Binaural sound source localization based on GASSOM and DNN

by

Shutao CHEN

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Industrial Engineering & Logistics Management

Mar 2022, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Shutao CHEN

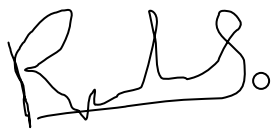
9 Mar 2022

Binaural sound source localization based on GASSOM and DNN

by

Shutao CHEN

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the PhD qualifying examination committee have been made.



Prof. Richard H. Y. SO, Thesis Supervisor

Prof. Jiheng Zhang, Head

Department of Industrial Engineering & Decision Analytics

9 Mar 2022

Acknowledgments

I am deep grateful to all companions that have helped and encouraged me during my PhD studies. Their continuous support gives me confidence and courage to face lots of challenges in discovering new areas and to complete this thesis.

In the first place, I would like to express my sincere thankfulness to my supervisors, Prof. Richard H. Y. So and his guidance and encouragements on my research journey. He is always willing to discuss with me patiently and share their insights and experience with me. I have learned a lot from him, not only the knowledge, but also the rigorous working attitude, and even the patience, the persistence in conducting experiment. I would also thank all the friends in my research group. We have many useful discussions on each other's research topics, which provides me new ideas and helps me to go through difficulties.

At the same time, I am also grateful to Prof. Bertram Shi and the students in his group. They gave me lots of help on the research work and I learned from them a lot.

Besides, I would also like to thank Prof. Xiangtong Qi, Prof. Allen Wu, Prof. Felix Chen for serving as my committee members and for their invaluable recommendations. I also want to thank Prof. Bertram Shi and Prof. Xiangtong QI for serving on the committee of my preliminary dissertation, and providing many suggestion for improving my thesis. Also thank other professor and staffs in the IEDA department that offer excellent courses and assistance for my study and life in HKUST.

I would never forget my dear friends in HKUST. They always believe in me and show their care and warmness when I come up with difficulties. We are colleagues, cooperators, friends and even like family members. I cannot image how I can overcome tough times without their accompany.

Finally and importantly, I would like to present my deepest appreciation and gratitude to my parents for their unconditional love and support. I also wanna say thank you to my gf Ms Ya Wang for her encouragement and love, it's my great fortune to own her accompany in my hard time.

Contents

List of Figures

List of Tables

Binaural sound source localization based on GASSOM and DNN

by Shutao CHEN
Department of Industrial Engineering & Decision Analytics
The Hong Kong University of Science and Technology

Abstract

Humans can localize sound source(s) with two ears - binaural sound localization. Conventional methods to model binaural localization focused on artificial spatial cues such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) to decode the locational information. In this work, we extracted spatial features with sparse coding algorithms and further mapped the features to predict sound locations with Deep Neural Network (DNN). The use of GASSOM (Generative Adaptive Subspace Self-organizing Map) and Independent Component Analysis (ICA) as the sparse coding algorithms were compared. Results indicate that GASSOM outperforms ICA. Map size and basis function length have been shown to affect the performance of GASSOM and the optimal selections of both parameters are reported in the thesis. In order to verify the ability of GASSOM-DNN sound localization model to simulate human binaural localization performance, benchmark studies with past reported empirical data were conducted. Factors investigated included: the influence of bandwidth, center frequency and duration of binaural cues; and the mismatch of non-individualized HRTFs. Performance of computational model was compared with previously reported human data and similarity was achieved. Future potentials on the use of GASSOM to model binaural sound localization are discussed.

Chapter 1

Introduction

1.1 Summary

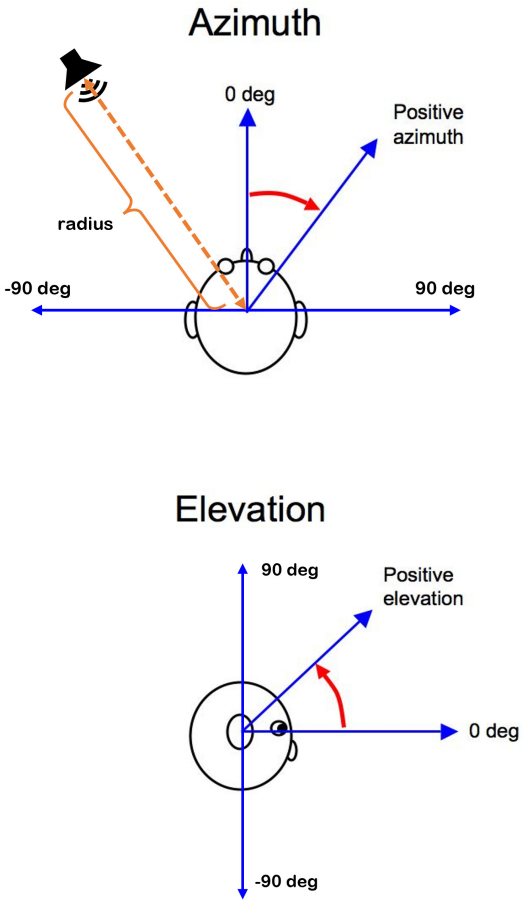
In chapter 1, we will give a brief introduction about the background of binaural sound localization, including the definition of binaural sound localization, the mechanism that human listeners determine the localization of a spatial sound source, the binaural sound construction techniques, and the Head-Related Transfer Function (HRTF) as well as the virtual auditory space (VAS).

1.2 Background introduction and motivations

Sound localization is an essential ability of mammalian animals for survival. Some animals need to track the location of prey for hunting (e.g., some echo-location bats), or recognize the presence of predators to escape from the dangers. For vertebrates, this process is always based on the audio signals received by the two ears. Compared with most computational models which utilize several microphones (microphone array technology), animals are still capable of localize the location of sound accurately. Inspired by this phenomenon, we decide to build a computational model that simulate the auditory system of human being. An accurate binaural sound locationaliztion model is also essential for good audio sound separation models (e.g., Chen et al. (2020); He et al. (2022), Hui et al. (2019))

For human being, binaural sound localization is the capacity to determine the location of sound sources in the 3-dimensional space relative to the center of the listener's head based on the signals perceived with two ears. The location of the sound source is usually

Figure 1.1: Polar coordinate illustration. Azimuth on the right side is defined as positive and negative vice versa.



described using polar coordinate, which consists of azimuth, elevation and radius as show in Figure 1.1,

Azimuth: the angle left or right of straight ahead (from $-180°$ to $180°$).

Elevation: the angle above or below the horizontal plane (from -90' to 90').

Radius: distance between sound source and the center of head.

In our work, our emphasis is placed on the direction of arrival (DOA) including azimuth and elevation, and the radius is omitted because it mainly affects the amplitude of the audio signal. To simplify the problem, we only consider the far-field condition, where the distance between the sound source and listeners is no less than 1.5 meters and the incidence angles of the waveform are almost parallel to each other.

1.3 Binaural localization mechanism in human being

1.3.1 Duplex Theory

There are many factors that can influence the perception of the sound location and many research has been investigated on this field. For examples, spectral content, reverberation and noise levels can all affect the perception of sound (e.g., Karunarathne et al. (2014, 2018); Mo et al. (2016), and Horner et al. (2004, 2006, 2009, 2011)). Several psychophysical hypotheses have been proposed, of which the most prominent hypothesis is 'Duplex Theory', which was proposed by Lord Rayleigh in 1800s (Rayleigh 1875). This theory states that our auditory system utilize interaural time difference and interaural level difference to localize sounds of low-frequency and high-frequency, respectively. It reveals the two physical factors underlying human binaural sound localization, that are interaural time difference (ITD) and interaural level difference (ILD).

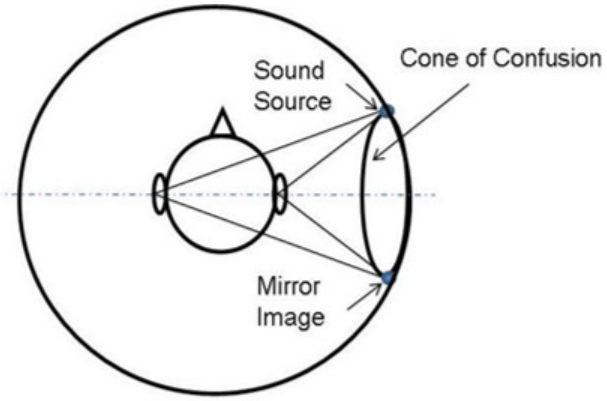
Interaural time difference is the time difference between the sounds arriving at two ears, which is mainly caused by the traveling time between the two ears. The interaural time difference focus on low-frequency between the maximum of interaural time difference for human being is calculated by $\pi \times d/(2 \times v_{speed}) = 1500$, where d is the diameter of human's head assumed to be around 17 cm, and v_speed is the speed of sound transmission in the air, which is assumed to be 340 m/s. Therefore, when the frequency is lower than 1500Hz, the interaural time difference is no more than one cycle, which is easy to determine. But when the frequency goes beyond 1500Hz, time alignment will occur. For example, given two pieces of 4000Hz sinusoidal tone with interaural time difference of $5\mu s$, it's ambiguous to determine whether the actual ITD is $5\mu s$ or $5 \pm 250\mu s$ because they look the same.

Interaural level difference is the sound level difference between the sounds that reach the listener's two ears. On the one hand, it is also caused by the distance for the further ear that distorts the energy during traveling. On the other hand, it comes from the head shadow effect because for higher frequency, the wavelength is shorter than the head size, and the wave will be reflected by the head, as well as the torso.

1.3.2 Cone of Confusion

ITD and ILD usually come up in conjugations, but localization merely with these two spatial will lead to ambiguity (So et al. 2006). In 3-dimensional space, a set of points

Figure 1.2: Cone-shape points in 3D space that shares the same distance difference to the left and right ear.



on a cone shaped location share the same interaural time difference and interaural level difference as the distance difference between these points to left and right ears are the same. When the listener's head is regarded as a sphere, it is more straightforward as shown in Figure 1.2

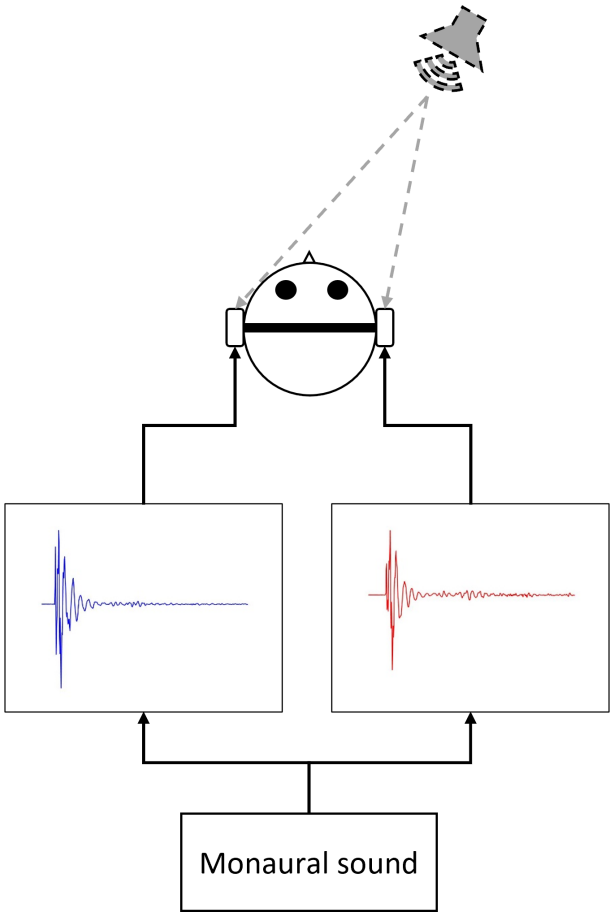
To overcome cone of confusion, spectral cues is utilized by human being to decode the exact location of sound source (**?**Au et al. 2011). Spectral cues come from the spectral changes result from physical effects including absorption of skin, baffle of head and torso and reflections of ear, especially the pinna. Spectral cues change with incident angle but no straightforward relation between spectral cues and incident angle. ITD, ILD and spectral cues are encoded in Head-Related Transfer Function (HRTF).

1.3.3 Front-back confusion

Front-back confusion is a special case of Cone of confusion, in which the listener hears a sound source from the front hemisphere and locate it to be the backward. The back-front confusion is defined similarly where the listener mis-locate the backward sound source to be in front.

Front-back/back-front confusion is a common psychophysical phenomenon in our daily life, and it has also been investigated in previous study, which we will review in the second chapter.

Figure 1.3: Monaural sound is filtered by a pair of HRTFS at azimuth 30°, elevation 0° to produce binaural sound. The listener would perceive it as coming from the corresponding direction. The blue and red curves in rectangles demonstrate the left and right channel waveform, respectively.



1.4 Head-Related Transfer Function

Head-Related Transfer function (HRTF) is a pair of transfer functions (including left and right transfer functions) that describe how a sound reaches the entrance of listener's two ears from a specific point in 3D space. It is defined as the amplitude response in frequency domain, and the corresponding time domain impulse response is call Head-Related Impulse Response (HRIR).

Monaural sound filter by HRTF is spatialized to binaural sound and when played through headphone, it is perceived as if it comes from the certain direction in the space. This is one of the most important applications of HRTF. The processing diagram is show in Figure 1.3.

1.4.1 Individualized and Non-individualized HRTF

HRTF varies across different people, and if we measure the HRTFs for the listener so that the synthesized audio stimuli would sound more realistic to the listener, which is called individualized HRTF. However, it's usually inconvenient to measure the HRTF of all the subjects, as a result, non-individualized HRTFs are always employed.

Non-individualized HRTFs come from measurement of other subjects. The utilization of non-individualized HRTFs will lead to degradation of the perception of binaural sound location. In most of the cases, azimuth location is less influenced. On contrary, the elevation is much more distorted. Therefore, it's of vital importance to select appropriate HRTFs for better immersion.

1.4.2 HRTF database

In the recent years, many famous HRTF database has been measured and released. The most well-known database is MIT KEMAR database, which was measure in 1994 in MIT's anechoic chamber (Gardner & Martin 1994). They utilized a KEMAR dummy head with two microphones mounted in both left and right pinna, which measure the left and right impulse responses, respectively. The distance between the dummy head and loudspeakers is 1.4 meters. The corresponding sampling frequency of recording is 44.1kHz. During the measurement, 710 different locations are selected which distributed between $-40°$ to $90°$ by every $10°$ in elevation and $-180°$ to $180°$ in azimuth by varying steps that depends on azimuth.

Besides KEMAR database, CIPIC HRTF database (Algazi et al. 2001) is employed in our work. It was measure in CIPIC interface laboratory anechoic chamber in UC Davis in 2001. It includes 45 different human subjects, and the KEMAR mannequin is also employed. The researchers measured the HRTF at 1,250 different locations, including 25 different azimuths from $-80°$ to $80°$ and 50 different elevations from $-90°$ to $270°$. The distance between the loudspeaker and subjects is 1 meter. The sampling frequency of recording is 44.1kHz.

For natural sounds, we employed speeches from TIMIT database, which contains American English sentences spoken by 630 speakers from various of major dialects. The sentences are designed to cover broad bands.

Database	Location	Notes
KEMAR	Azimuth 0° to 360°, Elevation −40° to 90°	Measured with KEMAR dummy head microphone in anechoic chamber
CIPIC	Azimuth −80° to 80°, Elevation −45° to 230°	Measured for 45 different human subjects in UCDavis
TIMIT	-	Phonetically rich sentences in American English wi th different dialects.

1.5 Binaural sound construction

There are many ways to construct binaural sounds. They are mainly divided into two categories: one is physical construction through loudspeakers and the other is virtual construction through headphones. These methods are summarized in Table 1.1.

For physical construction, one way is to attach the loudspeaker to a robotic arm and adjust the location by tuning the robotic arm with software. Another way is to mount loudspeakers on the surface of a sphere grid and play the stimuli through the speaker at target location. Physical construction can product the most realistic perception of spatialized sounds, but it also owns the disadvantage of inconvenience and high expenditure of facilities.

Virtual construction is another choice, which also consists of two methods. The first is to record the binaural sound with a dummy head with microphones mounted at the pinna. Afterwards, the recording is then played through headphones for reconstruction. This technology is widely used in order to provide better immersion and realism of the replay orchestra or chorus in theater. Therefore, it's not applicable for daily usage.

The second method is to construct with Head-Related Transfer Function (HRTF), in which the monaural sound that to be spatialized is filtered by the Head-Related Impulse Response at desired location. This method is more convenient as it can produce binaural sound without any physical configuration. The only requirement is the measurement of HRTFs at different locations. If the desired location is not included in the database, it can be estimated using interpolation with existing HRTFs. Besides, as the individualized HRTF is always unavailable, the utilization of non-individualized HRTF will cause

Methods	Explanation	Pros	Cons
Speaker	Placing a speaker at desired position	Easy and real	Inconvenient and costly
Headphone (Recording)	Record the audio at the entrance of listener's ear and play through the headphone.	Real	Ungeneralizable
Headphone (Virtual Auditory Space)	Filter the single channel audio with Head-Related Transfer Function at desired position	Repeatable and convenient.	Require individualized measurement of HRTFs.

Table 1.1: Summary of methods to construct binaural sounds

degradation on the location perception of the sound. To overcome the effect, selection of HRTFs with similar property as the subject's individualized HRTF would be appreciated.

1.5.1 Cochleagram

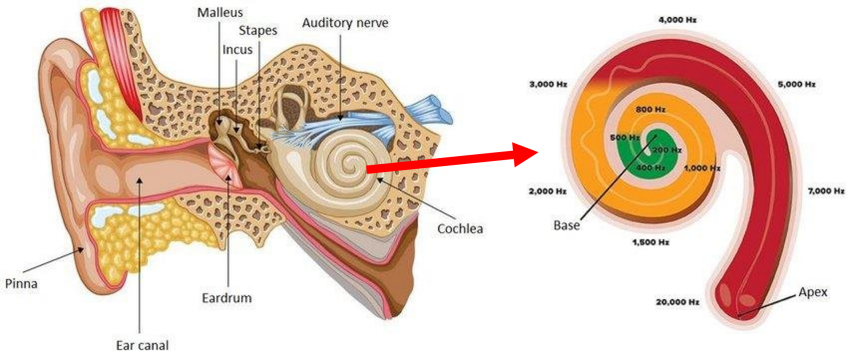
Cochleagram is a special version of spectro-temporal representation that mimics the response of auditory fibers in cochlea as illustrated in Figure 1.4. Cochlear is a spiral-shape organism in inner ear with different parts from base to apex response to different frequencies. To simulate this effect, gammatone filter bank is applied to measure the displacement of basil membrane of the hair cells in inner ear.

The gammatone filters take the following form,

$$G(t) = at^{n-1} \exp^{-2\pi bt} \cos(2\pi ft + \phi) \tag{1.5.1}$$

where a is amplitude, n is the order, b is the bandwidth, ϕ is the phase of the carrier and

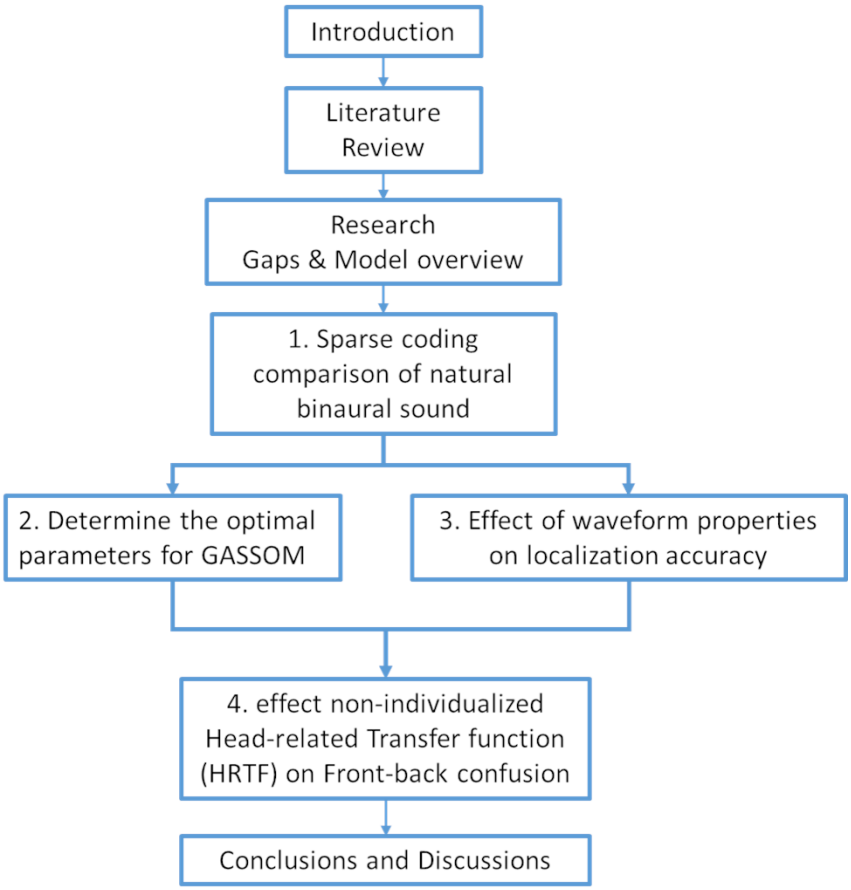
Figure 1.4: The left panel demonstrates the anatomical figure of human's auditory periphery. Modified from (İlik 2018). The right panel demonstrated the cochlear in inner ear.



f is the center frequency in Hz.

1.6 Outline of the thesis

Figure 1.5: Outline of thesis



Chapter 2

Literature Review

2.1 Summary

In this chapter, studies on binaural sound localization are firs reviewed, which uncover the underlying mechanism of sound localization and what kind of spatial cues are employed by mammalians to decode location. Computational models based on these artificial spatial cues are developed. Besides, sparse coding, including Generative Adaptive Subspace Self-Organizing Map (GASSOM) and Independent Component Analysis (ICA), as alternatives to extract spatial cues are introduced subsequently. Based on the spatial features extracted by GASSOM, decoding the sound source localization is modeled as a classification problem and Deep Neural Network is employed to solve it.

To evaluate the performance of the GASSOM-DNN based binaural sound source localization model, empirical data of psycho-acoustic sound localization experiment on human subjects are introduced, which investigated that effect of spectral properties of the stimuli on localization accuracy. Non-individualized HRTFs are widely used in both daily application and research field, which causes degradation of localization performance. Research on exploring the influence of non-individualized HRTFs on front-back confusion and how to eliminate the negative effect are also reviewed.

2.2 Duplex Theory

Binaural sound localization plays an important role in the daily life. It provides extra information in the so called cocktail party effect (Brown & Cooke 1994, Stecker & Gallun 2012). Studies on the mechanism of how sound localization is decoded by human being has attracted lots of attention and many hypothesis has been developed. Among these

studies the most well-known is the Duplex Theory, which was proposed by Lord Rayleigh in 1800s (Rayleigh 1875). Details about Duplex Theory can be found in chapter one and will not be repeated here. In summary, interaural time difference (ITD) and Interaural Level Difference (ILD) are believed to play an important role in the localization of lower frequency (< 1500Hz) sound and higher frequency (> 1500Hz) sound, respectively.

2.3 Computational models on binaural sound localization

Based on ITD and ILD, the azimuth location of the sound is easy to decode according to the monotonic relation between the ITD and ILD with azimuths. Many analytical computational models are developed to decode the location by estimation of sound location with spatial cues. A probabilistic model (Willert et al. 2006) was proposed to estimate the conditional distribution of ITD and ILD with respect to the sound source location using cochleagram generated by cochlear model in narrow bands, and the location that maximize the likelihood is determined as the sound location. Machine learning techniques such as Deep Neural Network (DNN)(Ma et al. 2019, May et al. 2015) are also employed for sound localization by mapping the ITD and ILD features into locations. Besides, Convolution Neural network (Zhou et al. 2019, Thuillier et al. 2018, Vecchiotti et al. 2019) is utilized for sound localization where the cochleagram/spectrogram are considered as 2D images. Therefore, similar techniques on images processing can be applied.

However, spectral cues resulted from the shape of pinna and canal are also of vital importance in localization(Middlebrooks et al. 1989), especially for discrimination of elevation including front-back confusion, while are always omitted by these models. Other methods was proposed based on HRTFs to extract spatial cues (Goodman et al. 2013), but HRTFs are usually unavailable for most of cases. Moreover, machine learning methods have the disadvantage that the feature extracted by the model is not visualizable.

To employ spectral cues for sound source localization, Source Cancellation Algorithm (SCA) was proposed for 3D sound sources are proposed (Keyrouz & Diepold 2006), in which they calculate the spectral cues by dividing the discrete FFT of left channel signal from the discrete FFT of right channel in frequency-wise. The spectral information carried by the sound source is canceled and the results are compared with corresponding

HRTF division collection, of which the direction with greatest similarity is selected as the predicted location. Nevertheless, this algorithm requires access of the corresponding database.

Traditional methods requires analytical form of spatial feature extractor in advance. Sparse coding provide a feasible way to solve this problem (Palakal et al. 1995).

2.4 Sparse coding on natural stimuli

It was proposed by Barlow (Barlow 2012) that the sensory system evolve to adapt to the environment statistics to convey as much information as possible with fewer neurons, this is so called efficient coding or sparse coding as efficiency means transmit more information with less neurons in a sense. Recent studies have demonstrated that the efficient learning of natural images and natural sounds (Lewicki 2002) leads to receptive fields of primary visual cortex neurons and auditory periphery, respectively. Independent component analysis has been applied to natural images and results in oriented, band-pass edge filter which is similar to receptive fields of simple cells that are found in humans' primary visual cortex (Olshausen & Field 1996).

Although there has been some experiment of auditory neural responses, the efficient coding of audition is less fruitful comparing with vision. (Lewicki 2002) shows that efficient coding of natural sounds or speech reproduce wavelet shape dictionary. Locally Competitive Algorithm (LCA) (Carlson et al. 2012, Klein et al. 2003) is also applied to encode introspect-temporal representation of audio stimuli and the results learned by these sparse coding algorithms were similar to the spectral-temporal receptive fields (STRF) of auditory cortex neurons that have been investigated in mammalians. But these studies only consider monaural sound, while binaural sounds are less investigated.

Moreover, recent work (Mlynarski 2014) shows that the efficient coding of spectrogram/cochleagram lead to emergence of Spectro-Temporal Receptive Fields (STRF) of neurons in Medial superior olive and Inferior Colliculus. In more recent work, Barlow (Barlow 2001) also hypothesis that informative features to make decisions can be learned by reduction of redundancy on input stimulus. However, the corresponding support is still sparse.

After attaining the representation of auditory signal, it is essential for organism to

recover important information about the environment to perform reaction. For example, during hunting or escaping from danger, the animal needs to extract location information of the surroundings to make decision. Eco-location bats (Palakal et al. 1995) have been investigated and the researchers found that within the auditory cortex of Eco-location bats, there are some topographical neurons that are functionally aligned with similar response to directional information of the echo they received. Lewicki (Lewicki 2002) demonstrates that applying ICA to binaural audio signals lead to emergence of spatial information, such as interaural time difference and interaural level difference. The learned basis functions are grouped into binaural and monaural according to spatial sensitivity analysis. Decoding of the location of sound sources based only on binaural basis functions reaches accuracy of more than 99%. Despite the success of ICA, there is no promising biologically inspired evidence for application of ICA. Meanwhile, the topography emerges in primary visual cortex (Olshausen & Field 1996) as well as primary auditory cortex is not emphasized. ICA also requires the full access to training data (Hyvarinen et al. 1998), and it's not task-specific as it's based on the independent and non-gaussianity of the subcomponents in the mixture (Hyvärinen & Oja 2000). Therefore, we investigate the application of generative adaptive subspace self-organizing map (GASSOM) to natural binaural sounds.

GASSOM (Chandrapala & Shi 2015) is an extension of Kohonen's adaptive subspace self-organizing map (ASSOM) with removal of constraint on input data. ASSOM (Kohonen 1990, 1996) is an invariant feature detector which try to find a manifold or subspace that the projection of input data to this subspace is maximize (that is the projection error, which is orthogonal to the projection, is minimized). Each subspace/node in ASSOM corresponds to certain pattern shared by data within an episode where that patter is not changed. When the transformation is in small scale, the pattern can be approximated by a linear transformation. Another advantage of ASSOM is that the manifold is aligned topographically according to function of each subspace. Usually, nodes in ASSOM that correspond to each manifold are aligned on a 2D space.

Temporal slowness as well as sparsity are two underlying principle for development of GASSOM: temporal slowness and sparsity. Temporal slowness(Lies et al. 2014) comes from the fact that the interaction between sensory system and environment is relatively stable, which is mostly true in our daily life like when people are fixing their eyes or

listening to someone talking. However, there is a constraint for the training of ASSOM that each input should be labeled clearly according to the common pattern, which is noted as an episode. The episodes are fed to the ASSOM sequentially. For GASSOM, the episodes are modeled using Hidden Markov Model to formulate the transition probability between them, and therefore remove the requirement of explicit label of episodes. This allows us to apply GASSOM to natural unlabeled data such as natural images generated by eye movement model or natural sound generated by static listener to cocktail party.

GASSOM has been proved success in prediction of the receptive fields of primary visual cortex neurons (Chandrapala & Shi 2015). It has also been successfully applied to encode spatialized bat echoes (Wijesinghe et al. 2021), but it has never been applied to encode natural sounds (e.g. speeches) and the time scale are always different from human's auditory system. In this study, we apply GASSOM to binaural natural speech from TIMIT database, as vision and audition are two major modality of human beings that share lots of similar properties. We simulate the movement of speaker on the horizontal plane of the listener by convolving the randomly selected speech with KEMAR head related transfer functions (HRTFs). HRTF quantify the transfer function from the sound source to the listener composite of spatial transfer and head and torso of human. During each episode, the location information, which is preserved in the HRTF, is constant. Therefore, the invariant feature of GASSOM will learn this stable feature during training. As a result, this leads to the emergence of basis functions that contains spatial cues of different direction embedded in HRTFs. The projection of binaural stimuli on the learned basis functions can produce most of the response as auditory neurons even thought it's linear operation (Schnupp et al. 2001).

2.5 Empirical data on human localization

Binaural sound localization has also been investigated with human listener about the factors that influence the localization accuracy. In 2010, William Yost conducted a series of experiment on the influence of temporal and spectral modulation. Effects of center frequencies together with bandwidth (Yost & Zhong 2014, Yost et al. 2013) , as well as sound duration, sound levels and sound envelope(Yost 2016, 2017) on sound localization accuracy on ear-level horizontal plane in frontal hemisphere are explored, in which human

subjects are asked to judge the location of sounds with different characteristics played through loud speakers at different discrete locations from the left side to the right. The results demonstrated that human's sound localization accuracy is heavily dependent on the frequency modulation of the stimuli while less dependent on the temporal modulation or amplitude.

Biological constraints on spectral limitation was examined (Jin & Carlile n.d.) with a neural system model. The sound localization model employed in this work achieves human-like performance when testing with stimuli of different bandwidth. It is in consistent with the recent study that auditory system compute spatial features within narrow band and integrate them across frequencies to predict the sound location. However, similar work on the effect of those stimuli features on the performance of computational localization model was still sparse.

Non-individualized HRTF also has a significant effect on the localization performance, especially for front-back confusions. Experiment on comparison of localization accuracy between stimuli synthesized with individualized and non-individualized HRTFs was conducted on human being (Wenzel et al. 1991, 1993). In their experiment, free-field speakers were utilized as replacement for individualized HRTFs as it has been proven (Wightman & Kistler 1989) that these two methods achieve similar performance. Significant increase in front-back confusion was discovered when listening to stimuli synthesized with non-individualized HRTFs. However, most of the computational model test the localization performance with the same HRTFs as training.

2.6 Complementary

At the same time of finishing the thesis, another paper was published (Francl & McDermott 2022) in which the author proposed to use Deep Convolutional Neural Network as localization model to benchmark with human being on localization performance. The author trained 100 neural network with different configuration on layers (including the number of layers and the number of nodes on each layer) and then selected 10 networks that performs best on localization accuracy. Then the selected 10 models were benchmarked with empirical data on human being experiments.

In this study, the benchmark consists of localization accuracy for both azimuth and

elevation separately, including front-back folding and front-back unfolding (folding means only consider frontal hemisphere by auto-correct the front-back confusion while unfolding means keep the original predicted location). They also considered the influence of frequency range using Gaussian white noise filtered by high-pass and low-pass filters. They also tested the model in multi-source condition and the model can estimate the number of sound sources that is quantitatively similar to human (around 4). The model also performs precedent effect in response to two subsequent stimuli with varying tiny interval that is similar to human being.

It should be noted that the training model is in natural environment that includes reverberation and reflections. They also investigated training the model in ideal anechoic environment, but found out that the performance of the model is less human-like in ideal environment, which reflected that human's behavior is a composite result of complex audio environment.

However, the model in the work utilized Deep CNN only and the spatial features learned by the model is hard to visualize, which prevent the researcher from learning the detailed underlying mechanisms for binaural sound localization. Our GASSOM-based computational model, however, owns the advantage that the spatial features learned is easy for visualization and well-organized. Besides, the Deep CNN model can only predict one sound source at the same time, and the discussion for precedent effect can be improved with more general stimuli such as speeches.

2.7 Research Gaps

Gap1: Current Binaural localization models are mainly focusing on artificial features, while investigation of sparse coding of natural sound for sound localization is less emphasized.

Gap2: GASSOM has been applied to encode visual stimuli and proven successfully to learn invariant features similar to the primary visual cortex, but it has not been applied to encode general audio stimuli. No comparison was made on which algorithm is more suitable for sound localization.

Gap3: Little attention has been paid to the performance of computational model under different physical constraint while empirical data on human being is fruitful.

Gap4: Most computational models only consider individualized HRTF to synthesize stimuli for testing while there's no attempt to simulate the effects of non-individualized HRTFs on front-back confusion and no benchmark between model output and empirical data.

2.8 Research Questions

According to the literature review, our study will solve the following research questions:

Q1: ICA has been applied widely in sparse coding of natural stimuli, particularly binaural sounds. However, ICA is not efficient because it could not discriminate the spatial characteristics of binaural sound from other acoustic features, such as speech content. GASSOM, as a competing sparse coding algorithm with topographic structure and visualizable basis functions, has been successfully employed to encode visual stimuli. Can GASSOM be a better replacement of ICA as a sparse coding alternative to simulate binaural sound localization?

Q2: If GASSOM were better replacement of ICA for simulating binaural sound localization, the next research question would be what are parameters associated with designing a GASSOM model that would significantly affect its ability to simulate sound localization responses? And are there optimal parameters?

Q3: It has been proved that human's auditory system integrates narrow band acoustic cues to predict the location of the sound, and therefore the localization accuracy is highly depending on the spectral properties of audio stimuli. The RQ is does the GASSOM model produce human-like sound localization performance dependency on spectral properties of audio stimuli?

Q4: As the basis functions of GASSOM is similar to human's auditory neuron receptive fields, what's the effect of non-individualized HRTFs on the front-back confusion performance of GASSOM-binaural localization model in comparison with empirical data on human being?

Chapter 3

Sparse coding of binaural sounds

3.1 Summary

In this chapter, we give a brief overview about the framework of GASSOM-based binaural sound localization model and compare the performance of Independent Component Analysis (ICA) with Generative Adaptive Subspace Self-Organizing Map (GASSOM) on the extraction of spatial features from binaural sound cochleagram.

The binaural localization model mainly consists of two stages: In the first stage, a set of spatial cue filters are constructed to extract spatial features. In the second stage, the location of the sound source is estimated based on the spatial features extracted by the filters constructed in previous stage. This process is usually modeled by a probabilistic model or deep neural network. We followed the previous study by employing a 3-layer deep neural network to decode the azimuth as well as the location.

The key problem in this chapter is to seek for the optimal sparse coding algorithm for binaural stimuli that would efficiently extract direction-related features. To determine the appropriate sparse coding algorithm, Generative Adaptive Subspace Self-Organizing Map (GASSOM) and Independent Component Analysis (ICA) are selected as two candidate algorithms, both of which have been proved to be capable of encoding natural visual stimuli and would produce an efficient representation of the signal that is similar to the receptive fields of simple cells in visual cortex.

Comparison was conducted between basis functions of ICA and GASSOM in the consideration of performance on extraction of spatial cues and topographical structures of the learned map. GASSOM was proved to outperform ICA on both metrics: Fisher information (FI) and Disorder Index (DI). Fisher information is calculated on each basis function

with respect to the sound directions to quantify to what extent the basis functions are about directions, and GASSOM basis functions were proved to be more informative about directions than ICA basis functions. Disorder index (DI) of the basis functions is also computed as metrics to quantify the similarity between the neighboring basis functions on the map, and GASSOM shows much better topographical smoothness than ICA.

3.2 Introduction

Unlike the widely-used microphone array technology which utilize several microphones to determine the location of sound, mammals localize the sound source with two ears and are able to achieve a relative high accuracy. In the past few years, many binaural sound localization models have been proposed that are motivated by human's superb performance on localizing sounds with merely two ears. Those models always contain two parts: spatial feature extraction and location estimation.

For the first part, traditional methods depend heavily on the disparities between left and right ears, which are known as interaural time difference (ITD) and interaural level difference (ILD) that are driven by the psychophysical experiments, as the spatial cues for the inference of location, while in fact spectral cue also plays an essential role in this task and is always omitted because its complex form is hard to determine.

In traditional methods, the filters are constructed based on psycho-acoustic experiments to extract the pre-determined spatial features such as the Interaural Time Different and Interaural Level Difference, where ITD is calculated by computing the correlation between the left and right channel and the delay corresponding to the maximum correlation is selected as Interaural Time Difference, while the power ratio of the left and right channel audio is calculated and transformed to decibel is defined as Interaural Level Difference. These two spatial features are based on rule of thumb and have some internal limitations. In this work, we remove the requirements for such a fixed analytical form of feature filters by applying sparse coding algorithm on binaural audio cochleagram as it has been proven that sparse coding on binaural stimuli leads to the emergence of spatial feature basis functions automatically without any manipulation or assumption about the filter form.

After construction of spatial feature extractor, the next step is usually modeled as a

multi-classification problem, where the training data are the outcome of previous stage and each class corresponds one position respectively. Probabilistic model such as Gaussian Mixture Model (GMM) can commonly model that utilized for this stage. With the development of computational power, machine learning were gradually hired by more and more researchers to solve this problem. Deep neural network (DNN) is one of the most popular tool for this multi-class classification task. In preliminary study, we compared the performance of localization accuracy between DNN and GMM based on the same spatial features, and found out that DNN outperform GMM in most of the cases. DNN also owns the advantage of wider application such as the Multi-Condition Training, which train the neural network under different Signal-to-Noise Ratio (SNR) to increase the generality and robustness of the model. Therefore DNN is selected for the second stage in this work.

In this chapter, our emphasize is put on the selection of appropriate algorithm for the extraction of spatial features. To dig up the spatial cues embedded in the audio stimuli, sparse coding is considered instead of the traditional analytical methods that is based on thumb of rule. Sparse coding has succeeded in producing bio-consistent neural coding patterns. ICA, as one of the most popular sparse coding algorithm, has been applied widely to encode natural stimuli such as audio patches, waveform sounds. More specifically, applying ICA to binaural sounds would lead to the emergence of basis functions that contain binaural cues (Mlynarski 2014). However, ICA is not task-specific and always results in miscellaneous outcomes because it could not discriminate the spatial characteristics of binaural sound from other acoustic features, such as onset, offset, pitch and harmonic stack. In the previous work, the corresponding ICA algorithm utilized by the author, Fast-ICA, also require the full access to the whole data for training. This make it difficult to apply the model for online learning. GASSOM, as a competing sparse coding algorithm with topographic structure and visualizable basis functions, has also been successfully employed to encode visual stimuli, which produces basis functions that are similar to the receptive fields of human's primary visual cortex neurons. Accordingly, the first research question is proposed as follow: Can GASSOM be a better replacement of ICA as a sparse coding alternative to simulate binaural sound localization?

Biological linkage

As we have introduced before, most of the sound localization model consists of two phases: the feature extraction and sound location estimation. The selection of these two phases is straightforward, but it also has biological explanation based on the investigation of auditory system in mammals. Here we will give a simple and intuitive analogy between the binaural localization model and auditory neural system.

The mammals ascending auditory system has a hierarchical structure. We can separate the processing of binaural cues into three stages: In the first stage, where the binaural sound arrived the ears, the waveform audio that is transmitted into sound pressure that perceived by the tympanum, is then analyzed in frequency domain by the cochlear to generate a spectral-temporal representation. In our model, as the HRTF is recorded at the tympanum, this process is simulated using Gammatone filter bank which outputs cochleagram. The following second and third stages are related to auditory ascending system.

The second stage corresponds to the pathways before Inferior Colliculus (IC), including IC; This part works for the primary process of auditory signals. For example, the neurons in Dorsal division of cochlear nucleus (DCN) are thought related to the process of spectral cues, that is to extract the spectral cues from binaural sounds that related to the perception of elevation; It also includes neurons in Lateral Superior Olive (LSO) and Medial Superior Olive (MSO) that are related to the extraction of Interaural Level Differences (ILD) and Interaural Time Difference (ITD), respectively. In our model, the spatial feature extractor, GASSOM or ICA, is analogous to this part. The basis functions learned by these algorithms serves as the functional neurons that could extract ITD, ILD as well as spectral cues for subsequent process in higher level.

The third stage corresponds to the pathways after IC, including Medial Geniculate Body and auditory cortex. The spatial cues learned from previous stages forms the perception of spatial location in this stage. However, the detailed mechanism in mammals auditory system is unrevealed. Therefore, we utilized a Deep Neural Network (DNN) to model this 'black box' as for both cases, the input is primary spatial cues and the output is predicted location. However, with the development of study on the underlying mechanism of auditory cortex, a more realistic and complex framework could be employed to model this process to achieve better performance in the future work.

3.3 Binaural sound localization model

In this chapter, we will give an overview on the GASSOM-DNN based binaural localization model before the comparison between ICA and GASSOM. The paradigm of the framework is illustrated in Figure 3.1. In the first step, monaural stimulus is filtered by HRTF at certain direction to produce binaural sounds. Both the left and right channels are passed through a set of Gammatone Filter bank that simulates the response of different frequencies along the structure of cochlea. For each frequency bin, half-wave rectification and cubic root were taken for the output of Gammatone filter bank to simulate the compression effect of the inner ear, and then divided into small frames, which results in cochleagram. Cochleagram of left and right channel were segmented into small blocks with overlapping, and then concatenated and normalized to zero mean and unit variance for GASSOM training. For each iteration, GASSOM was trained by one batch of blocks. After the training of GASSOM, new stimuli were generated and processed by the steps mentioned before to generate small blocks, which projected onto the basis functions of GASSOM. The output of GASSOM are fed into a Deep Neural Network (DNN) to train for the decoding of the sound location.

3.3.1 Preprocessing

The monaural stimulus is first spatialized by filtering with HRTF at specific location to generate binaural sounds. It should be noted before the whole process, the binaural signals are filtered by a first order high-pass filter,

$$f(t) = 1 - 0.99t; \tag{3.3.1}$$

which aims to pre-emphasize the high frequency elements of the signal to compensate for the high frequency energy distortion during the recording of audios.

After retaining the left and right channel waveform signals, they are first analyzed by the auditory periphery system to generate the Spectro-temporal representation, named Cochleagram, as the output is similar to human's inner ear cochlea organism frequency response. This process is simulated by gammatone filter bank followed by half wave rectification and cubic root that simulate the compression effect.

Figure 3.2 shows a demonstration of the frequency response of gammatone filters with each color corresponds to one gammatone filter, respectively. Different part along the cochlear structure from base and apex are sensitive to different frequencies. Accordingly, the gammatone filter bank consists of 128 gammatone-shape filters with center frequencies range between 100Hz and 20,000Hz. The bandwidth for each filter is dependent on its corresponding center frequency and can be simply approximated by the Equivalent Rectangular Bandwidth (ERB),

$$ERB(f) = 24.7 \times (4.37 * f + 1) \tag{3.3.2}$$

where the unit for f and $ERB(f)$ are kHz and Hz, respectively. The center frequencies are designed so that each filter has the same ERB.

The output of the gammatone filter bank is then half-wave rectified and smoothed with a low-pass filter to extract the envelope, after which the cubic root was taken to simulate the compression effect. These subsequent processes seek to simulate the non-linear transduction of the hair cell.

Temporally, the output in each frequency channel is then divided into small frames, within each frame it integrates the power energy of 8ms in duration, and the shift between successive frames is also set to 4ms. A logarithm transformation, which simulate the

Figure 3.1: Paradigm of the binaural localization model. The monaural sound is first spatialized with HRTF, which are then processed according to auditory periphery system to retain the cochleagram. Cochleagram are divided into small chunk that each contains 10 frames. The left and right chunks are concatenated together and passed through GASSOMs to extract spatial features and DNN to decode the location

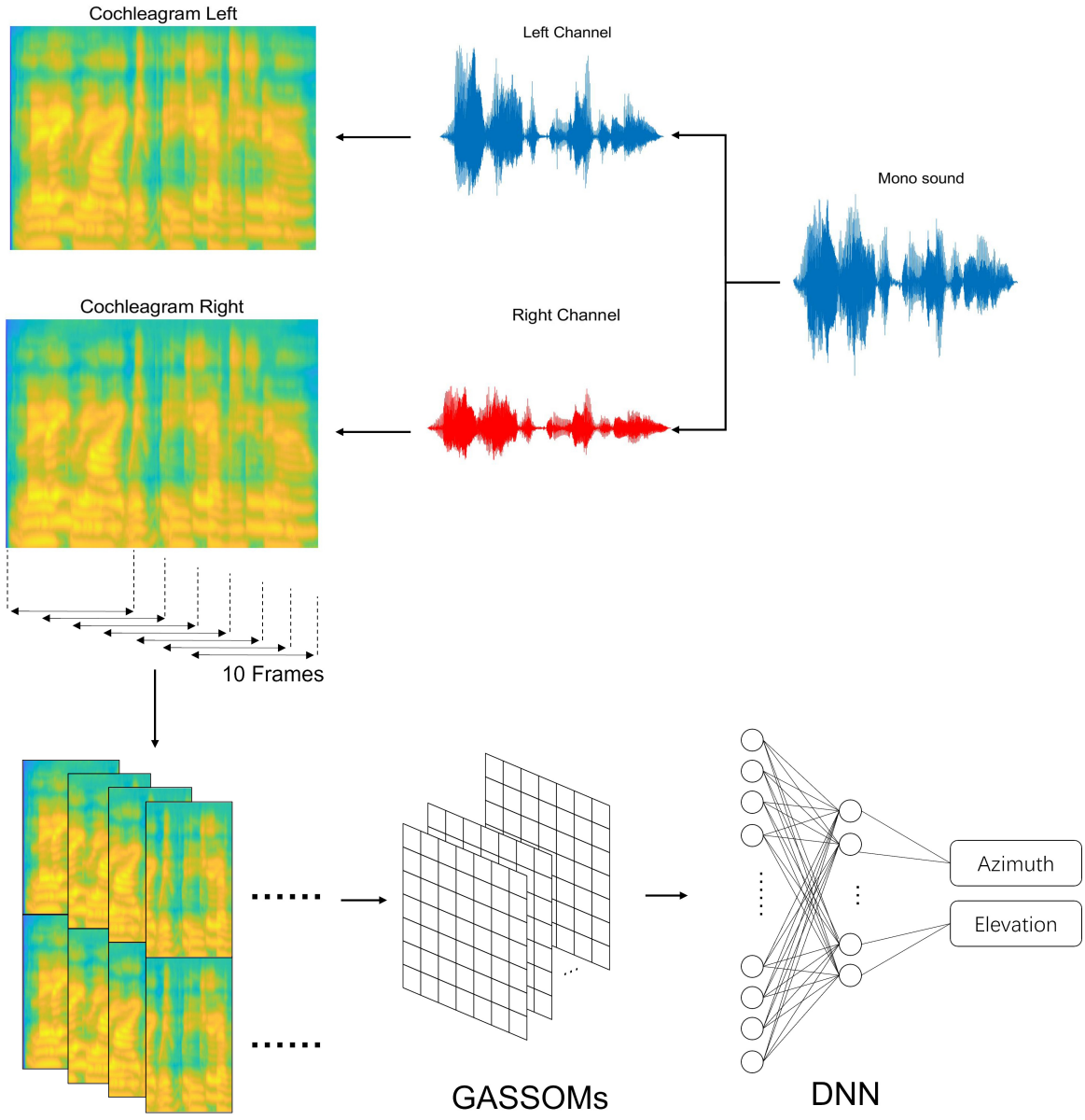
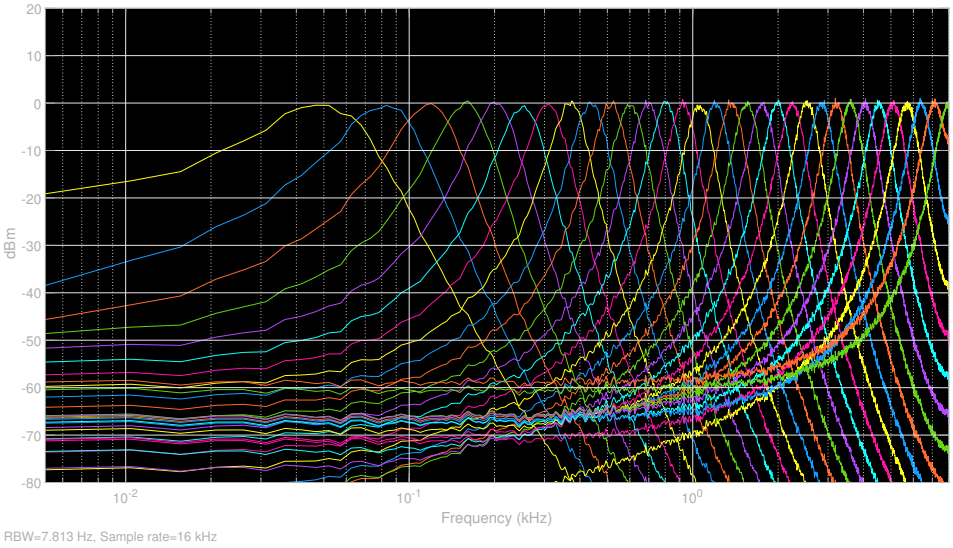


Figure 3.2: Frequency response of gammatone filter banks with different center frequencies (Plotted using MATLAB Spectrum Analyzer toolbox). Different color stand for different gammatone filters. The horizontal axis is log-scale frequency in Hz. The vertical axis is the frequency response in dB



cochlea compression effect, is then taken for the resulted spectro-temporal representation to get the eventual cochleagram. Similar preprocess was also employed by several previous study that focused on sparse coding of natural sounds.

3.3.2 Spatial feature extraction

After the acquirement of the spectro-temporal representation of the natural sounds for both left and right ear, they are then further divided into overlapping chunks for following process. Each chunk contains 10 successive small frames and for all the frequency channels. The stride between neighboring chunks is 10%, that is one frame. Such small stride is selected for the temporal slowness assumption, which will be discussed in detail in next chapter.

The chunks for left and right ears are then concatenated and reshaped to a vector for training. The vector is then normalized to zero mean with unit variance for robust performance. The left and right ear parts are normalized together because if we normalize them separately, the interaural level difference is eliminated. However, it should be noted that during this step, the sound level factor is eliminated.

The divided cochleagram are subsequently used for sparse coding. ICA and GASSOM

are considered in this work, and details are discussed in next section.

3.3.3 Classification for sound location

The deduction of the sound location is regarded as a classification problem, where the input is the spatial features, and the categorical classes are the sound locations. The spatial features are generated by passing the binaural sound through the extractors constructed by sparse coding. The extractor contains many basis functions of same length as the input stimuli, and the squared projection of the input stimuli onto all the bases are defined as the extracted features. Therefore, the length of the features is the same as the number of basis functions.

This classification problem can be solved by many probabilistic methods, of which Gaussian Mixture Model (GMM) and Deep Neural Network (DNN) are most widely used.

GMM depicts the distribution of the spatial features with respect to the location of the sound as a linear combination of Gaussian distributions,

$$p(s) = \sum_{l=1}^{L} p(s|D_l)p(D_l) \tag{3.3.3}$$

$$p(s|D_l) = \mathcal{N}(s|\mu_l, C_l) \tag{3.3.4}$$

Where $D_l, (l = 1 \dots L)$ is the label for the numbered location, and s is the latent coefficient, which is the spatial features in our case, and μ_l and C_l are the mean and covariance matrix of the Gaussian distributions, respectively. Without loss of generality, the probability for the present of different locations is equally balanced, so $p(D_l)$ can be simplified as a uniform distribution. The GMM is then reformulated as to find the location that maximize the posterior distribution of the latent coefficients,

$$\hat{D} = \arg\max_{l} p(s|D_l) \tag{3.3.5}$$

However, the distribution of the spatial feature is not always Gaussian distributed, so the performance of GMM in the decoding of location is unsatisfying. Besides GMM, DNN is also widely used in the sound localization model.

DNN is an artificial neural network that is comprised of multiple layers, including an input, an output layer, and many artificial-configured latent layers. It is inspired by

Layer	Configuration
Hidden Layer 1	200 nodes
Hidden Layer 2	200 nodes
Hidden Layer 3	50 nodes
Output Layer	19 nodes

Table 3.1: The number of nodes for each layer. Output layer number varied according to the task

the hierarchical structures of human neural systems. Compared with GMM, DNN owns much more sophisticated forms, and therefore can be used to solve more complicated classification problem with more layers. However, the increasing number of layers will lead to the over-fitting problem, which should be avoided.

In our work, we employed a DNN with 3 hidden layers. The configuration of the hidden layers is illustrated in the table 3.1.

The neural network employed here is feed-forward network. At the very beginning, the hidden layers are fully connected. After each hidden layer, a Rectified Linear unit (ReLU) layer and a Dropout layer are added. ReLU layer is used as an activation function in the following form,

$$f(x) = max(0, x) \tag{3.3.6}$$

Where x is the output from the previous layer. Dropout layer is used to avoid over-fitting problem, and the dropout rate is 0.2, which means in the end 20% of the neurons will be neglected.

Between the output layer and 3rd hidden layer, a SoftMax layer is inserted for multi-class classification. The output of SoftMax is the probability for different directions and is averaged over time, and the direction with greatest probability is selected as the predicted location. The number of output layer is task-dependent and 19 nodes is exemplary corresponding to 19 directions from $-90°$ to $90°$ by every $10°$.

The neural network is optimized with Stochastic Gradient Descent with Momentum (SGDM) optimizer. The data is shuffled by every epoch. The initial learning rate is 0.1 and drops by 0.1 after every 10 epoch for convergence. These parameters are tuned with rule of thumb.

3.4 Sparse coding on binaural sounds: ICA and GASSOM

3.4.1 Basic configuration

We first introduce the basic configuration for both ICA and GASSOM. All the parameters are kept the same for comparison.

The sampling frequency is 44.1kHz. It is selected as the upper limit of frequency for human's hearing system (e.g. cochlea) is around 20kHz. According to Nyquist sampling rule, The sampling frequency should be greater than 40kHz. Besides, the sampling frequency for database used in this study, TIMIT, KEMAR and CIPIC, all selected 44.1kHz as sampling frequency. In this case, we don't need to resample the data, which keep the originality and integrity of the spectrum of the sound and HRTFs.

3.4.2 GASSOM

GASSOM data generation

The training session contains 50,000 iterations. For each iteration,200ms monaural sounds are randomly picked from TIMIT database, and the HRTFs are from MIT KEMAR HRTF database. The directions include the front hemisphere azimuths on ear-level horizontal plane and range from $-90°$ to $90°$ by every $10°$. The direction is randomly selected uniformly. The binaural sounds are then processed to get the cochleagram chunk and fed to GASSOM to encode and update the basis functions.

GASSOM settings

The parameters for GASSOM is configured basically referred to (Wijesinghe et al. 2021) as both are applied for audio stimuli but of different scale. Therefore the parameters are then fine tuned according on rule of thumb.

All the basis functions are initialized with white noise that uniform distributed between -1 and 1. The topo-space of GASSOM is set to 10 by 10 for a trade-off between efficiency and localization accuracy, which is further discussed in next chapter. The learning rate is scheduled to starting at 8e-4 and decrease exponentially to 1e-5 according to the iteration, with time constant of 10,000 and maximum iteration of 50,000. The starting learning rate

is selected so that the basis functions will be updated from white noise to . The selection of ending learning rate is to ensure the convergence of learning amplitude. The standard deviation of the neighborhood function starts at 2 and decays exponentially to 0.2 with the same time constant. For different map size, these two parameters are scaled accordingly, not only for consistency, but also make sure that initially all basis functions have the chance to be selected to encode the input stimuli and topological convergence in the end (that is only a small portion of basis functions are updated to encode the input).

For transition probability matrix, it consists of a uniform distribution with a Gaussian distribution, where the uniform distribution is selected to represent the probability of the change of different winner nodes between different episodes, and the Gaussian distribution is selected to represent the change of winner node within each episode. The weight of uniform distribution is 0.4, and the standard deviation of the Gaussian distribution part is 2.25. The weight parameter is selected based on the portion of empirical data, which corresponds to the percentage of episode changes in all batch changes.

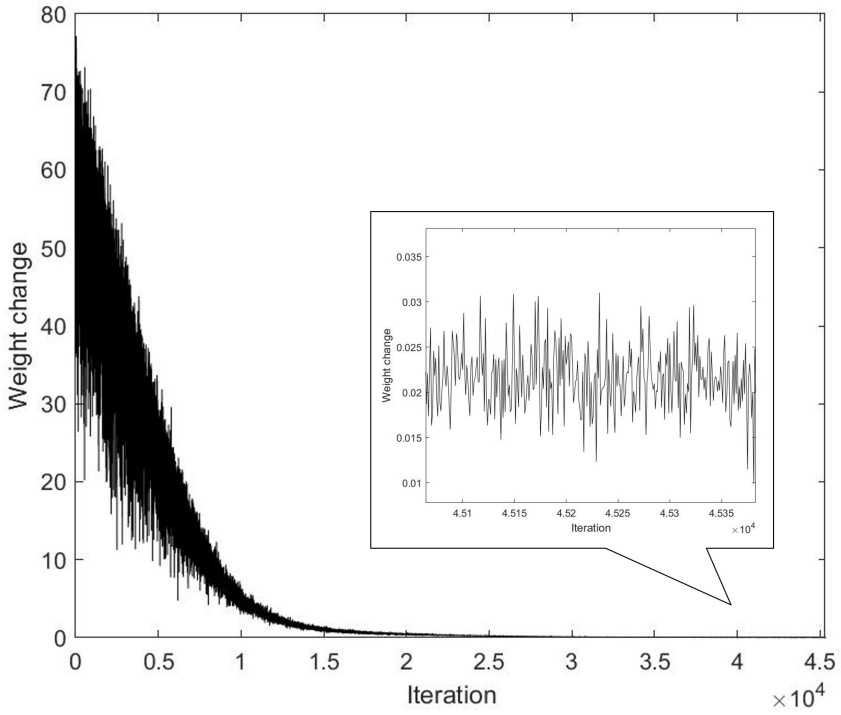
GASSOM Training convergence

To ensure the convergence of training process, the maximum iteration number is set to 50,000 and the convergence threshold for weight change is set to 0.01. The weight change is defined as the update of weight on successive iteration and averaged across all the basis functions,

$$\Delta\omega = \frac{\sum_{i=1}^{S} |\omega_{i,t-1} - \omega_{i,t}|}{S} \tag{3.4.1}$$

where S is the number of basis functions, and t is the iteration time/number. One of the training result is demonstrated in Figure 3.3

Figure 3.3: The weight change v.s. iteration number is illustrated and the training process is stopped at 45,381 where the weight change is less than 0.01. It's not beyond the maximum iteration but still shows good convergence as the weight change decreases exponentially and close to 0.

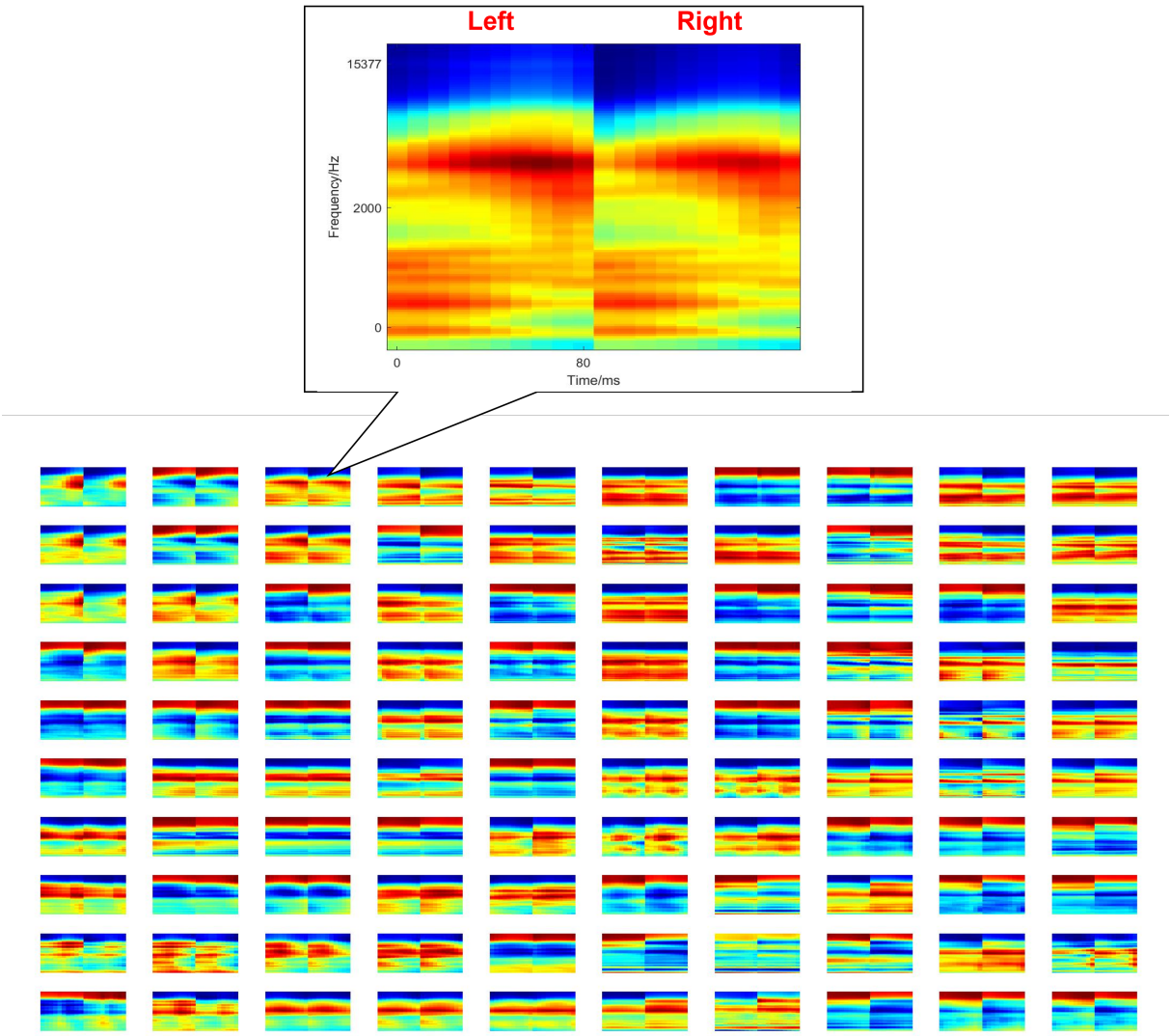


As illustrated in Figure 3.3, the training converged at iteration 45,381, which is not beyond the maximum iteration but still close to it. The slope is also close to zero. Both demonstrate that the selection of maximum iteration and convergence threshold is reasonable.

GASSOM result

The basis functions of trained GASSOM are illustrated in figure 3.4.

Figure 3.4: This figure illustrated the basis functions learned by GASSOM on cochleagram generated from binaural sound. It contains 10 by 10 basis functions. The top figure illustrated one of the scaled basis function from GASSOM. The horizontal axis contains 10 frames (that is 80ms) for both left and right ear and the vertical axis contains 128 frequency bins from 100Hz to 20,000Hz.



As can be seen from the figure, most of the basis functions learned by GASSOM are narrow band frequency modulated features. Disparities between the left and right parts can also be observed, especially for interaural level differences. However, the interaural time difference is less obvious relatively. This is due to the fact that the maximum of interaural time difference for human between is around 0.7ms, which is less than the length of each frame, but there's still slightly temporal difference in some basis functions.

3.4.3 ICA methods

ICA data generation

The generation of binaural stimuli data for training of ICA is similar to GASSOM, but the procedure is slightly different. The data contains 500 batches. For each batch, 200ms monaural sound is randomly picked and then filtered by HRTFs for all 19 different locations. The data are generated at one time and then used for ICA learning, while for GASSOM the process can be accomplished online.

ICA settings

In this work, Independent Component Analysis (ICA) (Hyvarinen et al. 1998, Hyvärinen & Oja 2000), which has been widely employed to solve the Blind Source Separation (BSS) problem, is considered as the baseline for the sparse coding algorithm as it has a long history of application on sparse coding and recent study (Mlynarski 2014) have successfully apply it to encode binaural stimuli and extract spatial features.

Binaural natural sounds can be regarded as a mixture of many auditory features add with additional spatial features. ICA, by assuming these features are independent and non-Gaussianly distributed, seeks to find a set of components that maximize the non-gaussianity among the sources.

Given the input observation samples $X \in R^{n \times m}$, where n is the dimension of the input observation and m is the number of observations, it seeks a de-mixing matrix $M \in R^{n \times n}$, to recover the sources $S \in R^{n \times m}$,

$$S = MX \tag{3.4.2}$$

Each row of the de-mixing matrix M is the basis functions, which is analog to the receptive fields of the auditory neurons, and each row of S is the corresponding neural responses to the input X. The discussion about the biological plausible interpretation is non-trivial and beyond the scope in our work.

The dimension of the input observation data is $128 \times 10 \times 2 = 2560$, where 128 is the number of frequency bins, 10 is the number of frames for each chunk, and 2 stands for left and right channels. To accelerate the computation and reduce the redundancy of the input data, the dimension of the data is reduced to 100 by principal component analysis

(PCA), which preserved more than 99% of the total variance. During the process of PCA, the data is also whitened and sphered so that each component has unit variance.

After PCA, the data is then fed to ICA for learning the basis functions. Fast-ICA algorithm is employed in this study. The learned basis functions are of dimension 100, which are then transformed back to original dimension 2560 using the PCA transforming matrix for visualization.

ICA result

The result for learned basis function are illustrated in figure 3.4. For each small patch, the horizontal axis is the temporal frame, which is 20 (160ms in duration) as the left and right ear stimuli are concatenated, and the vertical axis is the frequency bins.

The ICA basis functions are comprised of a great diversity of different audio features, such as the onset, offset, pitch, check board patterns, and so on. However, only small portion of the basis functions are related to spatial features. This reveals the fact that ICA is not task-specific and result in a miscellaneous audio patterns.

3.5 Result comparison and analysis

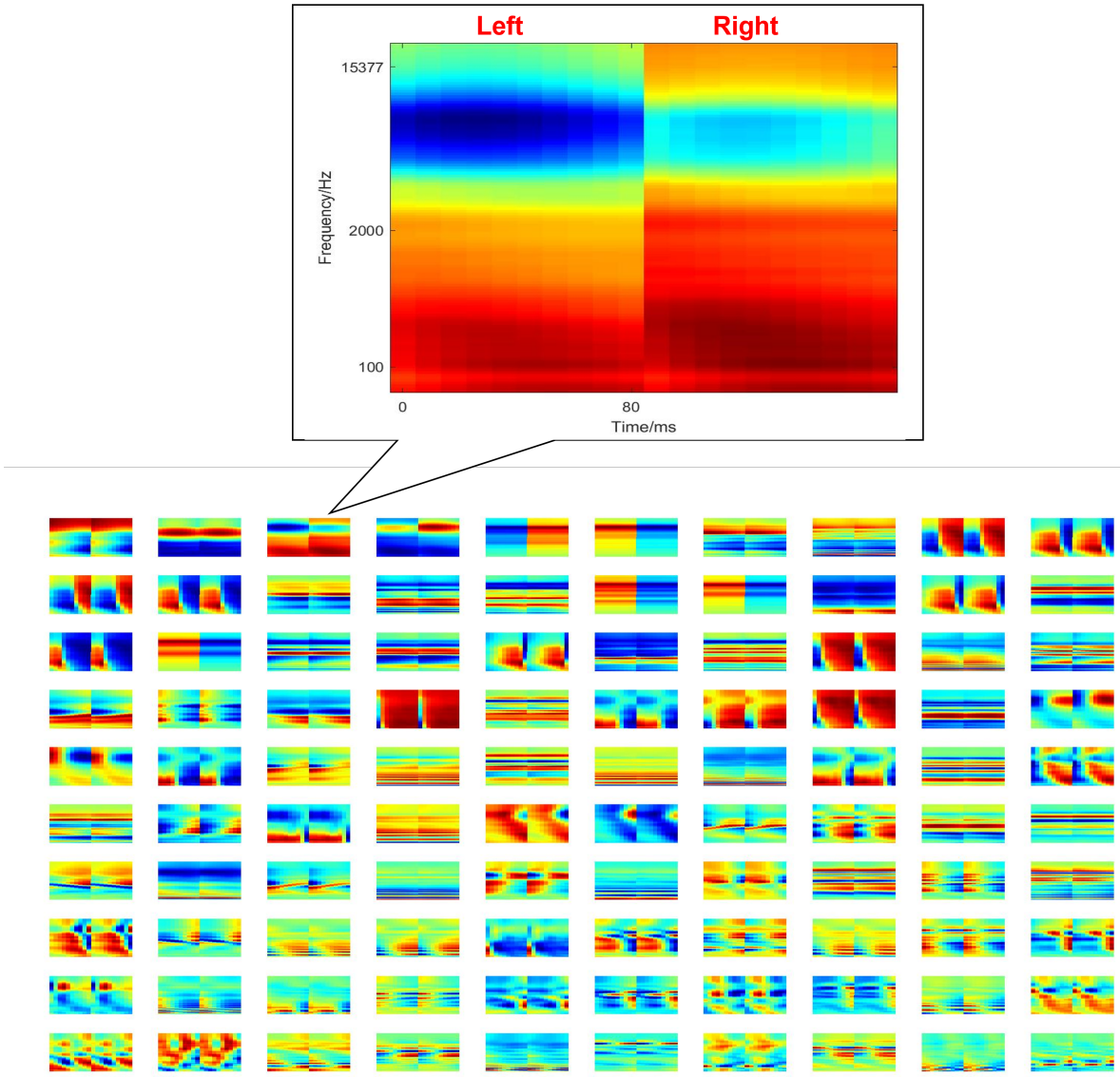
3.5.1 Metrics for comparison

In this study, we selected Fisher information and Disorder Index as two metrics to quantify the result and conduct comparison between GASSOM and ICA basis functions.

Spatial information

As our target in this study is to construct a binaural sound localization model, how informative the spatial features are about different directions. To avoid the influence of subsequent processing, i.e. location prediction with Deep Neural Network employed by our model, we merely analyzed the basis functions rather than computing the localization errors. Statistical methods are considered in this study.

Figure 3.5: This figure illustrated the basis functions learned by ICA on cochleagram generated from binaural sound. It contains 10 by 10 basis functions. The top figure illustrated one of the scaled basis function from GASSOM. The horizontal axis contains 10 frames (that is 80ms) for both left and right ear and the vertical axis contains 128 frequency bins from 100Hz to 20,000Hz.



Topographical smoothness

Another metric considered here is topographical smoothness. The author admitted that topographical smoothness is a self fulfilled property of GASSOM, but it is still meaningful to present the importance of topological structure here and the lack of topological structure would have substantial effect. This is motivated by the facts that many sensory systems in human being exemplify topological structures, such as the retinotopic map and tonotopic map in visual cortex and auditory cortex, respectively. Neighboring neurons share similar patterns would enable the formation of convergent basis from lower level auditory neurons to higher level neurons for process under the hierarchical structure of ascending auditory system. However, this property was not reflected in the Deep Neural Network employed in this work as the order, namely the topological structure of the input neurons in the input layer doesn't make any difference. But in future work, Convolutional Neural Network (CNN) might be utilized to decode the location, where the filters in the convolution layer make use of the topological structure of the input. Besides, better topographical smoothness allows better visualization the researcher.

Measurement

To quantify the directional information contains in basis functions learned by ICA and GASSOM, we compute the fisher information for each basis function with respect to different locations. Besides, the topographical smoothness is also considered, as the tonotopy structures have been discovered in the auditory cortex. The topographical structure can also be utilize to improve the acquirement of the spatial information.

3.5.2 Fisher Information

Fisher Information Definition

Fisher information quantifies to what extent a hidden parameters can be estimated from the observations. In this case the hidden parameters θ is the location of the sound and observations s are the spatial features extracted by the basis functions, which is computed as the squared projections of the binaural input stimuli onto the corresponding basis function.

Here we assume that each the distribution of the spatial features is distributed according to a Gaussian distribution with unit variance and centered at the μ_θ, which is determined by the corresponding location,

$$p(s|\theta) = \mathcal{N}(s|\mu_\theta, 1) \tag{3.5.1}$$

Then the fisher information can be reformulated as,

$$F(\theta) = (\frac{d}{d\theta} f(\theta)^2) \tag{3.5.2}$$

Fisher information quantifies to what extent the direction can be estimated from the projection of the basis function, the greater Fisher information is, the more informative the basis functions are about locations.

Fisher Information result comparison

The fisher information for each basis function with respect to different locations are computed, and then averaged over the locations. The result is shown in Figure 3.6.

The histograms of the Fisher information illustrate that basis functions learned by ICA contains little information about the locations, while basis functions learned by GASSOM are much more informative about locations relatively.

One-way ANOVA is also conducted, and the p-value is close to zeros, which means there's statistically significant difference of Fisher Information between GASSOM and ICA basis functions. We can observe from the figure that GASSOM basis functions are more informative about locations than ICA basis functions. The corresponding box plot is shown in figure 3.7.

Therefore in the following phase of decoding the locations with spatial features extracted by the sparse coding algorithms, GASSOM is expected to outperform ICA.

3.5.3 Disorder Index

Disorder Index definition

Disorder index is utilized to quantify the topographical smoothness. It measures the similarity between neighboring basis functions. Therefore, smaller disorder index stands

Figure 3.6: The histogram of Fisher information for each basis function on directions is demonstrated for both GASSOM and ICA. The greater the fisher information is, the more informative the basis functions are about directions. It's obvious that fisher information for GASSOM is much larger than ICA, therefore GASSOM basis functions are more informative about directions than ICA.

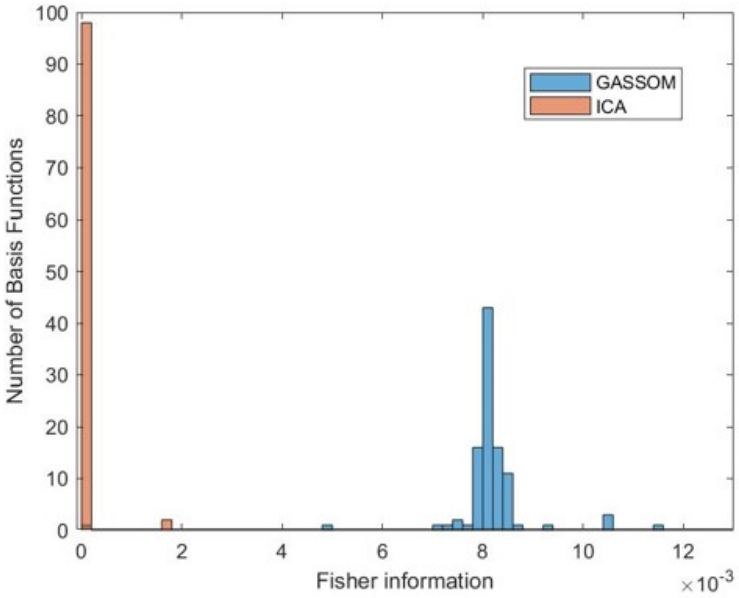
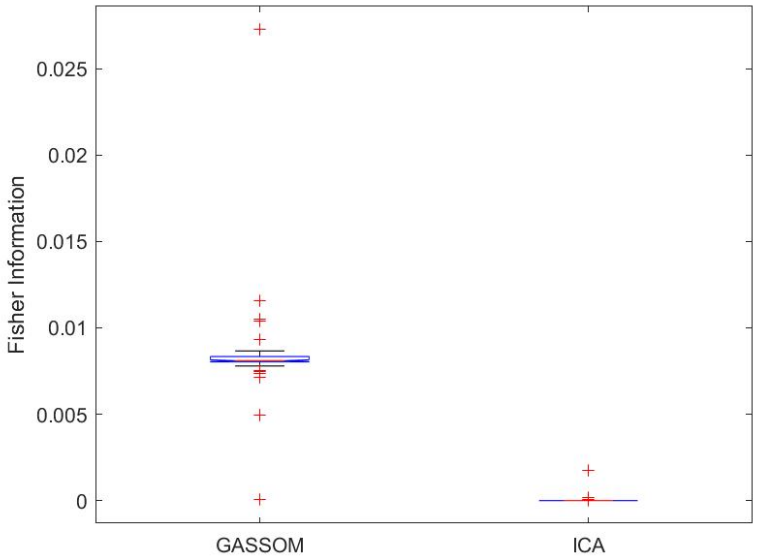


Figure 3.7: The left box is box plot for GASSOM and Right one is for ICA. Significant difference can be observed from the plot.



for better topographical smoothness. It is defined as the linear fitting residual errors of the current basis function with its neighboring basis functions using least squared algorithm.

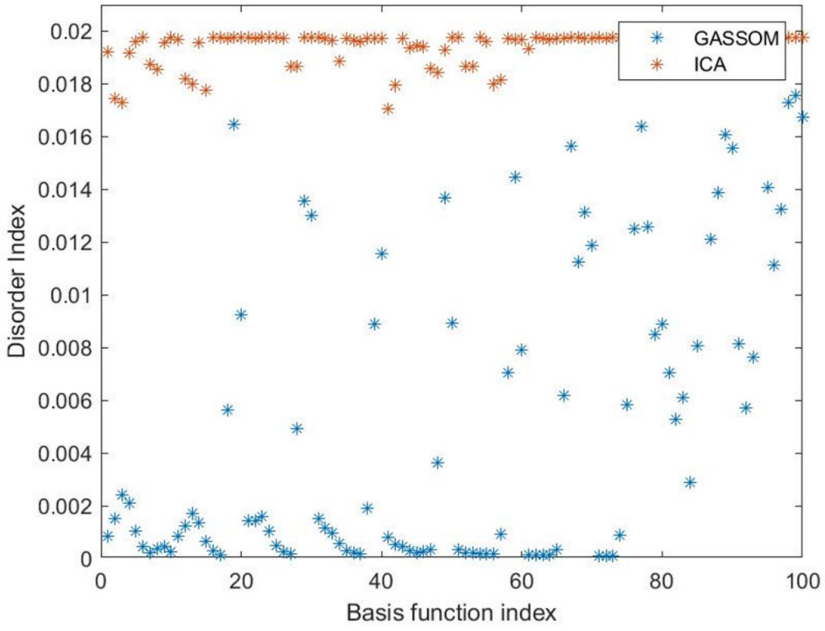
$$DI(i) = \sqrt{\frac{\sum_j h(i,j) r(j)^2}{N_{NB}}} \tag{3.5.3}$$

where $r(j)$ is the residual error of linear regression at j and N_{NB} is the number of units within the neighborhood window. $h(i,j)$ is the neighborhood function defined by the distance between i and j. For simplicity, we only consider the neighboring basis functions, and the results are averaged over all the neighborhood functions because for basis functions at edge or corner of the map, the number of neighboring basis functions are less.

Disorder Index result comparison

The results are shown in figure 3.8. One-way ANOVA is calculated between GASSOM and ICA disorder indices, and the p-value is less than 0.001, which means the GASSOM basis functions show significantly better topographical smoothness comparing with ICA basis functions.

Figure 3.8: The DI for ICA basis functions are concentrated around 0.02, while the DI for GASSOM basis functions are dispersed lower than 0.02. Smaller disorder index means better more similarity between neighboring basis functions. Therefore, GASSOM shows much better topographical smoothness than ICA.



It is evident from the plot that GASSOM basis functions have less disorder index in

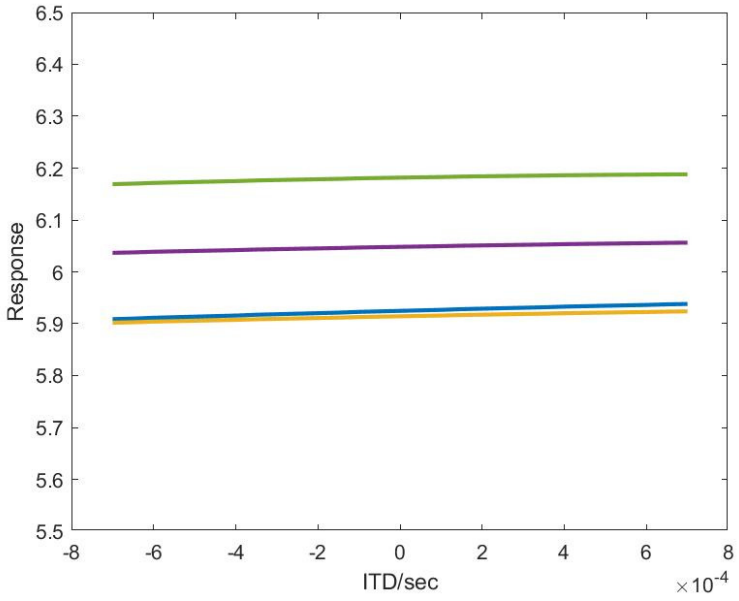
comparison with ICA. Therefore, GASSOM has much better topographical smoothness compared with ICA. This result is also straightforward illustrated in the visualization of basis functions.

It should also be noted that repetition of disorder index distribution is also observed, which is because the illustration of the plot is one-dimensional, while the structure of GASSOM is two-dimensional.

3.5.4 Sensitivity to ITD

During the computation of cochleagram, only the energy within each frame was utilized and the phase information are omitted, that is the finer temporal cues were neglected. To validate the existence of ITD information within each basis function, we computed the tuning curve for each basis function with respect to Gaussian white noise with different ITD. We selected the 5 most 'sensitive' basis functions to ITD and illustrated in Figure 3.9.

Figure 3.9: Tuning curve of ITD for all basis functions are calculated which is based on the projection Gaussian white noise with different Interaural Time Differences (ITD). To quantify the sensitivity, the variance of each tuning curve was calculated, and the 5 tuning curves that most sensitive to ITDs were illustrated (only 5 are selected for clearer illustration). Different colors stand for different basis functions. However, as illustrated in the figure, the tuning curves are flat across the ITDs, which means the basis functions are little sensitive to the ITDs.

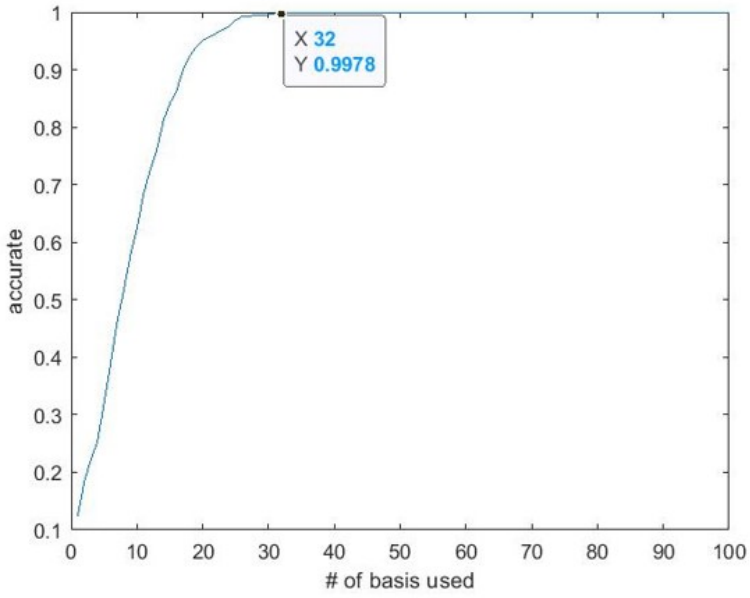


As demonstrated from the figure, the variation of response across different ITDs is very small, which means that the basis functions are little sensitive to ITDs. This is a natural result from the neglect of phase information. Another reason is due to the duration of each frame. In this study, we selected 8ms as frame length. Therefore, any time difference less than 8ms is not visible on the cochleagram. However, the head size is usually around 20cm, and the Interaural Time Difference is less than $700\mu s$. Therefore the ITD is hardly detectable from the figures. Besides, different slight slopes can also be observed from the figure, which means that the basis functions might be sensitive to different ITDs. However, these ITDs are beyond $700\mu s$ and is hard for biological explanation. Though there are hypothesis about bat auditory system that they amplify the ITD for better resolution, the linkage between our result and the hypothesis is still far-fetched. In the future work, a much shorter frame duration can be selected to reflect accurate ITD can be employed to solve the high-frequency problem.

3.5.5 MAE vs Number of basis functions

An alternative way to applied to quantify the spatial information contains in the basis functions. Different number of Deep Neural Network (DNN) were trained with increasing number of basis functions for ICA. Localization accuracy rate was calculated for different conditions. The result is illustrated in Figure 3.10

Figure 3.10: This figure illustrates the localization accuracy versus the number of basis function utilized to train the DNN. The basis functions are sorted according to Descending Fisher Information. An increase of localization accuracy can be observed with the increase number of basis function employed. It almost reached highest accuracy when the number of basis functions employed are more than 32.



It can be seen from the figure that as the number of basis function achieved certain number (around 32), the localization accuracy almost saturated. This means that 32% of basis functions contains spatial information, while the rest is not related to spatial location decoding.

3.5.6 Influence of background noise

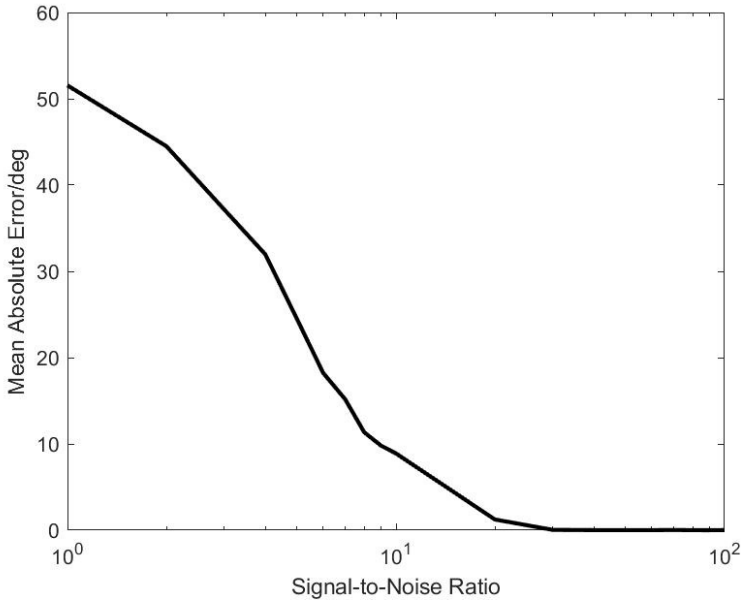
In previous study, the localization task was evaluated under anechoic condition. We also conducted a preliminary experiment when background Gaussian White Noise (GWN) appears. The noise ratio was quantified using Signal-to-Noise Ratio (SNR), which is

defined as the ratio of signal power and noise power and transformed to decibels.

$$SNR = 20 \times \log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{3.5.4}$$

Localization error was quantified using localization accuracy rate, which is the ratio of correct prediction of the sound sources locations. The result is illustrated in Figure 3.11

Figure 3.11: This figure illustrates the MAE in degree with respect to signal-to-noise ratio in dB. Significant decrease of localization error is demonstrated when the SNR increases. The localization error is close to zero when the SNR is greater than 30dB.



It can be seen from the figure that when the SNR reached above 3, the localization error is almost close to zeros. However, when the noise occurs, the localization accuracy is heavily distorted. Therefor, in the following study, we only considered anechoic condition.

3.6 Conclusion and Discussion

In this chapter, we introduced the paradigm of the binaural localization model and illustrated how the sparse coding of binaural sound is accomplished. For the localization model, we focused on the first stage, which is spatial feature extractor as the second stage is less need to be determined. As GASSOM has been successfully applied to encode audio stimuli, we determined to investigate its application on audio stimuli. ICA as another popular sparse coding algorithm, was proved to successfully produce basis functions that

would extract spatial features from binaural stimuli. Therefore ICA is selected as baseline algorithm and benchmark with GASSOM.

Comparison was conducted between GASSOM and ICA using the metric and Fisher Information and Disorder Index, which quantify the directional information and topographical smoothness, respectively. The results show that GASSOM is a better choice, either as a more informative spatial feature extractor, or in the consideration of topographical structure. The second conclusion is more straightforward as it's a natural result of the training process that embedded by the self-organizing topological structure in GASSOM.

The first conclusion is more abstract. Our explanation for is that GASSOM is capable to extract invariant features within each episode, and in this study, the spatial information is the invariant feature while other features carried by the monaural speech is casted aside. It is more concentrated on extraction of location-related features, though it may lead to slight redundancy, which can be observed from the similarities between the neighboring basis functions in GASSOM. But this is inevitable.

While for Independent Component Analysis, there's no specific task, and therefore it attempts to extract features that are less Gaussian and independent to each other. It is concentrated on the diversity of the audio features, while spatial information only takes small portion of the diversity. Even though the resulted basis functions would incorporate more information about the content of the binaural sounds, it is not preferred in the task-specific case. Besides, ICA require the full access to all data for training (we conclude based on the Fast-ICA algorithm, which is utilized by (Mlynarski 2014), though the author admits that a modified version of ICA can be applied online), while GASSOM can be trained online.

Therefore, GASSOM is selected as the sparse coding algorithm to construct the spatial feature extractor in the following study.

Chapter 4

Determining the optimal parameter for GASSOM training

4.1 Introduction

After comparison, GASSOM is selected to be spatial feature extractor for the binaural localization model. During the training of GASSOM, several parameters should be determined for better localization performance, among which the most important parameters are map size, chunk length and chunk shift. These three parameters are all related to the dimension of the basis functions learned by GASSOM. We will also related those parameters to its biological explanation.

To quantify the result, we selected Mean Absolute Error (MAE) and Best Matching Times (BMT) as the dependent variables.

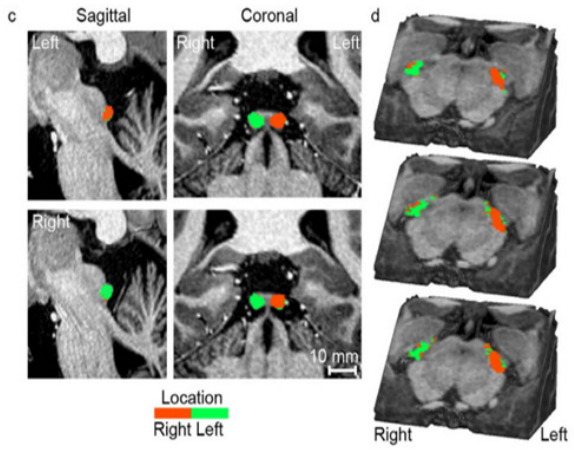
4.2 Methods

4.2.1 Variables

Map size

Map size stands for the number of basis functions in GASSOM 2D topo space. Usually, the map shape is square in order to reduce the influence of map edges and reach the balance on different topographical directions. A greater map size give rise to the information contained in GASSOM, but it will lead to inefficiency as illustrated in our previous example. It also leads to data redundancy and some of the basis functions may

Figure 4.1: The panel demonstrate the topographical distribution of the basis functions along the 10 by 10 map. Map size corresponds to the number of basis functions on the map. Different basis functions corresponding to different directions and larger map size would give more space to encode the 3D space. It's analog to the number of neurons in auditory cortex, but of different magnitude.



not have the chance to be selected as winner during the whole course of training. In this study, map size of 8×8,10×10,16×16,20×20 are considered.

The map size also has its biological meaning. It corresponds the number of neurons in audio cortex, though the magnitudes between them are different. The demonstration of map size and its corresponding biological linkage is illustrated in 4.1.

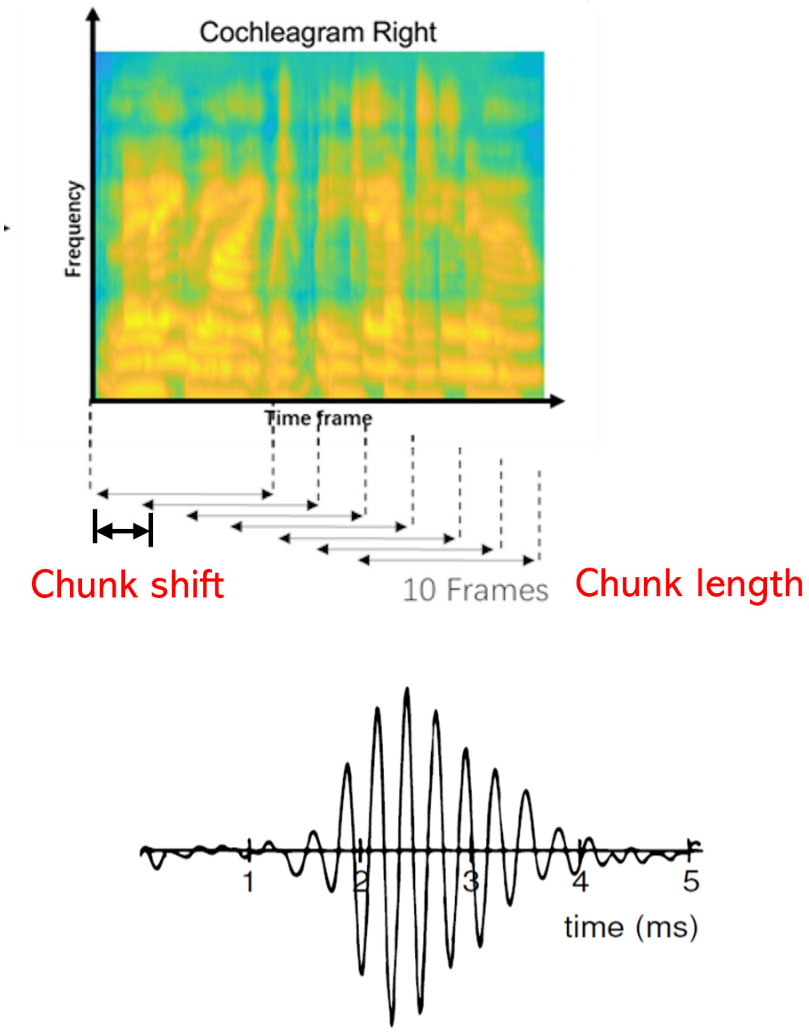
Chunk Length

the number of frames contains in each chunk during segmentation. It also determines the length of basis functions in GASSOM. Larger length will lead to more information captured by the basis functions, but it will also reduce the temporal resolution on the audio signals and increase the redundancy.

In this study, chunk length of 5, 10, 20 frames (corresponding to 40,80,160ms) are considered.

As illustrated in 4.2, the top panel is the definition of chunk length and chunk shift for the cochleagram. While in the bottom panel, the response of auditory fiber in cat brain measured with revcor function (de Boer 1978).

Figure 4.2: The top panel shows a cochleagram of 200ms speech and illustrates the definition of chunk length and chunk shift, where the chunk length is set to 10 frames (80ms) and the chunk shift is set to 1 frame(8ms) for example. The bottom panel illustrates the corresponding biological implication of chunk length, which is reconstructed impulse response representation of a fiber retained using reverse correlation (revcor) technique from cat auditory neurons (de Boer 1978). The auditory neuron response has finite length, although the duration of auditory fiber in cat shown here is only 5ms, which is much less than human being as a result of species difference.



Chunk Shift

the number of frame shift between successive chunks during the generation of data. A smaller shift will lead to better temporal slowness, which is one of the underlying principles, but too small shift will lead to data redundancy and computational inefficiency. Temporal slowness in our case is the slow movement of the sound sources relative to the listener.

In this study, chunk shift of 1,2,4 frames are considered. Chunk length and chunk shift cannot be decoupled separately, and therefore considered together. The interaction effect between Chunk Length and Chunk Shift is also tested.

4.2.2 Metrics

Mean Absolute Errors (MAE)

Mean Absolute Errors (MAE in degrees) is defined as the mean absolute difference between the predicted location and ground truth location. It aims to quantify the influence of different parameters on the localization accuracy. Smaller MAE stands for better localization accuracy. Even though it's not essential for

Best Matching Times (BMT)

Best matching times is defined as the number of time for each basis function being selected as the winner node during the training of GASSOM. It is designed to quantify the efficiency/utilization rate of the learned GASSOM. Too large or too small for the Best Matching Time are not appropriate. Large value means the basis function is always selected as winner during training, which makes the node hard to converge. While if the value is too small, it means the basis function is rarely selected as winner node, which means it's redundant.

It should be noted that Best Matching Times should be normalized with respect to the number of maximum iterations. After which the mean BMT is normalized to 1. We only need to consider the variance of the BMT.

4.2.3 Procedure

GASSOM Training

The GASSOM training phase follows the similar procedure as described in chapter 3. Binaural sounds are generated by convolving 200 ms monaural speech pieces that are randomly picked from TIMIT database with HRTFs from KEMAR HRTF database, where the direction of the HRTF is also randomly picked from -90 degree to 90 degree evenly. Cochleagram is then computed from the synthesized binaural sounds, which is afterwards divided into small chunks according to the variable configuration, and then used to train the GASSOM with different map size accordingly. The GASSOM training session is repeated for 50,000 iterations, during which the number of times for each basis function being selected as winner is counted.

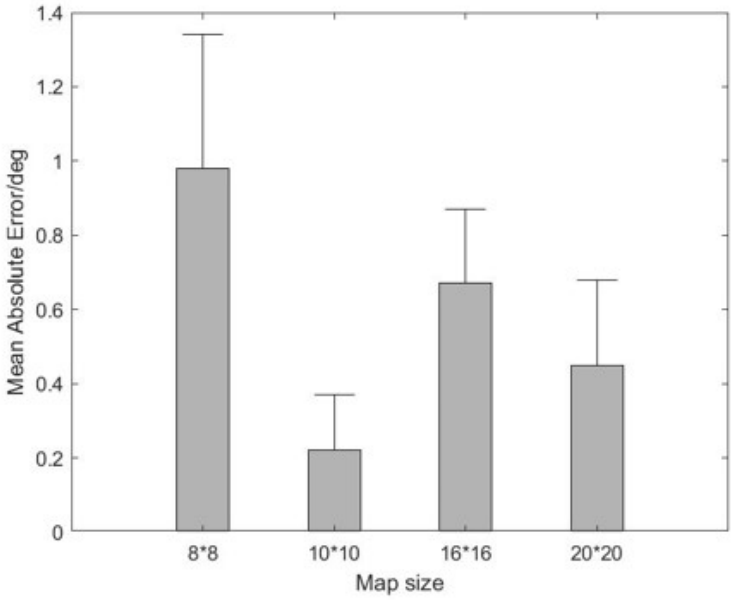
DNN Training

After the training of GASSOM, 200ms Gaussian white noise (GWN) is then selected for training of DNN. It has the advantage of broad bandwidth and easy generation. GWN is also filtered by KEMAR HRTFs for all different locations ranging from $-90°$ to $90°$. Cochleagram is then produced and divided into small chunks. The chunks are passed through GASSOM by computing the squared projections of it onto each basis functions, the output of which is then used as input for the training of DNN. The location is transformed into categorical indices ranging from 1 to 19, which correspond to the directions of $-90°$ to $90°$ degree by every $10°$, respectively.

Model Testing

the testing sounds are generated in the same way as those for DNN testing, which is also spatialized Gaussian white noise. The probability for different location is obtained by passing the sounds through the GASSOM and DNN, and then averaged over the chunks within the same signal. The location with the greatest posterior probability is selected as the predicted location. As the ground truth location is known as the direction corresponding to the selected HRTF, the mean absolute error in degrees between the ground truth location and predicted location can be calculated and average over all the directions.

Figure 4.3: The horizontal axis is map size and vertical axis is MAE in degree. The MAE is least when map size is 10 by 10, that is the localization is most accurate.



This process for each variable is repeated for 10 times for analysis of variance, and increase the robustness of the result.

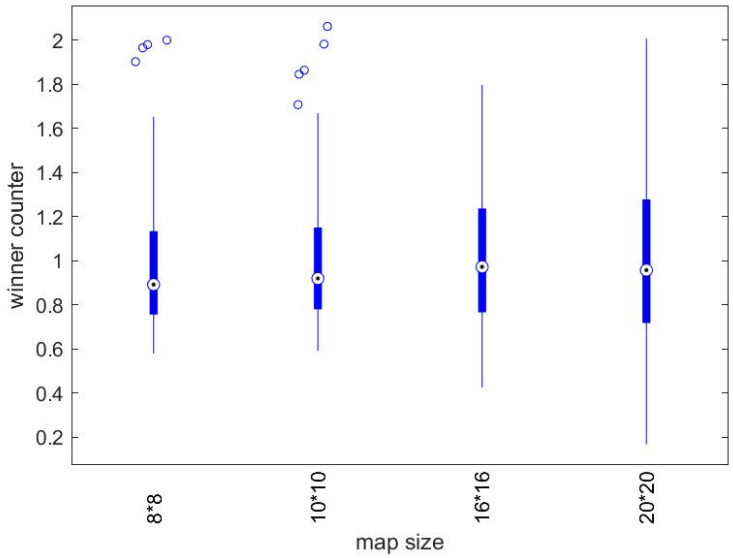
4.3 Result and analysis

The interaction of the map size and chunks are less correlated, so the effect of map size is tested separately, while the effect of chunk length and chunk shift is analyzed together.

Figure 4.3 demonstrate the Mean Absolute Errors GASSOM localization model with different map size. One-way ANOVA is conducted and $F(3, 16) = 8.02$, the p-value is $p = 0.0017 < 0.01$, which demonstrate that effect on map size on localization accuracy is statistically significant at level of 0.01. It can be seen from the figure that the MAE reach least value when the map size is 10 by 10.

However, comparable performance is also achieved when the map size is 20 by 20. To determine which is better, we also collect the Best Matching Times (BMT) as defined before to quantify the redundancy of the map. If the BMT of some basis functions is much smaller compared to other basis functions, it means that they are rarely selected as winner during the training. The result of the BMT for each map size is demonstrated in figure 4.4. It has been standardized because different map size contains different number

Figure 4.4: Box plot of BMT is illustrated here, where the horizontal axis is map size and vertical axis is BMT averaged over all the basis functions. The solid line is standard deviation and the bottom and top edge of the rectangle illustrates the 25 and 75 percentile respectively. The dotted circle is the median of the corresponding data. As the BMT has been normalized with respect to number of iterations, the mean value is 1, but the variance varies, which quantifies to what extent the balance is in training phase. Smaller STD means the nodes are more equally likely to be selected as winner node, and corresponds to better balance. Therefore 10 by 10 map size has a better balance.



of basis functions. This process is realized as follow,

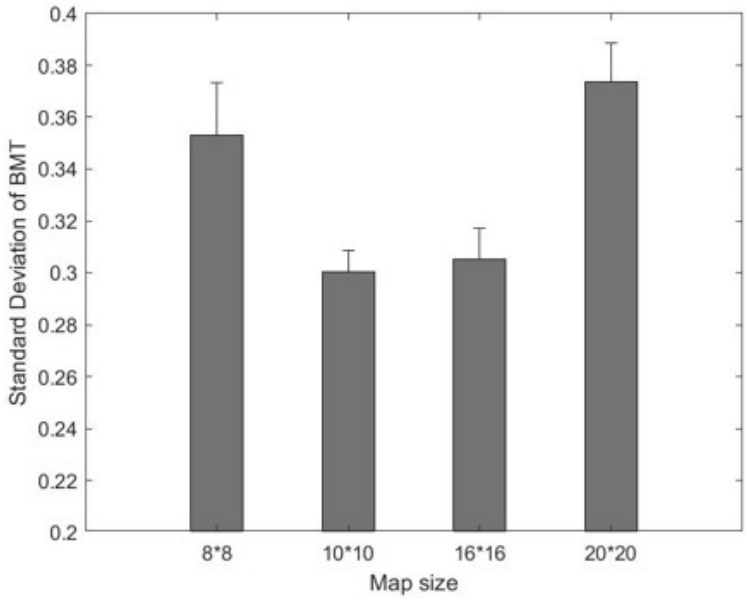
$$BMT_{standard} = \frac{BMT * N_{basis}}{N_{sample}} \tag{4.3.1}$$

Where N_basis is the number of basis functions for each map size, and N_sample is the total number of samples during the GASSOM training. Therefore, the BMT averaged over all basis functions is normalized to 1 as shown in figure 4.4.

The standard deviation of the BMT for each map size is also calculated for more straightforward illustration and is shown in figure 4.5. Smaller standard deviation means the chances for each basis function to be selected as winner is more likely to be equal.

From figure 4.5, map size 10×10 and 16×16 have more balanced BMT for each basis functions. Considering both BMT and localization accuracy, map size of 10×10 is a

Figure 4.5: To give a more straightforward illustration of the Standard deviation (STD) of the BMT, the STD for BMT of different map size is shown here. The variance is relatively smaller when the map size is 10 by 10. Smaller STD stands for better balance/map utilization rate and therefore map size 10 by 10 is a relatively better choice.



selected in the following study.

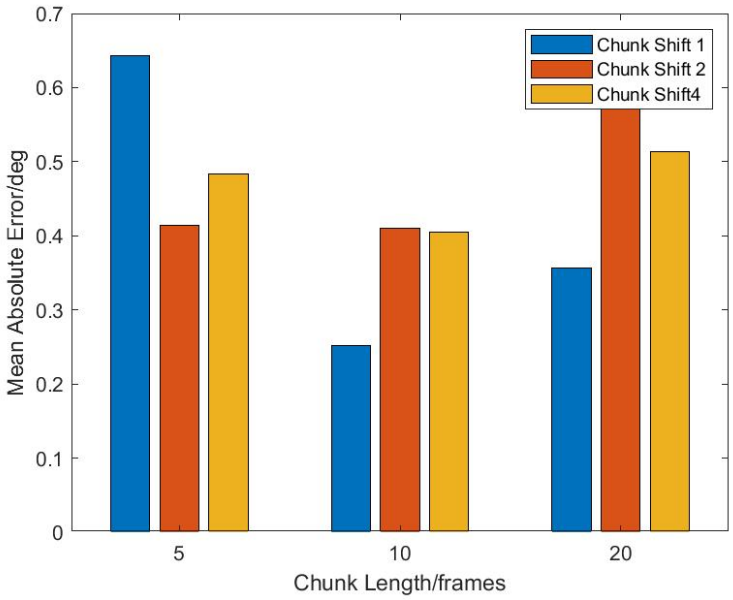
4.3.1 Chunk Length and Shift

Figure 4.6 demonstrate the bar chart of mean MAE for different configurations of chunk lengths and chunk shift.

It is illustrated from the figure that the localization accuracy is highest (least Mean Absolute Errors) when the chunk length is 10 frames, and the chunk shift is 1 frame.

Two-way ANOVA is conducted to analyze the result. For chunk length, $F(2, 28) = 8.95$, and p-value is less than 0.001, which means the chunks length has a statistically significant influence on the localization accuracy. For chunk shift, $F(2, 28) = 1.06$, and p-value $p = 0.3526$, which states that there's no significant effect of chunk shift on localization accuracy. For the interaction of chunk length and chunk shift, $F(4, 86) = 6.86$, $p = 0.0001$. This reveals the significant effect for the interaction of chunk length and chunk shift, which can be observed from the figure that when the chunk length is 5 frames, the localization accuracy is least when the chunk shift is 1 frame, while for other cases, the localization accuracy is highest when the chunk shift is set to 1 frame. As the

Figure 4.6: This figure illustrate the MAE for different configuration of chunk length and chunk shift. Different colors stand for different chunk shift. The MAE is smallest when the chunk length is 10 frames and chunk shift is 1. Therefore we selected chunk length to be 10 frames and chunk shift to be 1 for better localization performance.



main effect of chunk length is significant, we first select the chunk length to be 10 frames. Afterwards, the optimal chunk shift is selected as 1 frame due to the significant effect of the interaction effect.

4.4 Conclusion and discussion

The selection of map size is a complex result on the localization accuracy and utilization. A larger map size will give more space to capture different features, but it also has the disadvantage of inefficiency and low utilization rate. Besides, the localization accuracy drops when the map size getting too large. Our explanation for this phenomenon is that with the increase of map size, some basis functions are rarely selected as winner, and are therefore less informative about the spatial location. However, during the training of DNN, all the basis functions are treated equally. As a result, the performance is degraded by these less utilized basis functions.

For the selection of chunk length and chunk shift, the model performs best when the chunk length is set to 10 frames, which corresponds to 80ms in duration. When the

chunk length is too small, the localization accuracy drops because it's not long enough to capture all the information required to decode the location of the sound. On the other side, when the chunk length is too large, it will lead to computational redundancy in the sparse coding of GASSOM and cause degradation on the localization performance.

For chunk shift, a relative smaller value will lead to better performance in localization accuracy. Otherwise, it will reduce the temporal resolution of spatial features and eliminate the small variations of the audio signals, which might play an important role in the deduction of sound location. Smaller chunk shift also conforms the temporal slowness assumption of GASSOM, which results from the fact that natural stimuli in environment always varies slowly. When the chunk shift is too large, the features will vary vastly and cannot be captured by GASSOM.

However, it should be noted that the effect of chunk shift is dependent on the chunk length. When the chunk length is too small, the conclusion may not be applicable. Further study may be conducted to explore the reason for it.

Chapter 5

The effect of waveform properties on localization accuracy

In the next two chapters, we compare the performance of GASSOM based binaural sound localization model with empirical data on human subject on different tasks. The paradigm is shown in figure 5.1

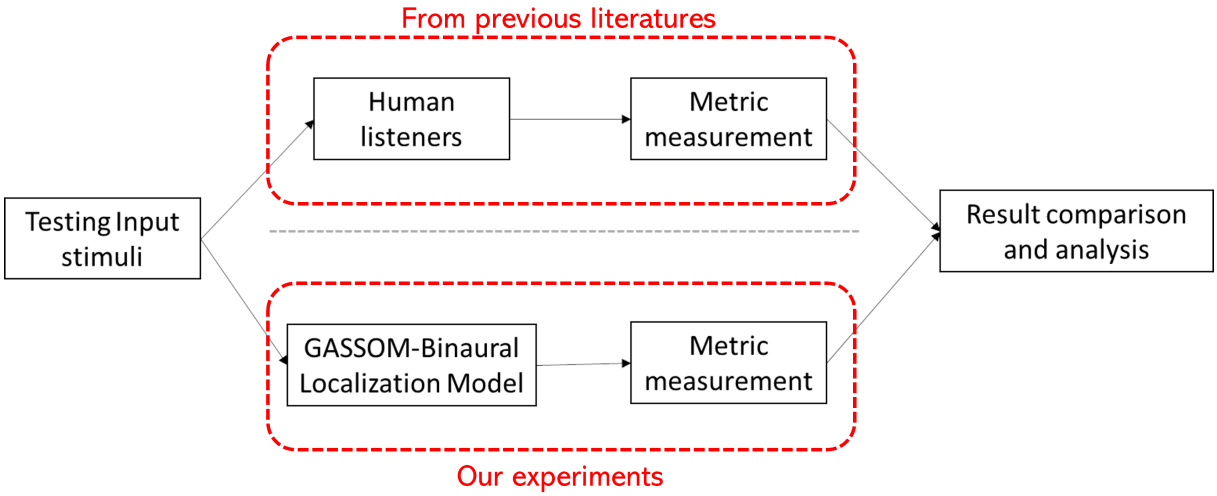


Figure 5.1: This flowchart demonstrates the paradigm of the following two chapters. Given the testing input stimuli that similar to previous work on human being experiment, we pass it through the GASSOM-based computational model and calculate the same measurement as the authors employed in theirh work. Then we compare the performance with the empirical data using the same quantification metric and analyzed the result. Similar performance was achieved by our computational model on several aspects.

5.1 Introduction

Psychoacoustic experiment on binaural sound localization on human listener has been investigated for a long time. It has been proven that the localization accuracy of human being is heavily dependent on the spectral and temporal properties of the audio signals. William Yost and his colleague have conducted a series of psychophysical experiment to explore the effect of many factors, such as center frequency, bandwidth (Yost & Zhong 2014), duration, sound level (Yost 2016), on localization performance. These work lead to deeper understanding on the localization mechanism in human auditory system.

However, little concentration has been focused on the effect of those parameters on the localization accuracy of bio-inspired computational model. It has been proved that those parameters have significant effect on the localization performance of computational models.

In comparison with the empirical data on human being, we conducted similar experiment on the GASSOM based binaural sound localization model, to investigate the effect of center frequency, bandwidth, and sound duration on localization performance. The sound level is omitted by during the computation, the noise level is normalized.

It has been demonstrated that human auditory system computes the spatial information within different narrow bands and integrate them together to decode the location of the sound. Audio stimuli with different bandwidth would contain different amount of spatial information. What's more, the different narrow band also plays different roles in the localization task, so it's of interesting to explore the effect of center frequencies on localization performance.

This study will not only fill this research gap, but also proving the capability of our GASSOM based computational model providing comparable performance as human being in sound localization tasks. This model therefore can be used as an alternative to the human subject in many psychoacoustic experiments.

5.2 Methods

5.2.1 Variables

Center frequency

the arithmetic/geometric mean of the lower cut-off frequency and upper cut-off frequency of the audio stimuli. In this study, center frequency of 250Hz, 2000Hz and 4000Hz are considered to keep in consistent with Yost's experiment on human being.

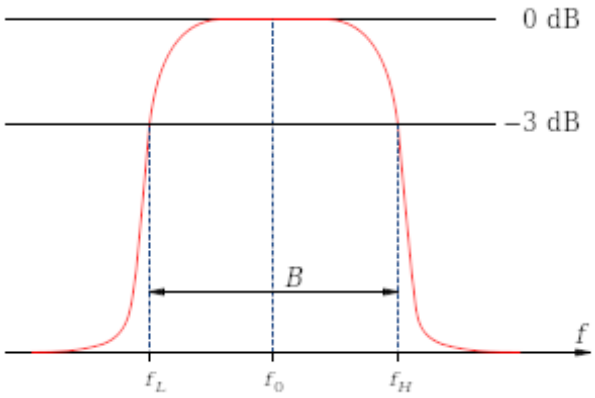
Bandwidth

the frequency difference between the upper 3dB cut-off frequency and lower 3dB cut-off frequency as shown in Figure 5.2. Bandwidth is always represented with octave as unit, which is defined as the logarithm of upper cut-off frequency divided by lower cut-off frequency. In this study, bandwidth of 1/6 octave, 1/3 octave, 1 octave and 2 octave are considered to keep in consistent with Yost's experiment on human being.

Sound duration

the temporal duration of audio signals in milli-seconds. In this study, 25ms, 150ms and 400ms are considered. If the sound duration (25ms here) is not long enough, zero padding is applied to the stimuli for subsequent process.

Figure 5.2: Band pass filter are illustrated here. B stands for bandwidth and f_0 stands for center frequency. f_L and f_H stand for the lower and higher 3dB cut-off frequency.



5.2.2 Metrics

For comparison with empirical data, we employed the same metric, Root Mean Square (RMS) errors, to quantify the localization accuracy of our computational model. It is defined as follow equation,

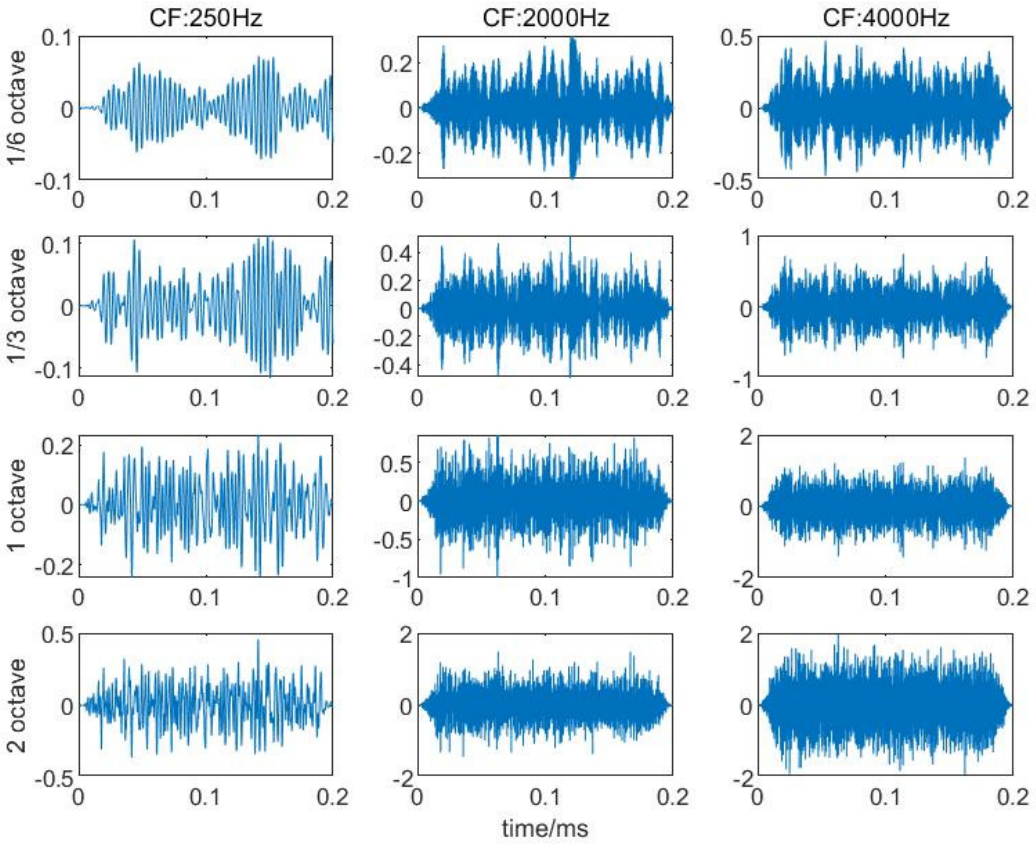
$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(\theta_{pred} - \theta_{true})^2}{N}} \tag{5.2.1}$$

Where N is the number of test samples, θ_{pred} is the predicted location in degree, and θ_{true} is the ground truth location in degree. The RMS is averaged over all locations for comparison.

5.2.3 Stimuli generation

The spectral-varying stimuli are generated by filter the 200ms spatialized Gaussian white noise with 20ms squared cosine rise-decay time, which is also synthesized with MIT KEMAR HRTFs at ear level horizon plane from $-90°$ to $90°$ degree by every $10°$ in azimuth, with 4-pole Butterworth filters designed for different center frequencies and bandwidths. The filtered audio signals are then processed to generate cochleagram and fed into the model to decode the location. Demonstration of the waveform signals are shown in figure 5.3.

Figure 5.3: The figure illustrates stimuli with different center frequency and bandwidth by filter 200ms Gaussian White Noise with Butterworth Filters that designed for different combination of center frequency and bandwidth. Different column stands for different center frequency and different row stands for different bandwidth.

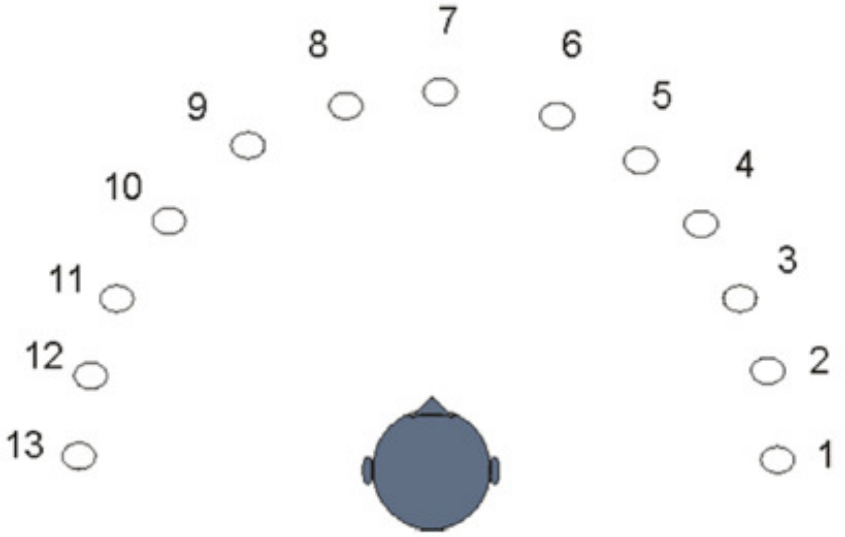


The duration varying stimuli are generated by changing the duration of the Gaussian white noise, which is not illustrated here.

5.2.4 Review of Yost's experiment

In Yost's experiment, 13 loudspeakers were utilized at different locations from $-90°$ to $90°$ with increment of $15°$ to produce sound sources at different locations in an anechoic chamber. All the speakers were numbered from 1 to 13 as illustrated in Figure 5.4.

Figure 5.4: The figure illustrated the setup of Yost et al.'s experiment on localization performance of human listener (Yost et al. 2013). 13 loudspeakers spaced at different locations were used in the experiment to construct spatialized sound sources by playing through the speaker at certain location. To determined the perceived location of the listener, the index of the speaker was reported as indicator.



The subjects were asked to report the index of the speaker as indicator for the perceived sound location. The RMS in degree between the ground truth location and reported location was computed for each trial.

5.2.5 Procedure

The GASSOM localization model is trained the same way as previous chapters. The spatial feature extractor is constructed by sparse coding binaural natural speeches with GASSOM, and the DNN is trained with spatialized broadband Gaussian White Noise. For training phase we employed the full frequency bandwidth because it simulates the human listener with normal hearing abilities. The spectral-temporal property of the stimuli only varied according to previous description in the testing phase.

Center Frequency and Bandwidth

Center frequency and bandwidth are coupled with each other; therefore both should be investigated together. For the testing phase, the stimuli are generated as mentioned above and the RMS about localization is computed for each condition. It is repeated for 8

times for each condition for further analytic for comparison with Yost and Zhang's (Yost & Zhong 2014) empirical data on human being.

Sound Duration

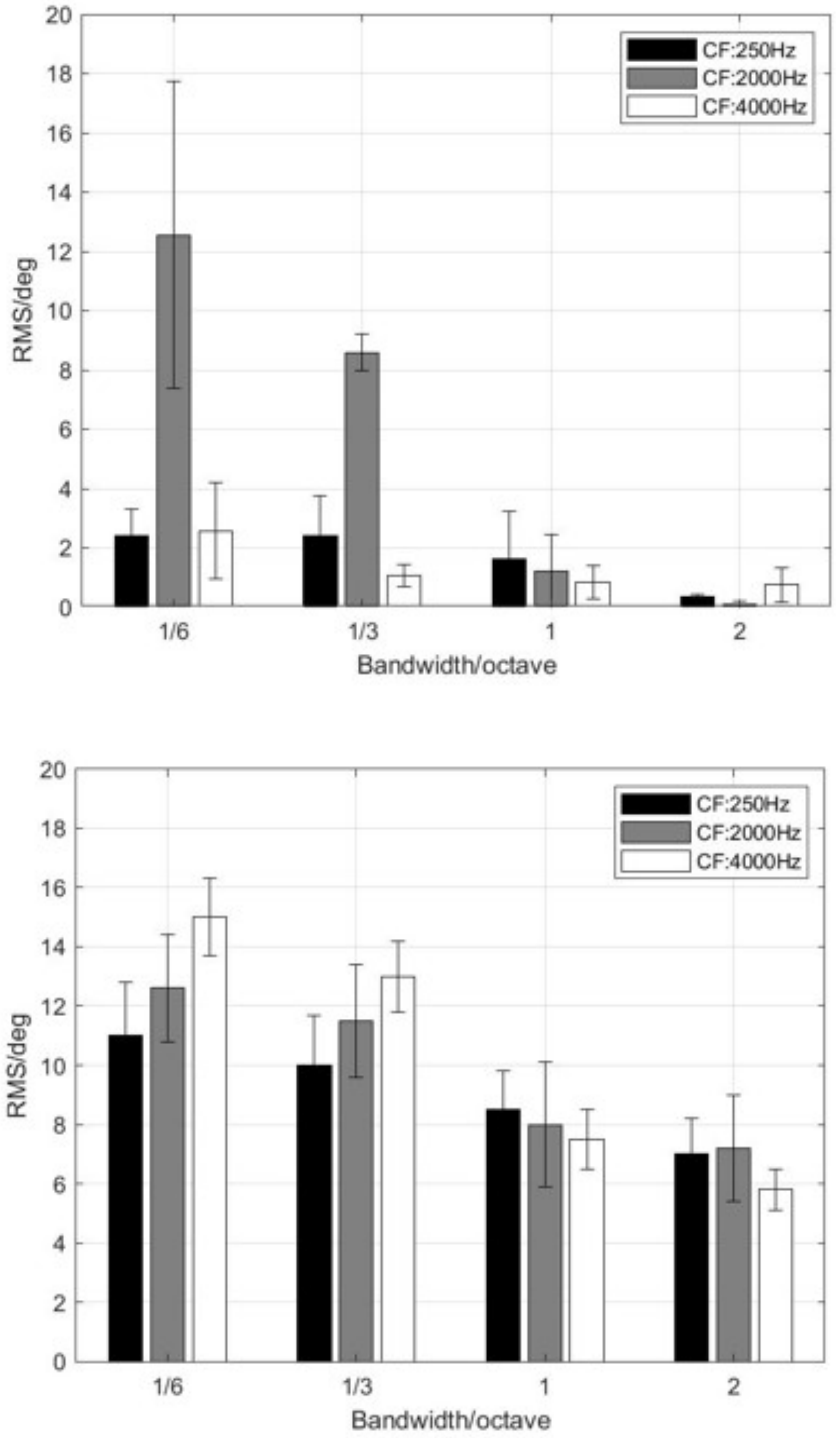
The effect of sound duration on localization accuracy is tested under different conditions, that is for different center frequency and bandwidths, to test the interaction effect of those parameters. In this study, center frequency of 250Hz, 2kHz and 4kHz are selected in consistent with Yost's experiment (Yost 2016) for comparison. 1/10 octave and 2 octaves are also selected as representation for narrow bandwidth condition and broad bandwidth condition, respectively. Different conditions are repeated for 8 times for analysis of variance.

5.3 Result and analysis

5.3.1 Effect of center frequency and bandwidth

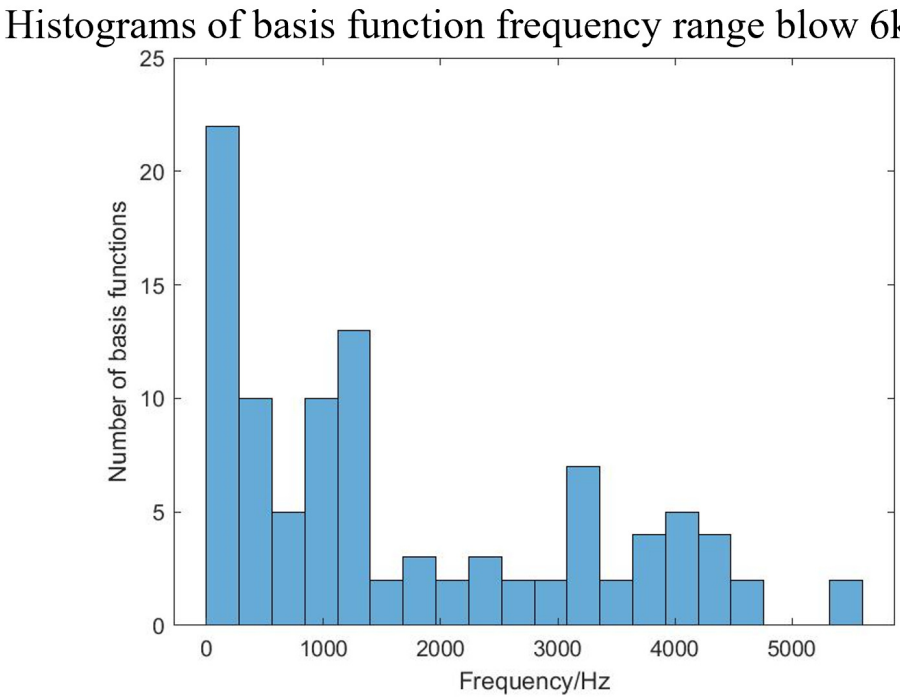
The error bar chart for Root Mean Squared (RMS) error of different center frequency and bandwidth conditions by the GASSOM-based computational model are illustrated in top panel in Figure 5.5, and the mean RMS errors are averaged over repetition trials. Bottom panel in Figure 5.5 is redrawn from the results in Yost and Zhang's experiment (Yost & Zhong 2014) for comparison.

Figure 5.5: The RMS for localization of stimuli with different center frequency and bandwidth by the GASSOM-based computational model are illustrated in top panel, where the bar illustrated the mean the error plots illustrated the standard deviation; Bottom panel are redrawn from Yost and Zhang's work (Yost & Zhong 2014)



To check the effect of the main effect of center frequency and bandwidth as well as

Figure 5.6: This figure illustrates the histogram of center frequency for basis functions learned by GASSOM below 6kHz. It can be seen from the figure that more basis functions have lower center frequency.



their interaction effect, Two-way ANOVA was conducted for the experiment.

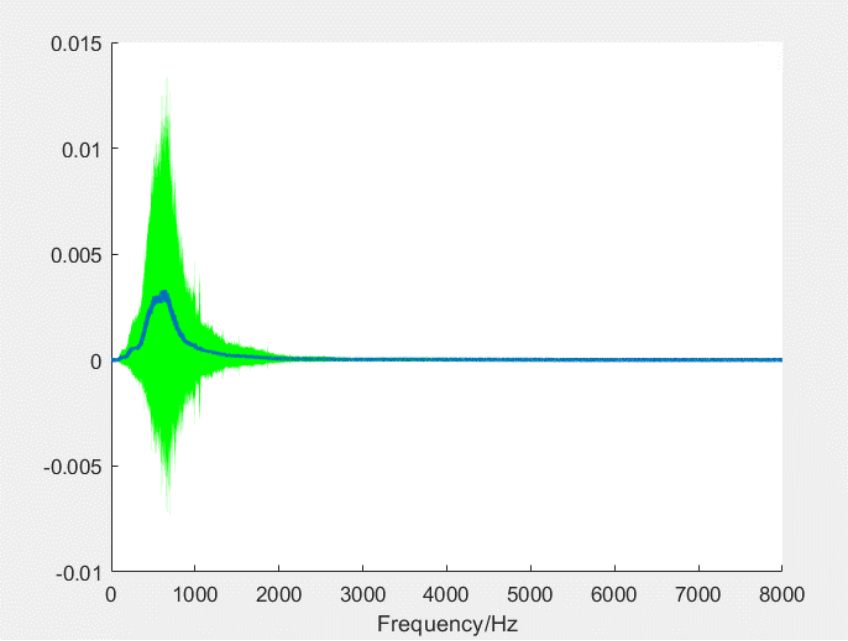
For main effect, the p value for both cases were less than 0.001, which states that both center frequency and bandwidth have statistically significant effect on the localization accuracy. For the interaction effect, the p value is also less than 0.01, which suggested that the interaction of bandwidth and center frequency has a significant effect on the localization accuracy at the level of 0.01.

It is illustrated from the figure that with the increase of bandwidth, there's an increase in sound source localization accuracy for both human and computational model, regardless of the center frequency. This is a natural result as the wider bandwidth provide more information about the location.

It can also be observed from the figure that higher localization errors occurred when the frequency of the stimuli is centered at 2kHz, especially for narrow bandwidth condition. To explore the reason, the frequency range of basis functions below 6kHz was also illustrated in Figure 5.6.

It can be seen from the figure that the number of basis functions with center frequency

Figure 5.7: This figure illustrates the distribution of the spectrum for monaural sound included in this study. The data is extracted from speeches from TIMIT database (Garofolo et al. 1993) by computing Fast Fourier Transform (FFT) and taking the amplitude . The blue curve stands for the mean of spectrum and green region covers the standard deviation. It can be seen from the figure that the spectrum contains more energy between 200Hz and 1500Hz.



at lower frequency (around 250Hz) is larger while the basis functions for 2kHz is less. The number for basis functions for 4kHz is also relatively larger than for 2kHz. The reason for the distribution might be due to the distribution of the spectrum of the training data, as illustrated in Figure 5.7. The figure illustrate the statistics of spectrum of all the speeches in TIMIT database. The blue curve stands for the mean of center frequency bin and green shadow stands for the standard deviation. It is illustrated in the figure that the database contains audio contents around 250Hz to 1,250Hz. Therefore, more basis functions tends to centered around this range.

However, the influence of the center frequency is non-trivial to determine. For the GASSOM localization model, the localization accuracy is heavily degraded when the bandwidth is narrow, and it becomes saturated when the bandwidth reaches 1 octave or above. Similar trend of localization accuracy reduction as increase of bandwidth can be observed from the figures redrawn from Yost's paper. However, the absolute value of localization error differs between the computational model and empirical data.

Difference between our computational model and empirical data on localization accuracy is also illustrated from the figure 5.5 that when the bandwidth is narrow, the human listeners perform better with lower center frequency while the computational model performs better with higher center frequency. The difference might result from the different HRTFs of subjects in Yost's experiment and the HRTFs used to train the GASSOM binaural localization model. However, the HRTFs for Yost's experiment is unavailable. Further study can be conducted for subjects whose HRTFs can be measured.

When the bandwidth becomes broad, the localization error for GASSOM-DNN based localization model reduced significantly. This indicate that the computational model heavily depends on the integration of spatial features from different narrow bands. When information from some of frequency bands is missing, the localization performance degrades significantly.

5.3.2 Effect of sound duration

The effect of sound duration was investigated in two condition: narrow bandwidth and broad bandwidth.

Narrow bandwidth

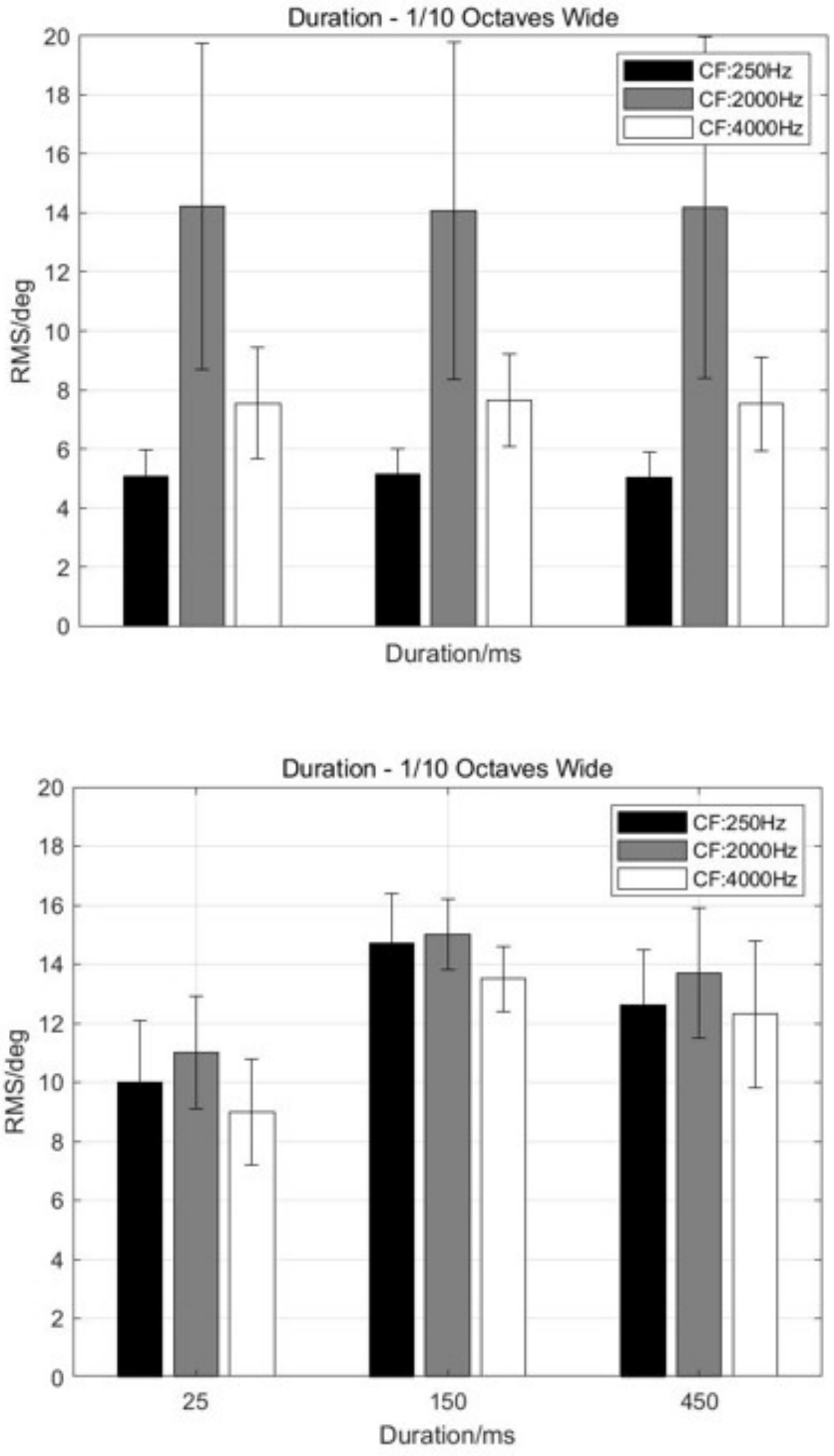
For narrow band, the result is illustrated in top panel in figure 5.8. Two-way ANOVA is conducted on the result to test the effect of sound duration and center frequency. There is no significant main effect of sound duration as $F(2, 22) = 0.5$ and neither significant interaction effect of sound duration and center frequency as $F(6, 66) = 0.2$ at 0.05 level of confidence. However, the main effect of center frequency is still statistically significant as $F(3, 33) = 30.9$, which is consistent with previous result.

Bottom panel in Figure 5.8 is result redrawn from Yost's experiment. Two-way ANOVA in Yost's work also demonstrated that there's no significant main effect of sound duration on the localization accuracy. Similar conclusions can be drawn from the empirical data on human listener, which illustrate that our computational model shows similar performance as human being in the localization task. However, for the GASSOM-based computational model, the localization error varied vastly across different center frequencies, but for empirical data, the variation across different frequencies are less obvious. This result illustrated that the amount of basis functions sensitive to different frequencies

are of similar magnitude, but in our computational model, basis functions for different frequency bins are different. This of one of the major disadvantage of our model that should be solved in the future. Possible solutions might be to increase the data amount or utilize more broadband training data.

What's more, the localization accuracy is best when the center frequency is 250Hz and worst when the center frequency is 2000Hz for both cases, which is consistent with previous experiment. The slight difference on localization accuracy can be explained by the fact that the HRTF used in our model comes from subjects that are different from Yost's experiment.

Figure 5.8: The RMS for localization of narrow band stimuli (1/10 Octave) with different sound duration are illustrated in top panel; Bottom panel are redrawn from Yost's work (Yost 2016). Different colors stand for different center frequencies. It can be observed from the figure that the localization error across different duration.



Broad Bandwidth

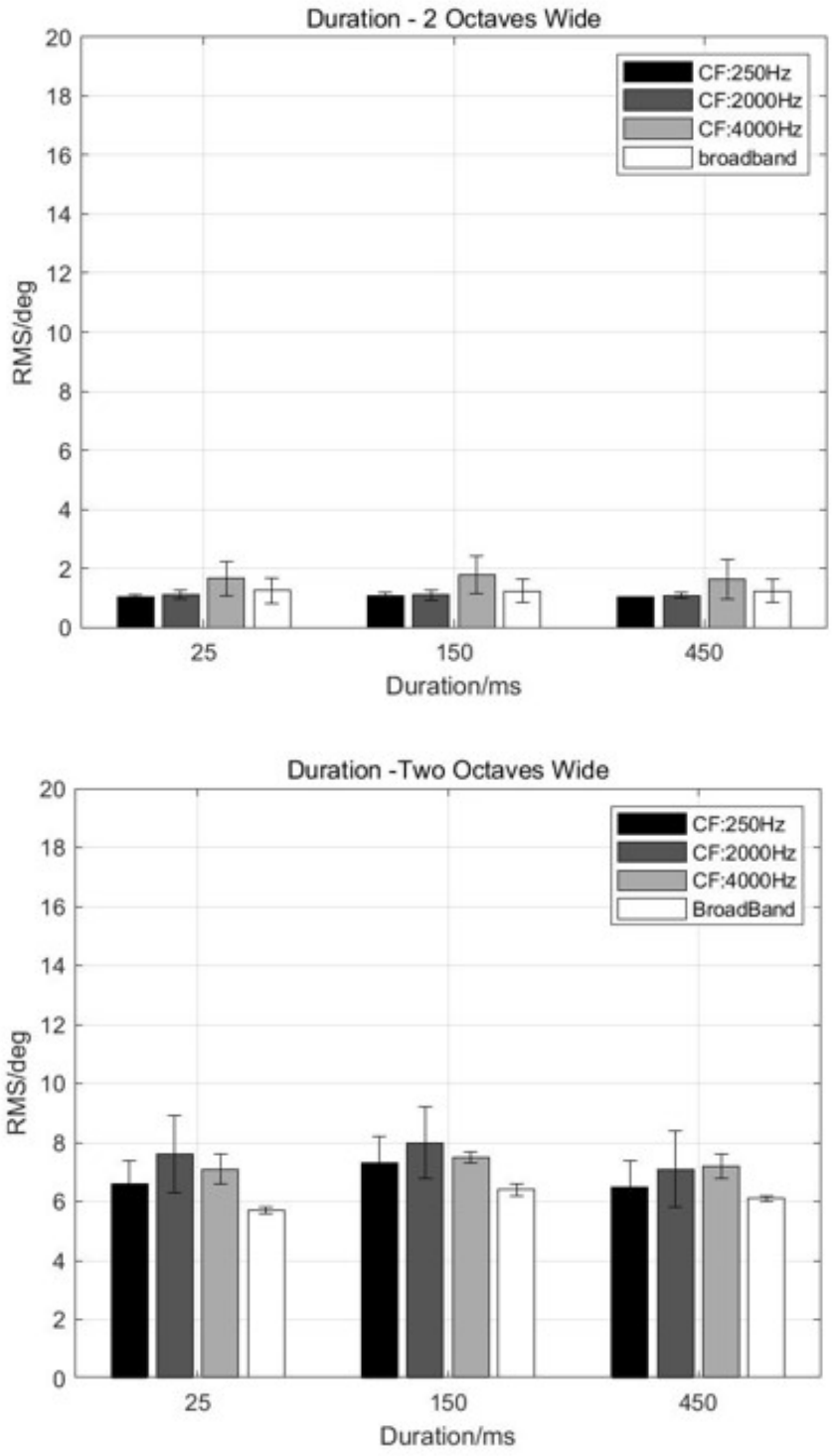
However, when it comes to broadband stimuli, the conclusion varied a little bit from previous one, where center frequency does not play such an important role as before.

The result is shown in top panel of figure 5.9. Two-way ANOVA is also conducted on the result. The main effect of sound duration is not significant as $F(2, 22) = 0.6$ and the interaction effect of sound duration and center frequency is not significant as well with $F(6, 66) = 0.7$. It is surprising that there's also no statistically significant main effect of center frequency on the localization accuracy, which is contradict to the narrow band case. Though the result is still consistent with the conclusion from Yost's empirical data as shown in bottom panel of figure 5.9, which is also redrawn from his paper.

Compared with the narrow band case, the main effect of center frequency is no longer significant. This might due to the fact that the increase of bandwidth made the localization performance less sensitive to center frequency as each testing stimuli contains more information in broader bandwidth. As illustrated in Figure 5.7, even though the number of basis functions centered around 2kHz is less compared with other region, the region within 1kHz and 4kHz contains much more basis functions. Therefore the localization performance is less influenced by the center frequency.

Despite of similar trend between GASSOM-DNN computational model and human listener, the localization is of great difference for broadband signal (2 octave), where the localization accuracy for computational model is much smaller compared with human listener. On the one hand, it might result from the fact that the HRTFs employed between the two experiments were different. On the other hand, it comes from the fact that human are more likely to make random errors during the experiment, while the computational model hardly made such kind of errors.

Figure 5.9: The RMS for localization of broad band stimuli (2 octaves) with different sound duration are illustrated in top panel; Bottom panel are redrawn from Yost's work (Yost 2016). Different colors stand for different center frequencies. It can be observed from the figure that the localization error across different duration. Besides, the localization error across different center frequency is also little, which indicated that both center frequency and duration have little effect on localization accuracy for broadband condition.



5.4 Conclusion and discussion

The adaption effect of the middle ear can be employed to give a better simulation on localization system and explore the effect of sound level on localization performance.

In this chapter we investigated the effect of center frequency, bandwidth, and sound duration on the localization performance of the GASSOM based binaural localization model in comparison with psychophysical experiments conducted by William Yost and his colleagues. For spectral properties, both center frequency and bandwidth have a significant effect on the localization accuracy. When the bandwidth get broader, the localization error decreases, which give computational explanation for the recent study that human auditory system integrate narrow band features for the prediction of sound location. The effect of center frequency depends on bandwidth. For narrow bandwidth, the center frequency has a significant effect. This effect drops with the increase of bandwidth. This might be due to the fact that different frequency channels plays various roles in the decoding of sound location, and lower frequency components seem to be more important in sound localization.

Besides the similarity, difference between human listeners' performance and GASSOM-DNN computational model is also observed. The localization accuracy of computational model is worse than human being when the bandwidth is narrow, and is better than human being when the bandwidth becomes broad enough. The reason is that when the bandwidth is narrow, spatial information from certain frequency bands is missing, and the response of corresponding basis functions is inaccurate. However the DNN network utilized the response as usual without any adjustment. This leads to great misjudgment on localization prediction. In the future work, weights of different frequency channel can be considered to correct the prediction.

As for sound duration, it has no statistically significant effect on the localization effect, which seems contrast to the intuition and longer duration provide richer information. One straightforward explanation for this is that longer duration just increases the present of repeated spatial features which is less informative for sound localization as all the stimuli are of similar bandwidth.

The effect of sound level is not studied in this work due to the limitation of our model. This can be improved by engaging the middle ear adaption, which amplifies the audio with less volume and inhibit audio with greater volume, to the model in the future work.

Difference between human listener and computational model is also observed. When the bandwidth becomes broader, the computational model significantly outperforms human listener. We credit this phenomenon to the fact that computational is less likely to make misjudgment on the localization task.

The experiment on the effect of different spectral and temporal audio properties on localization accuracy of our computational model shows great similarity with empirical data on human subjects. Sparse coding algorithms such as GASSOM is proved to be feasible for the extraction of spatial features like human being. This contributes to the motivation to build a bio-inspired binaural localization model that can be used as alternatives to human subjects in research experiment based on sparse coding and neural network. This will reduce the cost and make the experiment more convenient.

Congruence between the performance of computational model and human being provides insight about the mechanism underlying binaural sound localization. For example, the tonotopic basis functions in GASSOM suggests that auditory system integrate narrow band spatial features to predict location (Pavão et al. 2020). Further study can be conducted for this thesis in the future.

Chapter 6

Effect of non-individualized HRTFs on front-back confusion

6.1 Introduction

Recently, HRTF has been widely used in application of binaural sound for entertainment or scientific research. To provide a more realistic perception, it's admired to employ individualized HRTF for the rendering of spatialized sound. However, for most of the cases it is inconvenient to measure the individualized HRTF for the listener. Non-individualized HRTF is therefore used as an alternative. Due to the physical variation among human's auditory system, including pinna, canal, and torso, the HRTFs of different human listener are always diverse significantly. The performance on reconstruction of the binaural sounds is degraded vastly. Thence, the effect of non-individualized HRTF on human sound localization performance has become an important topic in the exploration of human localization mechanism.

Numerous scientific research has been conducted this area. In 1990s, Wenzel et al. conducted a series of experiment to compare the localization accuracy of human being when listening to virtual audio synthesized with non-individualized HRTFs with individualized HRTFs. The direction of the sounds come from omni-direction on ear-level horizontal plane, which is designed to explore the front-back confusion. Front back confusion is defined as the situation when human listener hears a sound source in the forward direction, he/she perceived it as coming from backward. This phenomenon is very common in our daily life. The localization accuracy degraded significantly when front-back confusion emerges.

Recently, in order to eliminate the influence of non-individualized HRTFs on the front-back confusion, Prof. So et al. proposed one affordable way to solve this issue by providing more selections of stimuli synthesized with HRTFs that come from spectral feature clustering analysis to the human subjects, who are asked to select one stimulus that give the best judgment of front-back perception. The result show great improvement with the provision of extra HRTF choices, which demonstrated that HRTF clustering is a potential way to reduce the occurrence of front-back confusion.

However, most of the computational binaural sound localization model are focused on the localization accuracy of stimuli synthesized with individualized HRTF, while little concentration has put on non-individualized HRTF. In fact, the localization performance of sound localization model is heavily dependent on the usage of different HRTFs. It has been proved (Wang et al. 2020) the different computational model trained with different HRTFs from CIPIC database testing with different stimuli showed diverse localization accuracy, which is due to the similarity between the HRTFs used for training and testing. A clustering analysis about HRTF can be applied to analyze the similarity between the HRTFs to provide a better localization performance when the individualized HRTFs is unavailable.

In the first part of this chapter, we explored the effect of individualized and non-individualized HRTF on the localization accuracy of GASSOM based computational model. The performance is compared with empirical data of human listener in Wenzel et al.'s work. In the second part, we conducted a similar procedure as Prof. SO's work that by providing optional HRTFs, the front-back differentiation of the computational model would be improved. We also conducted a follow-up experiment on the influence of cluster distance of front-back confusion rate of different computational model.

6.2 Effect of non-individualized HRTF on front-back confusion

6.2.1 Review

In Wenzel et al.'s experiment, human subjects were asked to judge the apparent location of broadband stimuli played through physical speak or headphone. The physical

speaker is used to simulate the individualized HRTFs as it has been shown that free-field stimuli would achieve similar performance as stimuli synthesized with individualized HRTF (Wightman & Kistler 1989). For non-individualized HRTF condition, stimuli are synthesized by filtering Gaussian white noise with non-individualized HRTFs and played through headphone. The directions of HRTFs ranged from $-180°$ to $180°$ on ear level horizontal plane. It should be noted that the backward directions are also included in this case for both training and testing. The localization performance was quantified using root mean square errors as defined in previous chapter.

Top panel of Figure 6.1 is an example of the predicted azimuth against target azimuth of subject SIK in free field (individualized HRTF), and the bottom panel of Figure 6.1 is for non-individualized case. The performance is relatively satisfying when the subject listening to free field stimuli, where only small portion of front-back confusion appear. But when it comes to non-individualized HRTF, front-back confusion rate increases significantly, especially when the stimuli are located near the sagittal plane.

6.2.2 Methods

Following the same procedure, we generate the testing stimuli by convolving 200ms Gaussian white noise with HRTFs of azimuths ranging from $-180°$ to $180°$ degree on horizontal plane. In this study, the HRTFs is selected from CIPIC database as it contains HRTFs measured from different subjects, which makes it possible to evaluate the influence of non-individualized HRTF from other subjects.

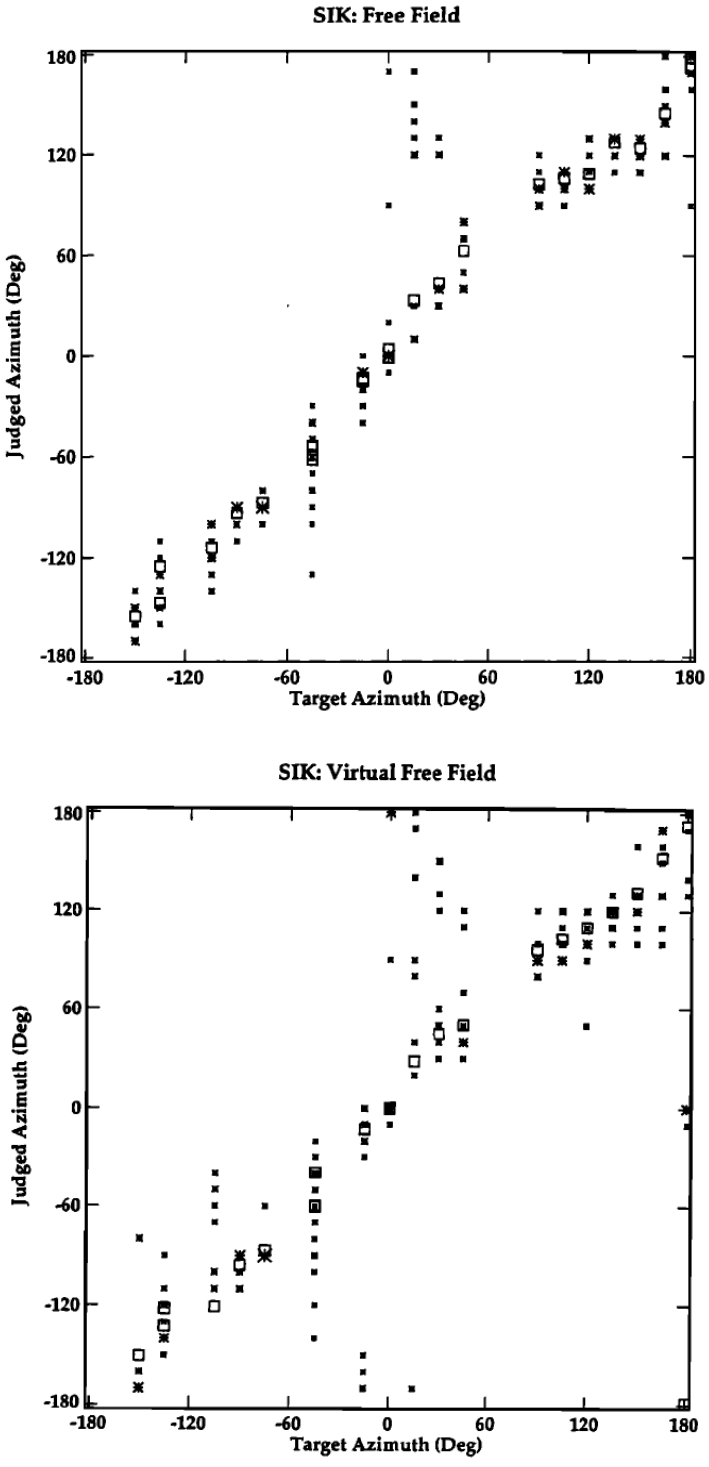
The GASSOM based localization model is first trained with HRTFs from a randomly picked HRTFs. After training, it is tested with stimuli generated with the same HRTF used in training phase for individualized HRTF comparison, and stimuli generated with HRTFs from other randomly picked subjects for non-individualized HRTF comparison.

Unlike Wenzel et al.'s experiment in which only 3 subjects are employed, we trained 10 computational model with 10 different HRTFs as virtual 'subject'. This is one advantage of computational model on convenience.

6.2.3 Result and analysis

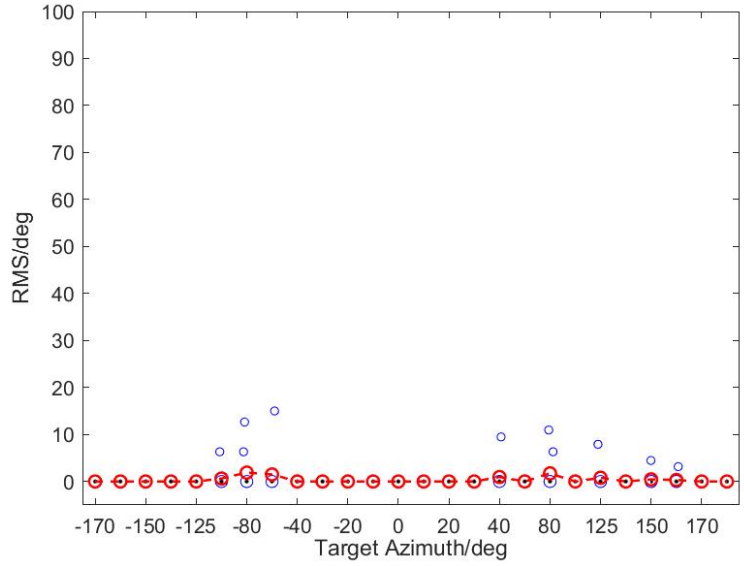
Top panel of Figure 6.2 illustrates the box plot of the RMS error of localization with individualized HRTF for the 10 GASSOM based computational models. Bottom panel

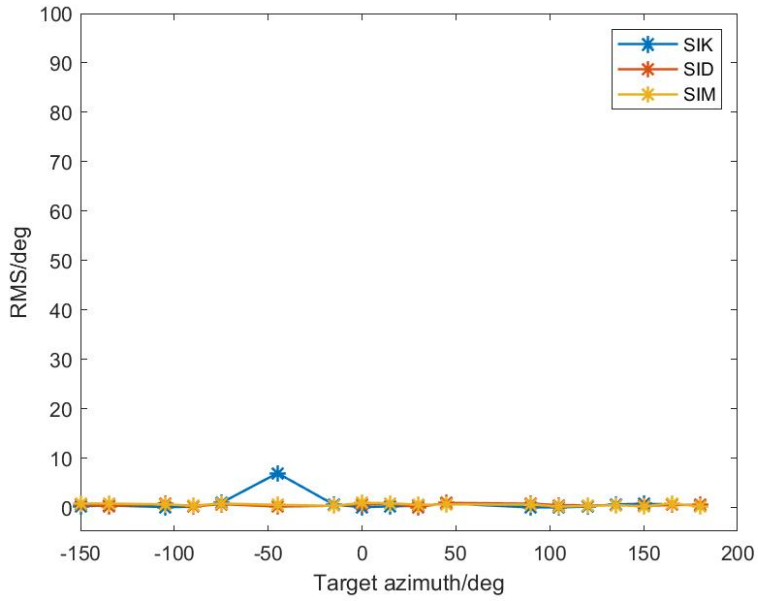
Figure 6.1: The RMS error for location prediction of stimuli in free field for subject SIK in Wenzel et al.'s experiment are illustrated in top panel (Wenzel et al. 1993), which is to simulate localization of synthesized stimuli with individualized HRTF; Bottom panel are results for virtual free-field, which is stimuli synthesized with non-individualized HRTF (Wenzel et al. 1993). It can be seen from the comparison that localization error increases when listening to non-individualized HRTF stimuli. In Wenzel et. al.'s experiment, they employed 3 subjects named as SIK, SID and SIM, and only SIK is demonstrated here.



in Figure 6.2 is redrawn from the empirical data on three human subjects SIK, SID and SIM. The target azimuth difference is due to azimuth measured by CIPIC HRTF database differs from Wenzel et al.'s experiment. We have applied HRTF interpolation (VBAP, bilinear interpolation) to synthesize the HRTF at the same directions as Wenzel et al.'s experiment, but the performance drops a lot. Therefor It is obvious from the data that when testing with stimuli synthesized from individualized HRTFs, the localization accuracy is close to zero for most of the azimuths for all the three computational models trained with different HRTFs. This is in consistent with human subjects' data in free field as shown in figure 6.2, which also give support to the conclusion that listening in free field retains comparable performance as listening to stimuli synthesized with individualized HRTFs.

Figure 6.2: In this study, we compared our results with those of Wenzel et. al.'s experiment . The top panel illustrate the RMS error for localization of stimuli synthesized with individualized HRTF from CIPIC HRTF dataset (Algazi et al. 2001) for GASSOM-based localization model; Bottom panel are corresponding results redrawn from Wenzel et al.'s experiment (Wenzel et al. 1993), where SIK, SID and SIM are different subjects. It can be seen that the RMS was small for both cases, which indicated that the localization performance was good for both computational model and human being.





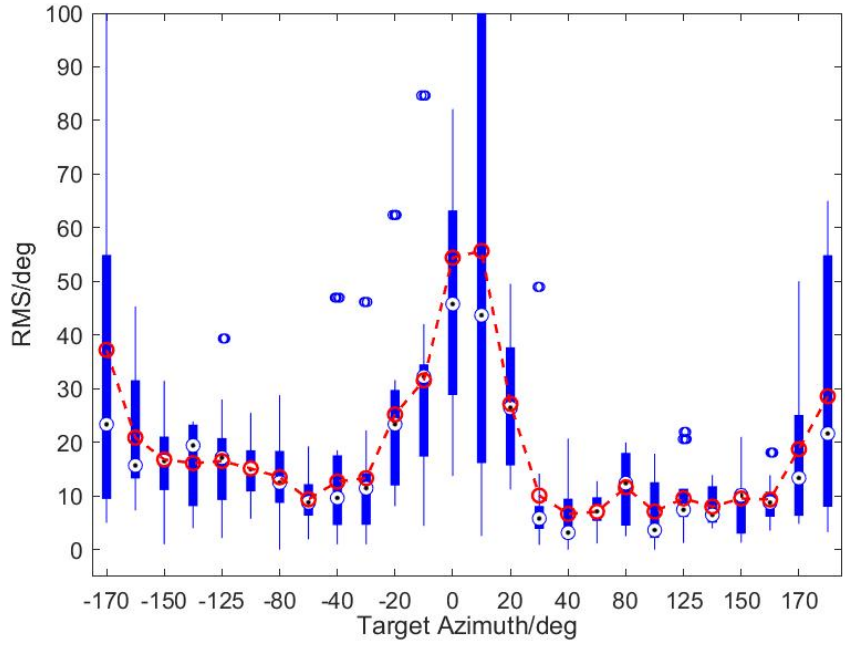
When it comes to non-individualized HRTFs, the performance drops significantly.

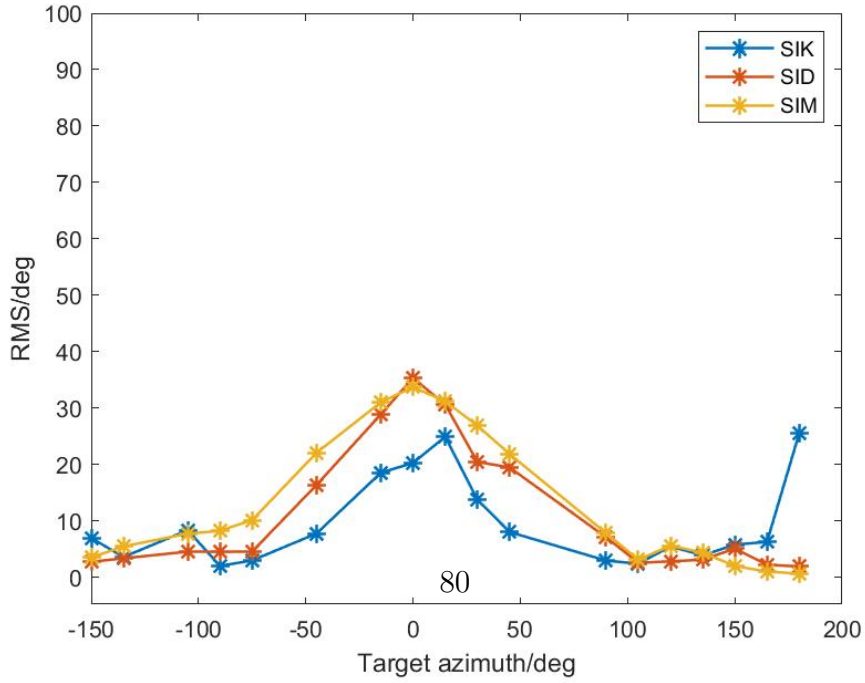
Bottom panel of Figure 6.3 is redrawn from Wenzel et al.'s experiment on human being. The localization error is also highest in the central front and central back among all the 3 subjects in the experiment. Comparison between computational model and empirical data on human illustrate that our model is competitive in the simulation of human binaural sound localization. Therefore the basis functions may be capable of catching the spectral cues that utilized by human being for front-back determination. This is in consistent with the advantage of sparse coding algorithm where the spectral filters will appear automatically without any manipulation.

Results are shown in top panel of figure 6.3, the RMS errors are relatively large when the target sound source is located close to the median plane, including both central front and central back. Our explanation for it is that chances of front-back error is more likely to happen in sagittal plane as the signals arrive at left and right ears are more symmetric, and the information provided by the binaural signal becomes more redundant. Hence less spectral cues can be utilized to discriminate the forward and backward stimuli. Besides, one mistake in the sagittal plane would lead to localization error of 180°, while it decreases as the sound location comes to lateral side in the occurrence of front-back confusion.

Another possible reason for the result might be due to when the sound source is located in central directions, the spectral cues difference between the frontal and backward directions tend to be more similar because it is mainly affected by the pinna, therefore the front-back confusion increases. But when it comes to the lateral side, the spectral cues difference between the frontal and backward direction tend to be more distinct as a result of the composition of pinna reflection and the shadow effect of head and torso. Therefore the front-back discrimination is improved for lateral side. These guess need to be carefully validated in future work.

Figure 6.3: In this study, we also compared our results for non-individualized HRTF case with those of Wenzel et. al.'s experiment (Wenzel et al. 1993). The top panel illustrate the RMS error for localization of stimuli synthesized with non-individualized HRTF from CIPIC HRTF dataset (Algazi et al. 2001) for GASSOM-based localization model; Bottom panel are results redrawn from Wenzel et al.'s experiment. It can be seen that the RMS was small for both cases, which indicated that the localization performance was good for both computational model and human being. For both cases, the localization error became much larger comparing with individualized HRTF, and the error is highest when the stimuli come from center front/back directions.





Biological linkage

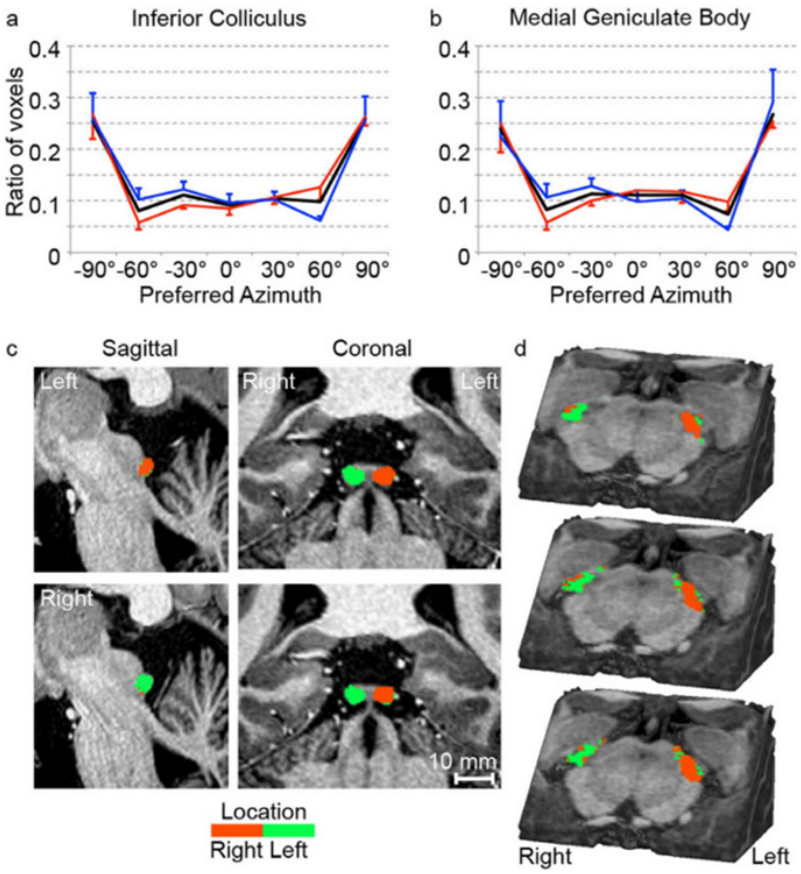
The relative higher localization accuracy for lateral side can also be explained by biological studies (Moerel et al. 2015). As illustrated in Figure 6.4, the ratio of voxel are both higher for binaural stimuli at Left ($-90°$) or Right ($90°$) sides. Which reflect that there are more neurons in subcortical area of auditory system that are sensitive to sounds from lateral side than central side. In our computational model, we can also observe more basis functions that correspond to stimuli that come from Left and right side. This also guarantees the congruence between the GASSOM and Biological neurons.

6.3 Effect of HRTF clustering analysis on front-back confusion

6.3.1 Review

Investigation on the spectral content has demonstrated that different critical bandwidths play an important role in the forward and backward discrimination. Fluctuations within certain spectral range would lead to notable change on perception of the direction where the sound source comes from. For example, the stimuli is more likely to be perceived as coming from the front if the energy within 0.28-0.56kHz is amplified. In SO's work, they review several critical bands that heavily influence the perception of front-back stimuli from previous study. These bands are summarized in table 6.1.

Figure 6.4: The tuning to stimuli of different locations for neurons in IC and MGN is illustrated in this figure (Moerel et al. 2015). The top two panel illustrate the voxel ratios in subcortical region for stimuli of different directions. The top left is for Inferior Colliculus and the top right is for Medial GEniculate Body (MGN). The red curve stands for low frequency and blue curve stands for high frequency. Therefore the localization accuracy is higher for 90° and −90°. The bottom two panels illustrate groups of neurons tuning to left and right stimuli in IC and MGN, respectively. The green regions correspond to right side and blue region corresponds to left side. Neurons tuning to both lateral directions take a large portion in the organism.



Band	Center Frequency
F1	3800-8000Hz
F2	13200-16000Hz
F3	150-540Hz
F4	1900-2900Hz
F5	3600-5800Hz
F6	8000-16000Hz
B1	10000-13000Hz
B2	720-1700Hz
B3	7400-11100Hz

Table 6.1: Critical band and frequency range for perception of sound direction (So et al. n.d.). F for forward and B for backward. 6 bands were selected for forward direction and 3 for backward direction.

Afterwards, different spectral quantities are extracted from these critical bands of HRTFs for center front and center back directions among different subjects as features for clustering analysis. Hierarchical agglomerative algorithm is employed to divide the HRTFs into 9 clusters, 3 of which are omitted as it's singleton.

Localization of stimuli synthesized with MIT KEMAR HRTFs served as baseline. Stimuli synthesized with HRTFs from the center of 6 clusters are provided to subjects as optional choices, from which the subject select the one with best front-back localization accuracy. The result demonstrated that with the provision of extra HRTFs, the localization errors are reduced significantly.

6.3.2 Exp.1

Inspired by this work, we first re-conduct similar experiment to examine the effect of extra HRTF options on the improvement of localization. Another subsequent experiment was designed to investigate the relevance of clustering distance on front-back confusions. Front-back rate was calculated as the percentage of center front stimuli that perceived as coming from backwards. Back-front confusion is defined as the percentage of center back stimuli that perceived as coming from forwards.

Methods

In the first experiment, GASSOM-based binaural localization model was trained following the identical procedure as before, where the HRTF is selected randomly from CIPIC database. HRTFs from the rest of the database are then employed for clustering analysis. There are 90 HRTFs in total, where the left and right ear for each of 45 HRTFs were considered separately. This is because we focused on front-back discrimination, where the spectral cues play an important role and the left and right HRTFs at central directions (front and back) tend to be more symmetric.

Spectral features from different critical bands are extracted and agglomerative algorithm with average linkage was used for clustering. We also selected 6 critical bands (F1 to F6) and 3 critical bands (B1 to B3) for frontal and backward directions, respectively. Afterwards, a 90×90 matrix was extracted from the CIPIC database and used for clustering analysis.

In the end, 5 clusters are selected from the result after clusters with few HRTFs and singletons are neglected for both frontal (azimuth 0°, elevation 0°) and backward (azimuth 0°, elevation 0°) directions as illustrated in Figure 6.5. Different colors stand for different clusters and the cluster index was marked in red in the figure.

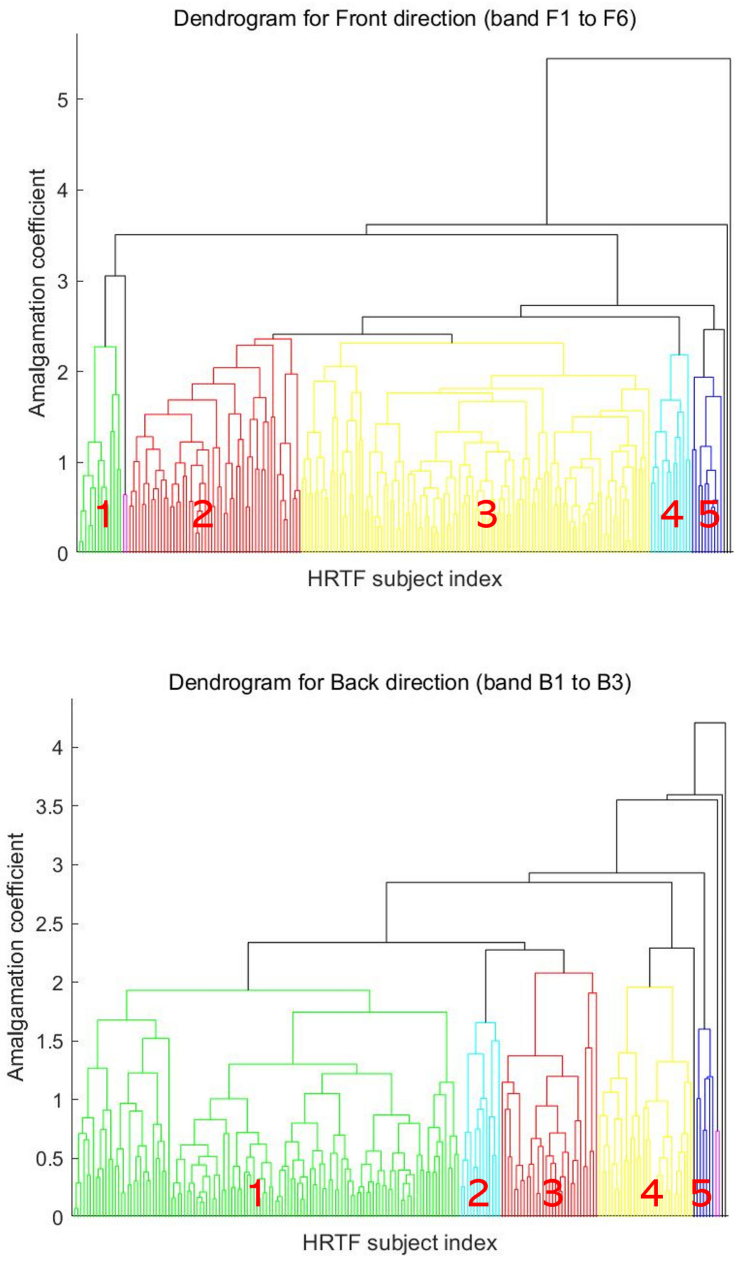
Stimuli synthesized with KEMAR HRTF in center front and center back are used as baseline stimuli and tested with the computational model. Afterwards, 5 HRTFs from the center of different clusters were used to generate stimuli and tested with the computational model, among which the HRTF with least front-back confusion rate was used to benchmark with the localization performance of baseline model trained with KEMAR HRTF.

Results and analysis

The results are illustrated with bar chart in figure 6.6 for both front-back confusion rate and back-front confusion rate. The confusion rate is significantly reduced with for the case with HRTF selection.

One-way ANOVA is conducted and $p_{front-back} = 1.19e-5$, $p_{back-front} = 0.0066$. It states that in both cases, the provision of optional HRTFs significantly reduced the front-back confusion. This is consistent with the conclusion in So's experiment on empirical data of human subjects. This result also ensured the congruence that our GASSOM based

Figure 6.5: The top panel illustrates the dendrogram of HRTF clustering for frontal direction; the bottom panel illustrates the dendrogram of HRTF clustering for backward direction. Different colors stand for different clusters. Cluster indices are marked in red. 5 clusters were selected for each direction after omitting the singletons and clusters with few HRTFs.



computational model can produce similar performance as human being.

6.3.3 Exp.2

Methods

The following experiment investigated the relation between HRTF clusters on front-back confusion. Neighboring clusters are believed to share more similarities on the spectral cues as a result of the intrinsic property of the clustering algorithm.

According to the index of the cluster, binaural sound localization model was trained with HRTFs from cluster 1 and tested with stimuli generated with HRTFs from other subjects in CIPIC database on center front and center back directions. The HRTFs might come from the same cluster as the training HRTF (cluster 1) or different clusters. The resulted front-back confusion rates that come from the same cluster are grouped together.

Result and analysis

The results of experiment 2 are illustrated in figure 6.7. In this experiment, cluster 1,2,3,4,5 are selected for comparison. One-way ANOVA is conducted among different clusters on the front-back confusions. $p = 0.0087 < 0.01$, which states that there's statistically significant effect of HRTF cluster on the front-back confusion rate.

It can also be observed from the chart that the front-back confusion rate is lower when the testing stimuli is synthesized with HRTFs from the same cluster as training HRTF (cluster 1), and increases as the cluster distance, measured by the absolute index value difference, increases. This result gives support to the validation of the selection of critical bands utilized by HRTF clustering analysis.

However, cluster 4 is abnormal from the conclusion. We conducted a subsequent test by training a GASSOM localization model with HRTFs from cluster 6 and test with stimuli synthesized with HRTFs from cluster 1. To our surprise, the front-back confusion is still close to 0. The double-sided test demonstrated that HRTFs from cluster 1 and cluster 4 resulted similar perception on front-back discrimination, but their spectral features in critical bands are different from each other.

One reason for this phenomenon is that as different critical bands have counter effect on the functionality, after these bands superimposed with each other, the influences

Figure 6.6: The top panel illustrate the front-back confusion rate for comparison between HRTF selection from 5 clustering analysis candidates and baseline KEMAR HRTF; Bottom panel are corresponding back-front confusion rate. It is illustrated from the figure that clustering analysis HRTF has less confusion rate than baseline KEMAR HRTF in both front-back confusion discrimination and back-front confusion discrimination.

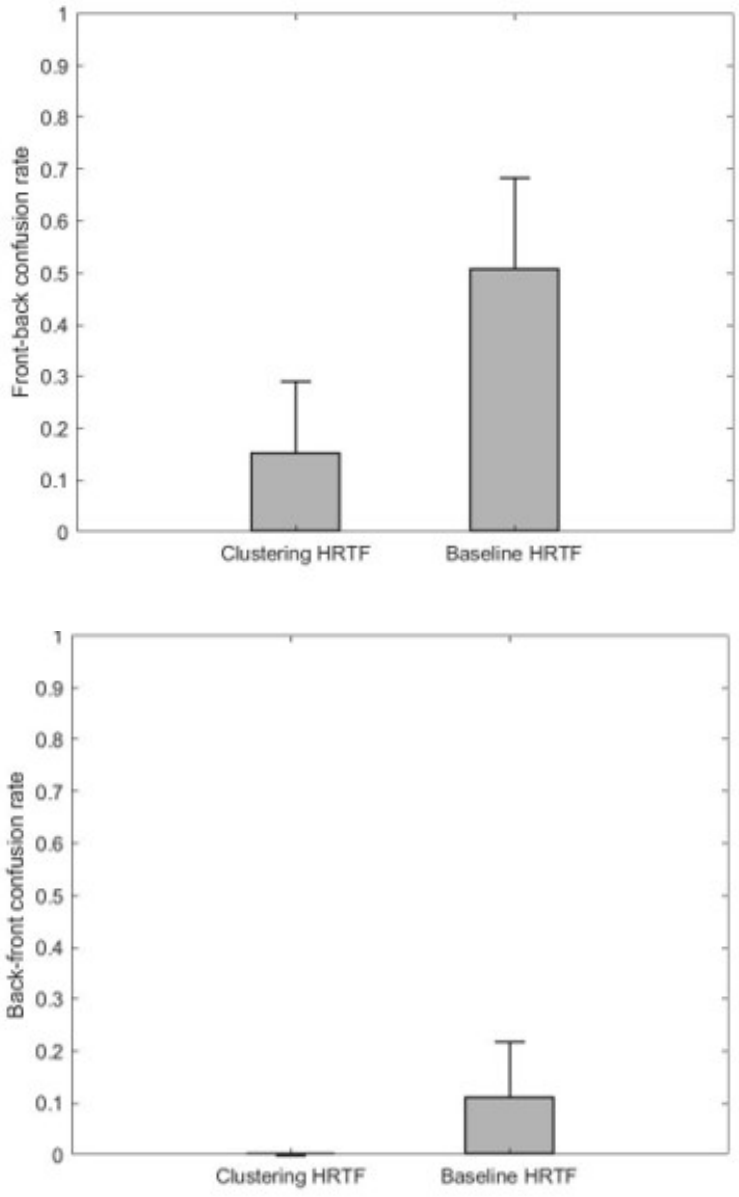
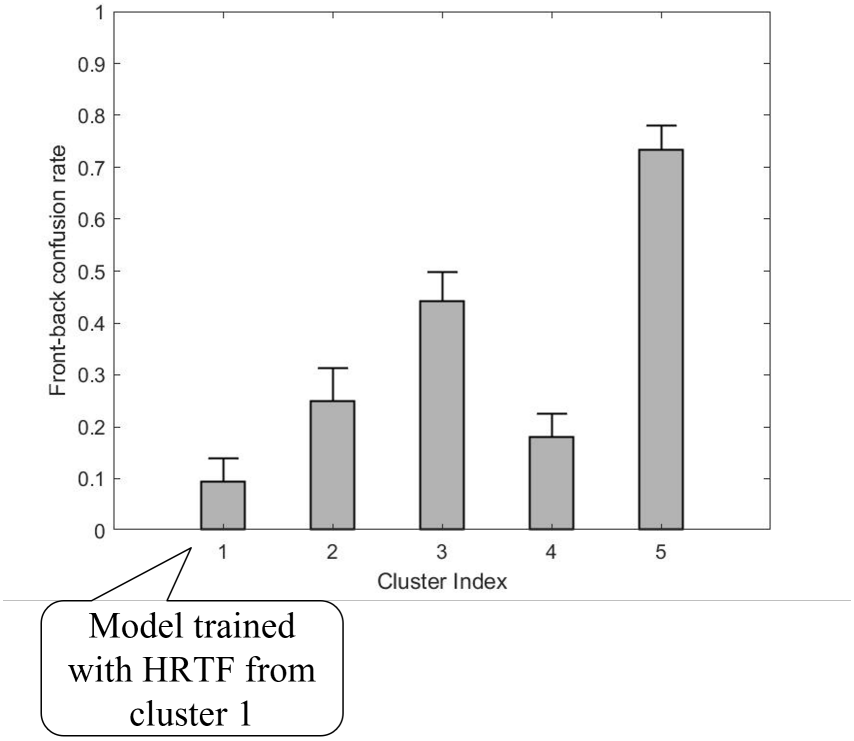


Figure 6.7: The front-back confusion of testing stimuli synthesized with HRTFs from different cluster center are illustrated. The horizontal axis is cluster index, and vertical axis is the front-back confusion rate. As distance among the clusters increase, the front-back confusion also increases.



were canceled. Another explanation is that the selected critical bands are not representative enough. There might be potential critical bands that have some influence on the perception but undiscovered yet, and this requires future work on the exploration.

6.4 Conclusion and discussion

In this chapter, we explored the influence of non-individualized HRTFs on the front-back confusion by repeating experiments that haven been conducted on human subjects to our computational model. Similar conclusions were drawn from the results which demonstrated that our GASSOM based binaural localization model achieves great similarity with human being.

During the first experiment, we tested the localization performance of computational model with individualized HRTF situation by using the same HRTF as training phase,

while in the psychoacoustic experiment, free-field condition was employed as alternative to individualized HRTF situation because the individualized HRTF is inconvenient to access. This is one of the potential advantages of developing a human-like binaural localization model.

In the second experiment, HRTF clustering analysis was conducted with respect to spectral features extracted from critical bands that have been proposed in previous studies. Our binaural localization model also shows great similarity with empirical data on human listeners, which gives support to the conclusion that provision of optional HRTFs to the listener can significantly reduce the occurrence of front-back confusion.

The subsequent experiment validated the feasibility of clustering analysis and critical bands, which suggests that the front-back confusion can be reduced by providing HRTFs that share similar spectral features as the listener when the individualized HRTF is unavailable. This provide another way to solve the front-back confusion issue.

Chapter 7

Summary and discussion

Modeling binaural sound localization was investigated in this work. GASSOM and DNN based sound localization model was developed which consists of two phases: spatial feature extracted with sparse coding and sound location decoder with DNN. Sparse coding performance of GASSOM and ICA on binaural sound spectral-temporal representation was first compared and GASSOM was selected as the preferred spatial feature extractor. Experiments were conducted to determine the optimal parameter for GASSOM training. These optimal parameters were applied to develop and train a GASSOM-DNN binaural localization model (BLM). Two further experiments were conducted to compare the model predictions with empirical data collected in previous listeratures using human listeners. Results indicated that similar to human listeners, our GASSOM-DNN BLM's localization performance also suffered from front-back confusions and were significantly influenced by waveform parameters of the binaural cues.

7.1 Sparse coding of binaural sounds

In the first study, we give an overview of the computational model. It comprises two parts: the first part aims to extract spatial features. Unlike the conventional methods which explicitly calculate the Interaural Time Difference (ITD) or Interaural Level Difference (ILD), sparse coding was employed as the extractor of sound directional information. The form of the basis functions learned by sparse coding algorithm has no pre-fixed form, and all the process are unsupervised.

The result basis functions are similar to receptive fields of auditory neurons. It can also extract spectral features that utilized by human listeners to discriminate frontal and backward directions (see Chapter 6). Different basis functions have special frequency

range, which is consistent with the tonotopic structure of auditory neurons in auditory system (e.g. Inferior Colliculus).

Two of the most representative sparse coding algorithms, Generative Adaptive Subspace Self-Organizing Map (GASSOM) and Independent Component Analysis (ICA), were compared on the encoding of cochleagram of spatialized binaural sounds. Fisher information and disorder index were used to quantify the informative about sound location and topographical smoothness of the two algorithms, and GASSOM outperforms in both measurements. Therefore GASSOM was selected as sparse coding algorithm in the following study. It is natural that GASSOM achieve more smoothness across neighboring basis functions as a result of its intrinsic self-organizing structure. The explanation for its out-performance in spatial information extractor is non-trivial. But the diversity of basis functions learned by ICA suggests the sparse coding process of ICA is not task-specific and it seeks to maximize the independence among the basis functions it extracted, while for GASSOM it aims to encode the invariant features within each episode, and in this case the spatial information is fixed in each episode, therefore the directional information is well-preserved.

The author admit that ICA is more appropriate for extractor of more diverse acoustic features, while GASSOM is more task specific. Besides, topographical smoothness is intrinsic property of GASSOM, though the advance of topographical structure is not reflected here. It may not be fair to compare with Independent Component Analysis. Other metrics should be explored for complete evaluation.

GASSOM was proved successful to encode audio stimuli as it has been successfully applied to encode visual stimuli. This result also gives support to the uniform principle of sensory system in response to stimuli of different modality.

7.2 Determining the optimal parameters for GASSOM training

As GASSOM is served as sparse coding algorithm, one of the major problem is to determine the parameters for training to attain better performance. In the second study, map size as well as chunk length and chunk shift are tested. A larger map size would give more space for GASSOM to encode the stimuli. But it cannot be too large because some

basis functions will be unused in the end. The result verified our guess. The localization performance increases when the map size increase from 8 by 8 to 10 by 10. But when it's 16 by 16, the performance drops. According to the resulted basis functions, some of them are rarely selected as winner during the training process, therefore is redundant. But in the following phase of DNN training, they are still uitlized as input features, therefore caused degradation on the performance.

Chunk length and chunk shift are another two factors considered in this study. Chunk length determines the length of the basis functions in GASSOM and chunk shift determines the temporal slowness, which is one of the underlying principles of GASSOM. The localization error was least when the chunk length was set to 10 frames and the shift between successive chunks was set to 1 frame. The result verifies the temporal slowness that when the shift is small enough, the spatial information is more likely to be captures. But for the chunk shift, it's also non-trivial to explain the reason why 10 frames would be a better choice. Our guess is that if the length is too small, the temporal feature cannot be integrated by the basis functions to decode the location. But if it's too large, the temporal resolution is too coarse to capture the finer variation of the spatial features. The variance of spatial cues such as ITD was also proved to be utilized by the auditory system to decode the location. Our experiment also supported this conclusion.

7.3 Effect of waveform properties on localization accuracy

The 3^{rd} and 4^{th} study seeks to bridge the gap between psycho-acoustic experiments and computational modeling. It describes the development of GASSOM-DNN model that integrates spectro-temporal cues for binaural sound localization and bench-marked with empirical data on human's psychophysical experiments, including experiment on different spectro-temporal experiments and the effect of non-individualized HRTFs.

As little attention has been put on the effect of waveform studies, we conducted this experiment to fill this gap and investigate to what extent similarity our computational model can achieve comparing with human listener. Center frequency together with bandwidths and sound duration were investigated in this study to benchmark with empirical data on human being. The sound level is not investigated because the sound level was

normalized during the preprocessing of binaural stimuli.

For spectral properties, the result indicated that both center frequency and bandwidth are important for the decoding of sound location. This is in consistent with the empirical data on human being conducted by Yost and his colleagues. Broader bandwidth will increase the localization accuracy as more information is captured for broader bandwidth, which suggested that the model integrate information in different frequency bins to decode the location. One explanation comes from recent study which revealed that human auditory system compute the spatial information within each narrow band (as defined in cochleagram) and integrate them to predict the sound location. Our model provided another way to verify the judgment.

However, the relation between the localization accuracy and center frequency is more difficult to determine, though the One-way ANOVA suggested that center frequency did play an important role in location estimation. For our computational model, it's due to the fact that the number of basis functions response to different frequencies vary and is highly dependent on the training data spectrum. We admit that if we increase the categories of audio stimuli for the training, this phenomenon might be improved. We selected speeches as representation for natural sounds (Lewicki 2002) is based on the fact that it contains vowels and consonant, which correspond to harmonic and white noise, respectively. For human being, the relation between localization accuracy and center frequency is still a black box. Our guess is that the number of neurons with receptive fields around different frequency range vary accordingly. This hypothesis is based on analog from basis functions in GASSOM. It also demonstrates that GASSOM provides us another way to make assumption about natural phenomenon.

For temporal properties, that is the sound duration, no statistically significant effect was found in our experiment. This is contrast to the intuition that with longer duration, more information is provided to the localization system to dig up. Our explanation is that as the enlarging of duration is just repetition of the single sound piece, no extra information can be extracted. Also this result is consistent with the empirical data on human listeners about sound duration. Both experiments showed similarity between our computational model and human being.

7.3.1 Phase locking

Phase-locking, as part of the famous volley theory, has been proposed to play an important role in encoding temporal fine structure (TFS, that is the higher frequency part) of stimuli as the maximum vibration of cochlea fibers for human being is relative lower as a result of the stiffness of auditory hair cells. It mainly occurs at lower frequency, and less frequently when the frequency becomes higher as the phase-locking tends to be random and the phase is hard to align accurately.

In this study, however, we only employed as relatively simple model that treats all the frequency equally without paying attention to the phase locking. This is because many mechanism has not been explored, such as the upper limit frequency of phase locking is not yet well determined by the researcher, which might range from 1,500Hz to 10,000Hz (Verschooten et al. 2019). It is suggested that the upper limit of phase-locking varies for different functions, and for binaural processing the possible phase-locking upper limit is around 1,500Hz, which might be utilized in building the phase-locking module for our binaural sound localization model in the future. Besides, the lack of phase-locking might be an explanation for the lack of ITD sensitivity of basis functions learned by GASSOM as illustrated in chapter 3 due to the fact that phase information is omitted during calculation of cochleagram.

As phase-locking model occurs before GASSOM learning, it's compatible to employ a more human-like auditory periphery model for preprocessing stage of the framework. Many computational models have been proposed, among which recently, Perterson et. al. proposed a model cascaded a Boltzman function with a low-pass filter and a static exponential transfer function to simulate the phase-locking of auditory neuron fibers (Peterson & Heil 2020). It owns the advantage that parameters are independent of sound level and the model can account for the instantaneous change of spiking rate with respect to variation of stimulus level. This model can be employed in our future work to improve the sensitivity to temporal fine structures, which is essential in ITD encoding.

7.4 Effect of non-individualized HRTF on front-back confusion

Front back confusion has bothering the application of spatial audio for a long time. Many studies have been conducted about the cause to this issue. It has been found that the usage non-individualized HRTF would increase the localization error, especially front-back confusion rate. However, it is unrealistic to measure the individualized HRTF for each subjects. Great effort has been placed to find a better to reduce the influence non-individualized HRTF on front-back confusion. One solution is to provide the listener with HRTF selections that come from different HRTF clusters. The clustering analysis were conducted based on the spectral properties of critical bands that play important role in the front-back directional sound perception.

In this study, we first validate the influence of non-individualized HRTFs on GASSOM-DNN based computational model. Similar experiment was conducted following procedures in Wenzel's experiment (Wenzel et al. 1993). The results demonstrated that our computational model also suffers from front-back confusions. For this reason, a following experiment inspired by the HRTF clustering analysis. The front-back confusion rate was validated by testing the model with HRTFs from different clusters. The result indicated that the front-back confusion rate would be reduced if the cluster distance of the training and testing HRTFs decreased. This provided us with another idea to reduce the detriment of front-back confusion. We can quantify the similarity using the spectral cues in critical bands between the target HRTF and optional HRTFs when the target HRTF is not available. The selection of HRTFs with more similar spectral properties would promote the performance in front-back discrimination. This experiment also verified the feasibility of the spectral cues employed in the clustering analysis.

7.5 Application

The GASSOM-based binaural localization has the advantage of unsupervised learning based on sparse coding with topographical structure. The basis functions enable a more straightforward visualization on the spatial features learned by sparse coding algorithm. Based on the analysis of basis function, more specific hypothesis can be proposed in

future study. Much is still unknown about binaural localization by human listeners. In particular, localizing sound in the presence of echo is still a topic of research. The lack of knowledge on binaural localization in the presence of echo cause performance degradation in most audio separation models. With the GASSOM-based binaural localization models, empirical experiments can be simulated to study and understand how human listeners can localize sound under heavy reverberation.

In our work, we only trained and tested the model in anechoic environment with only one source. It can also be applied to localization multiple sound sources simultaneously in noisy environment with reverberation, which is more realistic in our daily life. It is expected to learned spatial features that related to the perception or separation of several simultaneous speakers or suppression of subsequent echos comes after the first arriving of waveform.

7.6 Limitations and future work

Limitations

The model has only been trained with directional speeches as representation for natural sounds. More varieties of audio stimuli (such as instruments, background noise, and other languages) should be used for training to make the model more general and robust. Benchmarking with human being's frequency sensitivity can also be check with the increasing usage of different types of audio stimuli.

In this work, many outcomes are not well-explained and the discovery for the intrinsic reason is non-trivial, such as the selection of GASSOM training parameters. We admit that the optimal parameters for different configuration of GASSOM might lead to different results. A unified framework should be developed to evaluate the performance of sparse coding results under different conditions.

In the third study, many factors that might influence the performance of localization accuracy that were tested in Yost's experiments on human being have not been explored yet due to the limitation of the current work, such as sound level. We believe with the gradual improvement of the current model, those factors can be explored in the future work.

In terms of the second phase in our model, DNN might not be the best choice as it is considered as a 'black box' and DNN is just one of choices that give a relative better performance. Other framework such as spiking neural network (Wu et al. 2018), which is inspired by the fibers and synapse mechanisms in neural system, should be explored as an alternative on this task, which would portray the localization auditory system in a more biological plausible way and probably produce more human-like performance.

Future work

As this is the first attempt that GASSOM has been applied to encode binaural speech, future studies should be conducted. The phase information (finer temporal structure or ITD for high frequency, more specifically) was omitted during the calculation of cochleagram, which might cause the degradation of localization performance. Even the degradation is not significant in our work which is ideal condition, it might help to resolve the reverberation of source source through precedent effect as finer temporal structure

is important to the precedent effect according to previous work. In future work, those ITD information can also be included. For example, we can train another GASSOM based on phase information of cochleagram, and estimate the location based on both two GASSOMs. The encoding of binaural stimuli into GASSOM is a non-trivial topic to be discussed.

In this work, we train the data with randomly picked speech and convolved with HRTFs from randomly picked directions. In future work, the directions might be selected with smoothly successive, and HRTF interpolation might be employed to compensate for directions not measured. Different training methods can be investigated as comparison with the developmental process of human being (such as modeling the process that infants learn to localize sound source with GASSOM). It is also suggested that training in more natural environment will lead to more human-like result.

Future study on the selection critical bands that influence the front-back discrimination should also be improved based on psycho-acoustic experiments as some results in the current experiments is hard to explain, that is the low front-back confusion rate between cluster 1 and cluster 4 in the last study, which suggests that there might be other more dominant critical bands that influence the perception of front-back. It can also be explored using statistical methods to find out the most significant critical bands on front-back determination.

In this study we only consider single sound source in anechoic environment without reflection or reverberation. The performance of such computational model can also be tested under multi-sound sources conditions with background noise. Multi-Condition Training (Ma et al. 2019) and precedent effect can also be included in such a model to make it more robust.

Bibliography

Algazi, V., Duda, R., Thompson, D. & Avendano, C. (2001), The CIPIC HRTF database, *in* 'Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)', IEEE, New Platz, NY, USA, pp. 99–102.
URL: *http://ieeexplore.ieee.org/document/969552/*

Au, J., So, R. & Horner, A. (2011), Toward mass-customizing up/down generic 3-d sounds for listeners: A pilot experiment concerning inter-subject variability, *in* 'Audio Engineering Society Convention 130', Audio Engineering Society.

Barlow, H. (2001), 'Redundancy reduction revisited', *Network: Computation in Neural Systems* **12**(3), 241–253.
URL: *https://www.tandfonline.com/doi/full/10.1080/net.12.3.241.253*

Barlow, H. B. (2012), Possible Principles Underlying the Transformations of Sensory Messages, *in* W. A. Rosenblith, ed., 'Sensory Communication', The MIT Press, pp. 216–234.
URL: *http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262518420.001.00*
9780262518420-chapter-13

Brown, G. J. & Cooke, M. (1994), 'Computational auditory scene analysis', *Computer Speech & Language* **8**(4), 297–336.
URL: *https://www.sciencedirect.com/science/article/pii/S0885230884710163*

Carlson, N. L., Ming, V. L. & DeWeese, M. R. (2012), 'Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus', *PLoS Computational Biology* **8**(7), e1002594.
URL: *https://dx.plos.org/10.1371/journal.pcbi.1002594*

Chandrapala, T. N. & Shi, B. E. (2015), 'Learning Slowness in a Sparse Model of Invariant

Feature Detection', *Neural Computation* **27**(7), 1496–1529.
URL: *https://direct.mit.edu/neco/article/27/7/1496-1529/8097*

Chen, B., Wang, H., Wei, Y. & So, R. H. (2020), Truth-to-estimate ratio mask: A post-processing method for speech enhancement direct at low signal-to-noise ratios, *in* 'ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 7509–7513.

de Boer, E. (1978), 'On cochlear encoding: Potentialities and limitations of the reverse-correlation technique', *The Journal of the Acoustical Society of America* **63**(1), 115.
URL: *http://scitation.aip.org/content/asa/journal/jasa/63/1/10.1121/1.381704*

Francl, A. & McDermott, J. H. (2022), 'Deep neural network models of sound localization reveal how perception is adapted to real-world environments', *Nature Human Behaviour* **6**(1), 111–133.
URL: *https://www.nature.com/articles/s41562-021-01244-z*

Gardner, B. & Martin, K. (1994), HRTF Measurements of a KEMAR Dummy-Head Microphone, Technical report, MIT Media Lab Perceptual Computing.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Pallett, D. S., Dahlgren, N. L., Zue, V. & Fiscus, J. G. (1993), 'TIMIT Acoustic-Phonetic Continuous Speech Corpus'. Artwork Size: 715776 KB Pages: 715776 KB Type: dataset.
URL: *https://catalog.ldc.upenn.edu/LDC93S1*

Goodman, D. F., Benichoux, V. & Brette, R. (2013), 'Decoding neural responses to temporal cues for sound localization', *eLife* **2**, e01312.
URL: *https://elifesciences.org/articles/01312*

He, Y., Wang, H., Chen, Q. & So, R. (2022), Harvesting partially-disjoint time-frequency information for improving degenerate unmixing estimation technique, *in* 'ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE.

Horner, A. B., Beauchamp, J. W. & So, R. H. (2006), 'A search for best error metrics to predict discrimination of original and spectrally altered musical instrument sounds', *Journal of the Audio Engineering Society* **54**(3), 140–156.

Horner, A. B., Beauchamp, J. W. & So, R. H. (2009), 'Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones', *The Journal of the Acoustical Society of America* **125**(1), 492–502.

Horner, A. B., Beauchamp, J. W. & So, R. H. (2011), 'Evaluation of mel-band and mfcc-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds', *Journal of the Audio Engineering Society* **59**(5), 290–303.

Horner, A., Beauchamp, J. & So, R. (2004), 'Detection of random alterations to time-varying musical instrument spectra', *The Journal of the Acoustical Society of America* **116**(3), 1800–1810.

Hui, J., Wei, Y., Chen, S. & So, R. H. Y. (2019), Effects of base-frequency and spectral envelope on deep-learning speech separation and recognition models., *in* 'INTERSPEECH', pp. 634–638.

Hyvarinen, A., Oja, E., Hoyer, P. & Hurri, J. (1998), Image feature extraction by sparse coding and independent component analysis, *in* 'Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)', Vol. 2, IEEE Comput. Soc, Brisbane, Qld., Australia, pp. 1268–1273.
URL: *http://ieeexplore.ieee.org/document/711932/*

Hyvärinen, A. & Oja, E. (2000), 'Independent component analysis: algorithms and applications', *Neural Networks* **13**(4-5), 411–430.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0893608000000265*

İlik, B. (2018), Mems thin film piezoelectric acoustic transducer for cochlear implant applications, Master's thesis, Middle East Technical University.

Jin, C. T. & Carlile, S. (n.d.), 'Neural System Model of Human Sound Localization', p. 7.

Karunarathne, B., So, R. H. & Kam, A. C. (2014), Alarm vigilance in the presence of 80dba pink noise with negative signal-to-noise ratios, *in* 'Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014, Southampton, UK, 7-10 April 2014', CRC Press, p. 443.

Karunarathne, B., Wang, T., So, R. H., Kam, A. C. & Meddis, R. (2018), 'Adversarial relationship between combined medial olivocochlear (moc) and middle-ear-muscle

(mem) reflexes and alarm-in-noise detection thresholds under negative signal-to-noise ratios (snrs)', *Hearing Research* **367**, 124–128.

Keyrouz, F. & Diepold, K. (2006), An Enhanced Binaural 3D Sound Localization Algorithm, *in* '2006 IEEE International Symposium on Signal Processing and Information Technology', pp. 662–665. ISSN: 2162-7843.

Klein, D. J., König, P. & Körding, K. P. (2003), 'Sparse Spectrotemporal Coding of Sounds', *EURASIP Journal on Advances in Signal Processing* **2003**(7), 902061.
URL: *https://asp-eurasipjournals.springeropen.com/articles/10.1155/S1110865703303051*

Kohonen, T. (1990), 'The self-organizing map', *Proceedings of the IEEE* **78**(9), 1464–1480.
URL: *http://ieeexplore.ieee.org/document/58325/*

Kohonen, T. (1996), 'Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map', *Biological Cybernetics* **75**(4), 281–291.
URL: *http://link.springer.com/10.1007/s004220050295*

Kohonen, T., Kaski, S. & Lappalainen, H. (1997), 'Self-Organized Formation of Various Invariant-Feature Filters in the Adaptive-Subspace SOM', *Neural Computation* **9**(6), 1321–1344.
URL: *https://direct.mit.edu/neco/article/9/6/1321-1344/6082*

Lewicki, M. S. (2002), 'Efficient coding of natural sounds', *Nature Neuroscience* **5**(4), 356–363.
URL: *http://www.nature.com/articles/nn831*

Lies, J.-P., Häfner, R. M. & Bethge, M. (2014), 'Slowness and Sparseness Have Diverging Effects on Complex Cell Learning', *PLoS Computational Biology* **10**(3), e1003468.
URL: *https://dx.plos.org/10.1371/journal.pcbi.1003468*

Ma, N., Gonzalez, J. A. & Brown, G. J. (2019), 'Robust Binaural Localization of a Target Sound Source by Combining Spectral Source Models and Deep Neural Networks', *arXiv:1904.03006 [cs, eess]* . arXiv: 1904.03006.
URL: *http://arxiv.org/abs/1904.03006*

May, T., Ma, N. & Brown, G. J. (2015), Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues, *in* '2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, South Brisbane, Queensland, Australia, pp. 2679–2683.
URL: *http://ieeexplore.ieee.org/document/7178457/*

Middlebrooks, J. C., Makous, J. C. & Green, D. M. (1989), 'Directional sensitivity of sound-pressure levels in the human ear canal', *The Journal of the Acoustical Society of America* **86**(1), 89–108. Publisher: Acoustical Society of America.
URL: *https://asa.scitation.org/doi/abs/10.1121/1.398224*

Mlynarski, W. (2014), 'Efficient coding of spectrotemporal binaural sounds leads to emergence of the auditory space representation', *Frontiers in Computational Neuroscience* **8**, 26.
URL: *https://www.frontiersin.org/article/10.3389/fncom.2014.00026*

Mo, R., So, R. H. & Horner, A. (2016), 'An investigation into how reverberation effects the space of instrument emotional characteristics', *Journal of the Audio Engineering Society* **64**(12), 988–1002.

Moerel, M., De Martino, F., Uğurbil, K., Yacoub, E. & Formisano, E. (2015), 'Processing of frequency and location in human subcortical auditory structures', *Scientific Reports* **5**(1), 17048.
URL: *http://www.nature.com/articles/srep17048*

Olshausen, B. A. & Field, D. J. (1996), 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature* **381**(6583), 607–609.
URL: *http://www.nature.com/articles/381607a0*

Palakal, M., Murthy, U., Chittajallu, S. & Wong, D. (1995), 'Tonotopic representation of auditory responses using self-organizing maps', *Mathematical and Computer Modelling* **22**(2), 7–21.
URL: *https://linkinghub.elsevier.com/retrieve/pii/089571779500107D*

Pavão, R., Sussman, E. S., Fischer, B. J. & Peña, J. L. (2020), 'Natural ITD statistics predict human auditory spatial perception', *eLife* **9**, e51927.
URL: *https://elifesciences.org/articles/51927*

Peterson, A. J. & Heil, P. (2020), 'Phase Locking of Auditory Nerve Fibers: The Role of Lowpass Filtering by Hair Cells', *The Journal of Neuroscience* **40**(24), 4700–4714.
URL: *https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2269-19.2020*

Rayleigh, L. (1875), 'On Our Perception of the Direction of a Source of Sound', *Proceedings of the Musical Association* **2**, 75–84. Publisher: [Royal Musical Association, Taylor & Francis, Ltd.].
URL: *http://www.jstor.org/stable/765209*

Schnupp, J. W. H., Mrsic-Flogel, T. D. & King, A. J. (2001), 'Linear processing of spatial cues in primary auditory cortex', *Nature* **414**(6860), 200–204.
URL: *http://www.nature.com/articles/35102568*

So, R. H. Y., Ngan, B., Horner, A., Braasch, J., Blauert, J. & Leung, K. L. (n.d.), 'Toward orthogonal non-individualised head-related transfer functions for forward and backward direct', p. 16.

So, R., Leung, N., Braasch, J. & Leung, K. (2006), 'A low cost, non-individualized surround sound system based upon head related transfer functions: An ergonomics study and prototype development', *Applied ergonomics* **37**(6), 695–707.

Stecker, G. C. & Gallun, F. J. (2012), 'Binaural Hearing, Sound Localization, and Spatial Hearing'.

Thuillier, E., Gamper, H. & Tashev, I. J. (2018), Spatial Audio Feature Discovery with Convolutional Neural Networks, *in* '2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, Calgary, AB, pp. 6797–6801.
URL: *https://ieeexplore.ieee.org/document/8462315/*

Vecchiotti, P., Ma, N., Squartini, S. & Brown, G. J. (2019), 'End-to-end Binaural Sound Localisation from the Raw Waveform', *arXiv:1904.01916 [cs, eess]* . arXiv: 1904.01916.
URL: *http://arxiv.org/abs/1904.01916*

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C., Joris, P. X., Heinz, M. G. & Plack, C. J. (2019), 'The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints', *Hearing Research*

377, 109–121.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0378595518305604*

Wang, J., Wang, J., Qian, K., Xie, X. & Kuang, J. (2020), 'Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition', *EURASIP Journal on Audio, Speech, and Music Processing* **2020**(1), 4.
URL: *https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-020-0171-y*

Wenzel, E. M., Arruda, M., Kistler, D. J. & Wightman, F. L. (1993), 'Localization using nonindividualized head-related transfer functions', *The Journal of the Acoustical Society of America* **94**(1), 111–123.
URL: *http://asa.scitation.org/doi/10.1121/1.407089*

Wenzel, E. M., Wightman, F. L. & Kistler, D. J. (1991), Localization with non-individualized virtual acoustic display cues, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91', ACM Press, New Orleans, Louisiana, United States, pp. 351–359.
URL: *http://portal.acm.org/citation.cfm?doid=108844.108941*

Wightman, F. L. & Kistler, D. J. (1989), 'Headphone simulation of free-field listening. I: Stimulus synthesis', *The Journal of the Acoustical Society of America* **85**(2), 858–867.
URL: *http://asa.scitation.org/doi/10.1121/1.397557*

Wijesinghe, L. P., Wohlgemuth, M. J., So, R. H. Y., Triesch, J., Moss, C. F. & Shi, B. E. (2021), 'Active head rolls enhance sonar-based auditory localization performance', *PLOS Computational Biology* **17**(5), e1008973.
URL: *https://dx.plos.org/10.1371/journal.pcbi.1008973*

Willert, V., Eggert, J., Adamy, J., Stahl, R. & Korner, E. (2006), 'A Probabilistic Model for Binaural Sound Localization', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **36**(5), 982–994. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).

Wu, J., Chua, Y. & Li, H. (2018), A Biologically Plausible Speech Recognition Framework Based on Spiking Neural Networks, *in* '2018 International Joint Conference on Neural

Networks (IJCNN)', IEEE, Rio de Janeiro, pp. 1–8.
URL: *https://ieeexplore.ieee.org/document/8489535/*

Yost, W. A. (2016), 'Sound source localization identification accuracy: Level and duration dependencies', *J. Acoust. Soc. Am.* p. 7.

Yost, W. A. (2017), 'Sound source localization identification accuracy: Envelope dependencies', *The Journal of the Acoustical Society of America* **142**(1), 173–185.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5509470/*

Yost, W. A., Loiselle, L., Dorman, M., Burns, J. & Brown, C. A. (2013), 'Sound source localization of filtered noises by listeners with normal hearing: A statistical analysis', *The Journal of the Acoustical Society of America* **133**(5), 2876–2882.
URL: *http://asa.scitation.org/doi/10.1121/1.4799803*

Yost, W. A. & Zhong, X. (2014), 'Sound source localization identification accuracy: Bandwidth dependencies', *The Journal of the Acoustical Society of America* **136**(5), 2737–2746.
URL: *http://asa.scitation.org/doi/10.1121/1.4898045*

Zhou, L., Ma, K., Wang, L., Chen, Y. & Tang, Y. (2019), 'Binaural Sound Source Localization Based on Convolutional Neural Network', *Computers, Materials & Continua* **60**(2), 545–557.
URL: *http://www.techscience.com/cmc/v60n2/28382*

Appendix A

ANOVA tables

Table A.1: ANOVA table for the main effect of ICA and GASSOM on Fisher information about directions (see Chapter 3 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	0.0034	1	1.4e3	4e-92
Error	4.8e-4	198		
Total	0.0039	199		

Table A.2: ANOVA table for the main effect of ICA and GASSOM on Disorder Indices averaged across all the basis functions (see Chapter 3 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	0.0099	1	548.28	5.91e-59
Error	0.00358	198		
Total	0.01348	199		

Table A.3: ANOVA table for MAE of GASSOM model with different map size on localization error in degree (see Chapter 4 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	1.6686	3	8.5686	0.0013
Error	1.0386	16		
Total	2.7072	19		

Table A.4: Two-way ANOVA table for MAE with different Chunk length and Chunk shift. The columns correspond to chunk length main effect and row correspond to chunk shift main effect (see Chapter 4 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	0.4217	2	8.95	0.0003
Rows	0.04913	2	1.06	0.3526
Interaction	0.64578	4	6.86	0.0001
Error	1.90732	81		
Total	3.02452	89		

Table A.5: Two-way ANOVA table for MAE with different Center frequencies and bandwidths. The columns correspond to center frequency main effect and rows correspond to bandwidth main effect (see Chapter 5 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	2.5813	3	18.1211	2.47e-7
Rows	1.9402	2	23.1559	3.44e-7
Interaction	2.2012	6	8.7632	6.60e-6
Error	1.5107	36		
Total	7.9234	47		

Table A.6: Two-way ANOVA table for MAE with different Center frequencies and Duration for narrow band. The columns correspond to center frequency main effect and row correspond to duration main effect (see Chapter 5 for more details).

Source	*SS*	*df*	*F*	*p*
Columns	9.7489	2	30.63	0
Rows	0.0001	2	0	0.9996
Interaction	0.0009	4	0	1
Error	4.2972	27		
Total	14.0472	35		

Table A.7: Two-way ANOVA table for MAE with different Center frequencies and Duration for broad band. The columns correspond to center frequency main effect and row correspond to duration main effect (see Chapter 5 for more details).

Source	SS	df	F	p
Columns	0.0306	3	5.21	0.0043
Rows	0.0002	2	0.06	0.9443
Interaction	0.0003	6	0.03	0.9999
Error	0.0704	36		
Total	0.1015	47		

Table A.8: ANOVA table for front-back confusion rate of baseline HRTF and optional HRTFs from clustering analysis (see Chapter 6 for more details).

Source	SS	df	F	p
Columns	2.8373	1	18.14	5.13e-5
Error	13.7675	88		
Total	16.6048	89		

Table A.9: ANOVA table for back-front confusion rate of baseline HRTF and optional HRTFs from clustering analysis (see Chapter 6 for more details).

Source	SS	df	F	p
Columns	0.2767	1	7.62	0.007
Error	3.1956	88		
Total	3.4723	89		

Table A.10: ANOVA table for front-back confusion rate of different HRTF clusters. The model is trained with HRTF from cluster 1 and tested with all HRTFs, of which the results are grouped according to the cluster index (see Chapter 6 for more details).

Source	SS	df	F	p
Columns	1.9106	4	3.41	0.0124
Error	11.6216	83		
Total	13.5322	87		

Appendix B

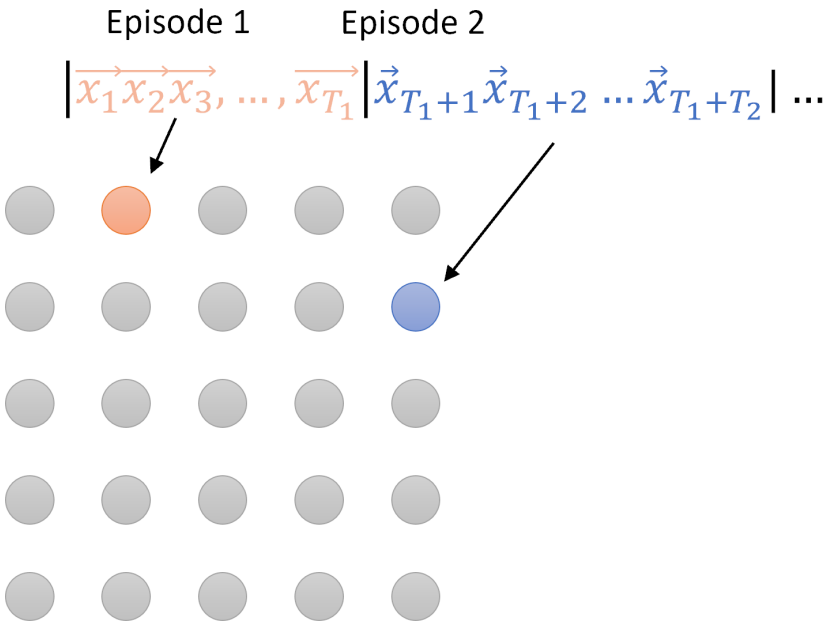
GASSOM Introduction

GASSOM (Chandrapala & Shi 2015) is a generative version of ASSOM (Kohonen et al. 1997), therefore ASSOM should be introduced first.

ASSOM

ASSOM consists of a 2D map with nodes (basis functions) on squared lattice, of which each is associated with a subspace $B_i \in \Re^{n \times h}$ that is spanned by h vectors of the same dimension as input, suppose to be n (e.g. with binaural stimuli of 4ms and 44.1kHz sampling frequency, the input dimension n is 352.). Figure B.1 for a demonstration of 5×5 ASSOM, where x_t is the input training sample and t is the time lag.

Figure B.1: The ASSOM contains 5×5 nodes on squared lattice, each node is associated with a subspace. ASSOM is trained with several episodes. Different episodes are explicitly labeled with different colors.



Given the input x_t, the projection onto ASSOM is defined as the squared inner product between the input vector and each subspace of ASSOM,

$$P_i = \sum_{h=1}^{H} (b_{ih}^T x_t)^2 \tag{B.0.1}$$

The projection is used as features for subsequent processing. In our work, the projection onto all the nodes is fed into DNN to predict the location. Another choice is 'Winner-takes-All' (WTA), which select the greatest projection across all nodes as input feature to next stage. However, the performance is not well in comparison with previous one. That's because one direction might corresponds to several nodes simultaneously. WTA only consider the most significant feature ignore the rest, which causes the degradation of performance.

During training, for each episode of input data, it computes the projection error $\hat{x}_t$ across each node,

$$\hat{x}_t = x_t - B_i B_i^T x_t \tag{B.0.2}$$

The node with least projection error is defined as the 'winner node'. Then the winner node and its neighborhood nodes are updated towards the corresponding data by a certain amount. The neighborhood function follows a multi-variate Gaussian distribution based on the Euclidean distance between the nodes on the map,

$$d_{i,c} = \mathcal{N}(I_i|I_c, \sigma^2 I) \tag{B.0.3}$$

However, during the training all the episodes should be explicitly labeled and fed to ASSOM subsequently. This is usually unrealistic in application and therefore prohibited ASSOM from encoding unknown natural stimuli. To solve this problem, GASSOM is proposed.

GASSOM

GASSOM is generative version of ASSOM, which is able to encode input stimuli without the requirement of explicitly labeling of different episodes. It assumes that each input sample is generated by one of the nodes and models the transition between different winner nodes with Hidden Markov Model (HMM). The generation node is indicated by $z_t \in \{0,1\}^S$, where $z_{ti} = 1$ means the current input x_t is generated by z_i. At each time t, only one node is set to 1 and the rest are set to 0. The transition probability is represented by a_{ij},

$$a_{ij} = P(z_{tj} = 1|z_{t-1,i} = 1) \tag{B.0.4}$$

The transition probability is model with a summation of Gaussian distribution and uniform distribution. The Gaussian distribution corresponds to node transition within the same episode, and uniform distribution corresponds to node transition between different episodes as the common pattern of different episodes changed significantly.

$$a_{ij} = \rho S^{-1} + (1-\rho)\frac{\mathcal{N}(l_j|l_i, \sigma_{Tr}^2)}{\sum_k \mathcal{N}(l_k|l_i, \sigma_{Tr}^2)} \tag{B.0.5}$$

where S is the number of nodes, ρ is the parameter to adjust the weight of uniform and normal distribution, whose covariance matrix is defined by a diagonal matrix σ_{Tr}^2.

After the fixation of GASSOM training parameters, the probability of hidden nodes based on the input observation can be formulated explicitly. Given the input data, the nodes that generate each input can be estimated both offline and online.

For offline estimation, the whole training data is available and Forward-backwards algorithm can be applied to estimate the winning node of the whole training data by maximizing the likelihood for each time lag recursively. For online estimation, only data on and before the current time lag is available, therefore the conditional probability based on previous time lag and current observation can be utilized to estimated the winner node subsequently from the very beginning.