, 823-831.

WARD, D. B., DING, Z. & KENNEDY, R. A. 1998 Broadband DOA estimation using frequency invariant beamforming. IEEE Transactions on Signal Processing, 46, 1463-1469.

Wenzel, E. M., et al. 1993 Localization using nonindividualized head-related transfer functions. The Journal of the Acoustical Society of America 94(1): 111-123.

WIDROW, B. & HOFF, M. E. 1960Adaptive switching circuits. IRE WESCON convention record,. New York, 96-104.

Wightman, F. L., and Kistler, D. J. 1989 Headphone simulation of free field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. 85, 858–867.

WOODRUFF, J. & WANG, D. 2012 Binaural localization of multiple sources in reverberant and noisy environments. IEEE Transactions on Audio, Speech, and Language Processing, 20, 1503-1512.

Xiao, X., et al. 2015 A learning-based approach to direction of arrival estimation in noisy and reverberant environments. Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE.

XIE, B. 2013. Head-related transfer function and virtual auditory display, J. Ross Publishing.
Yalta, K. N., T Ogata ,2017 Sound Source Localization using Deep Learning Models. Journal of Robotics and Mechatronics Vol.29 No.1: 37-48.

YOUSSEF, K., ARGENTIERI, S. & ZARADER, J.L.2012 A binaural sound source localization method using auditive cues and vision. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012. IEEE, 217-220.

Yu Q, Yan R, Tang H, Tan KC, Li H. 2016 A spiking neural network system for robust sequence recognition. IEEE transactions on neural networks and learning systems. 2016 Mar; 27(3):621-35.

Zemlin, W. R. 1968 Speech and Hearing Science, Anatomy and Physiology.

Zhang, J., et al. 2014 Dependency of the Finite-Impulse-Response-Based Head-Related Impulse Response Model on Filter Order. 10.14279/depositonce-4103.

Zhao, L. and L. Zhaoping 2011 Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles." PLoS Computational Biology 7(8): e1002123.

ZHANG, X.-S. 2000 Introduction to artificial neural network. Neural Networks in Optimization. Springer.

Zhong, X.-l. and B.-s. Xie , 2014 Head-related transfer functions and virtual auditory display. Soundscape Semiotics-Localization and Categorization, InTech.

ZIEGELWANGER, H., MAJDAK, P. & KREUZER, W. 2015 Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization. The Journal of the Acoustical Society of America, 138, 208-222.

Ziegelwanger, H., and Majdak, P. 2014 Modelling the direction continuous time-of-arrival in head-related transfer functions, J. Acoust. Soc. Am. 135, 1278–1293.

ZOTKIN, D., HWANG, J., DURAISWAINI, R. & DAVIS, L. S.2003 HRTF personalization using anthropometric measurements. Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on., 2003. Ieee, 157-160.