Creation and Perception of Sound Source Width in Binaural Synthesis

Hengwei Su

Research Area of Creativity of Music and Sound
Division of Musicology and Music Studies
Tokyo University of the Arts

March 2019

Abstract

Techniques able to create and control extent of sound source imagery are necessary to produce complex auditory events as usually experienced in natural auditory environment. However, methods proposed in previous study are mainly for loudspeaker reproduction. In this dissertation, a method to create source width of monophonic source in binaural synthesis is proposed. Parameters of processing and its relevant effect on perception were investigated by conducting a series of subjective listening experiments. In addition, the proposed widening processing was implemented as a DAW effect for sound effects mixing to verify its effectiveness in the application of audio production.

The approach of the processing method was to process source signals by frequency bands division and distribution to achieve widen perceived source width. The frequency components of source signals were divided into sub-bands through filterbank. The sub-bands were distributed to different directions within the range of the width which was intended to synthesize. The distribution was performed by convolution with HRTFs of corresponding directions. The influence on the perceived source width of the processing method was investigated. Three listening experiments were conducted to examine various parameters and aspects of the processing. The results showed suggested that under some conditions the widening effect can successfully create and control width for monophonic source. Method to distribute frequency bands had a significant influence on effectiveness of widening and the localizations of the width. Bandwidth of the sub-band was found to have influence on the stability of the widening effect. Furthermore, investigations of naturalness of the signals after proposed were also conducted. The results suggested that there was no severe degradation by the processing method. However, there were still limitations including individual differences and dependence on source signals, suggesting that further investigation and improvement are needed.

The overall spatial impression when using the widening effect in binaural audio production was also investigated. The result suggested that the spatial impression could be improved when using the widening effect to synthesize width for monophonic sound effects. Although the effectiveness may depended on the nature of the sound effect and the subjective criterions of the listener.

概要

現実的に存在する音源は単純な点音源ではなく、広さ、大きさを持っている。そのため、より自然な、複雑な空間印象をもたらすため、音像を広げる処理が必要である。過去の音源の広がり処理に関する研究は、主にスピーカー再生を対象とする。そのため、この論文では、バイノーラルシンセシスにおけるモノラル音源の音像幅を合成する処理方法を提案した。処理のパラメータとその知覚に対する効果について調査するため、一連の主観聴覚実験を行った。また、提案された音像を広げるエフェクトをプラグインとして実装し、実際にオーディオ制作に応用する場合に、空間印象を向上させることができるかどうかを検証した。

処理のアプローチとしては、音源信号を帯域バンドの分解と分布の処理によって音像幅を広げる方法である。音源の周波数成分の分解はフィルタバンクによって行い、分解された各バンドは合成しようとする幅の範囲内に異なる方向に、頭部伝達関数との畳み込みで分布させた。音像を広げる処理が知覚された音像幅に与える影響を調査するため、三つの聴取実験を行い、様々なパラメータと側面から処理の有効性を検証した。結果により、処理が音像幅において有意な効果があり、そして音像幅とシンセシス幅との間に正の相関が見られた。それらの結果は、この処理方法がモノラル音源の幅を合成と制御することができると示唆した。帯域バンドの分布方法は、音像を広げる効果と音像幅の定位に有意な効果があった。帯域幅は広げる効果の安定性に影響を与えると示した。また、処理による自然さの変化も調査した。結果により、広げる効果による顕著な劣化がなかったと示唆した。しかし、個人差と音源特性への依存性があり、さらに調査と改善する必要があると意味した。

バイノーラルオーディオ制作に音像を広げる効果を使用する場合に、全体的な空間印象について調査した。結果により、モノラルの効果音の音像幅を合成することによって空間印象を向上させることができると考えられる。しかし、その効果が効果音の特性と、聴取者の主観的な嗜好による異なると示された。

Table of Contents

List of Figures **ix**

List of Tables **xiii**

Nomenclature **xv**

1 Introduction **1**

1.1 Aims and Scope of the Thesis 2

1.2 Organization of the Thesis 2

2 Background **5**

2.1 Binaural Technique 5

2.1.1 Introduction of Binaural Technique 5

2.1.2 Head-Related Transfer Function 6

2.1.3 Binaural Recording, Synthesis, and Reproduction 9

2.1.4 Binaural Auralization of Channel-based, Object-based, and Scene-based Audio 11

2.2 Sound Source Width 12

2.2.1 Auditory Perception of Extent 13

2.2.2 Perceived Source Width in Audio Reproduction and Sound Source Widening Effect 14

2.2.3 Source Widening Effect in Binaural Reproduction 16

2.3 Experiment Design and Analysis Method in Psychoacoustics Research . . . 17

2.3.1 Scheffé’s Pairwise Comparison 17

2.3.2 Analysis of Covariance 18

3 Sound Source Widening Effect for Binaural Synthesis **21**

3.1 Introduction 21

3.2 Experiment 1: Virtual source width in binaural synthesis with frequency-dependent directions . . . 22
3.2.1 Methods . . . 22
3.2.2 Results . . . 28
3.2.3 Discussion . . . 35
3.2.4 Conclusion . . . 36
3.3 Experiment 2: Frequency bands distribution for virtual source widening in binaural synthesis . . . 37
3.3.1 Methods . . . 37
3.3.2 Results . . . 40
3.3.3 Discussion . . . 47
3.3.4 Conclusion . . . 50
3.4 Experiment 3: The Source Widening Effect in Binaural Synthesis with Spatial Distribution of Frequency Bands . . . 51
3.4.1 Methods . . . 53
3.4.2 Results . . . 58
3.4.3 Discussion . . . 67
3.4.4 Conclusion . . . 69
3.5 Summary . . . 69

4 Spatial impression of source widening effect for binaural audio production 71
4.1 Introduction . . . 71
4.2 Method . . . 72
4.2.1 Plugin Algorithm . . . 72
4.2.2 Mixing Experiment . . . 73
4.2.3 Mixing Results . . . 75
4.2.4 Stimuli . . . 75
4.2.5 Listening Experiment . . . 75
4.3 Result . . . 76
4.4 Discussion . . . 76
4.5 Summary . . . 78

5 Summary 79

References 81

Appendix A List of Publications 85

Appendix B Experiment Instructions **87**

List of Figures

1.1 Research flow in this thesis 3

2.1 The interaural differences 7

2.2 The transfer function from the emitted position to the ear, and the transfer function from the source to the center of the head. 8

3.1 Widening processing algorithm to generate stimuli: (a) Stimulus with 60° synthesis width, (b) Stimulus with 30° synthesis width. 23

3.2 1/12 octave smoothed spectrums of the three source signals. The power was normalized so that the maximum value of each signal equals to 0 dB. 24

3.3 Distribution result of different distribution method for stimuli with 60° synthesis width. The power of each band converted to dB respecting to the maximum value of each stimuli is represented by gray gradient color. For *order1*, *order2*, and *random* method, the power of stimuli of cello signals is represented. 26

3.4 Setup of subjective listening experiment 27

3.5 The GUI for *perceived width* evaluation. 28

3.6 Histogram of azimuth angles of sound source width and azimuth angles of centers. 4 subplots from top to bottom represents different source types, panels in horizontal directions represents different synthesis widths, and panels in vertical directions represents different distribution methods. The vertical lines (red) indicate mean azimuth angles of centers of source widths. 30

3.7 How the histogram was plotted based on the reported perceived widths . . . 31

3.8 Mean source widths of different synthesis width across responses of all participants, with error bars representing standard errors. Panels in horizontal directions represent different source types, and panels in vertical directions represent different distribution methods. 32

3.9 Mean center computed from experiment results (black) with error bars representing standard error, and mean center weighted by power (red circle) and A-weighting applied power (blue triangle). The significances of *t*-test analyzing whether the mean = 0 were marked by asterisks. 34
3.10 The frequencies of *in-head localization* and *sound source moving*. 35
3.11 GUI constructed in Cycling'74 Max for pairwise comparison in the listening experiment. 40
3.12 Average relative ratings of perceived widths as a function of synthesis width for three types of source signals. Symbols denotes different distribution method as described in the legend. 41
3.13 Relative perceived width as a function of $1 - \text{IACC}$ and Pearson's correlation coefficients between them for each source signal 45
3.14 Relative ratings of naturalness from subjective listening experiments. The upper panels are the relative naturalness of spatial attributes for cello and xylophone sources, and the lower panels represent naturalness of timbre. . . 46
3.15 Dendrogram of hierarchical cluster analysis. Height represents intergroup dissimilarity between two groups, the number of each individual observation represent each participant. Dotted lines show how groups were divided. . . 47
3.16 Average relative perceived width for each group. 47
3.17 Average relative naturalness of spatial impression for each group 48
3.18 Average relative naturalness of timbre for each group 49
3.19 The widening processing method. (a) Stimuli with 60° synthesis width with center at 0° azimuth. (b) Stimuli with 30° synthesis width with center at 15° azimuth. 55
3.20 Synthesis widths of each participants tested. Each dot indicates the synthesis width on horizontal axis was tested by the participant on vertical axis. . . . 56
3.21 The GUI for the timbre degradation evaluation. 59
3.22 Mean scores for *perceived width* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively. 60
3.23 Mean scores for *perceived width* ratings of stimuli of the pink noise source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively. 60
3.24 IACC of stimuli from each condition as a function of synthesis width 62

3.25 Mean scores for *timbre degradation* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively. 64

3.26 Mean scores for *spatial quality degradation* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively. 65

4.1 The processing algorithm of the widening effect plugin. The example is for the parameter of *sound source width* set to 30° and the *sound source center* set to 15°. 72

4.2 The interface of source widening plugin. 73

4.3 The setting for the mixing experiment. 74

4.4 GUI constructed in Cycling'74 Max for the subjective listening experiment. 76

4.5 Box plot of the ratings for each stimulus 77

List of Tables

2.1 Example of statements of 7-point scales 18

3.1 Statements of 7-point scales for the width evaluation and the naturalness evaluations. 39

3.2 ANOVA table of width ratings for cello source 41

3.3 Stimuli pairs with significant differences for cello source. * represents a 0.05 and ** represents a 0.01 significance level. 42

3.4 ANOVA table of width ratings for xylophone source 42

3.5 Stimuli pairs with significant differences for xylophone source. * represents a 0.05 and ** represents a 0.01 significance level. 43

3.6 ANOVA table of width ratings for white noise source 43

3.7 Stimuli pairs with significant differences for white noise source. * represents a 0.05 and ** represents a 0.01 significance level. 44

3.8 Analysis of Covariance Table for cello source with 0° center position 61

3.9 Analysis of Covariance Table for cello source with 15° center position . . . 61

3.10 The results of regression analysis. 63

3.11 Pearson correlation coefficient between $1-\text{IACC}$ and ratings of perceived widths for each participant under four conditions. The coefficients >0.5 are emboldened. 64

3.12 Analysis of Covariance Table of timbre degradation ratings for cello source at 0° center position 65

3.13 Analysis of Covariance Table of timbre degradation ratings for cello source at 15° center position 65

3.14 Analysis of Covariance Table of spatial quality degradation ratings for cello source at 0° center position 66

3.15 Analysis of Covariance Table of spatial quality degradation ratings for cello source at 15° center position 66

4.1 Values or ranges of values of the *sound source width* parameter of each sound effect used by the two participants. 75

Nomenclature

Roman Symbols

a_i	azimuth angle in degrees of the band distributed to
C	common transfer function
$*$	convolution operator
$y_l(t)$, $y_r(t)$	signals measured at the entrance of the left and right ear canals
E_R, H_L	transfer functions from the position of emitted sound to the right and left ears
H_R, H_L	head related transfer functions of right and left ears
max	maximum value
p_i	power of each band
$s(t)$	monophonic source signal
(i, j)	comparison pair in Scheffé pairwise comparison
x_{ijk}	corresponding value of the score for the kth judge for (i, j) pair in Scheffé pairwise comparison
x_{it}	the tth observation of covariate under ith level in analysis of covariance
Y_ε	"yard-stick" for the confidence interval of $\alpha_i - \alpha_j$ in Scheffé pairwise comparison
Y_{it}	the tth observation of dependent variable under ith level in analysis of covariance

Greek Symbols

α_i, α_j	main effects of i and j in Scheffé pairwise comparison
$\hat{\alpha}_i$, $\hat{\alpha}_j$	estimates of α_i, α_j in Scheffé pairwise comparison

β the slope of covariate in analysis of covariance

ε error probability of the confidence interval in Scheffé pairwise comparison

ε_{it} error in the model of analysis of covariance

μ represents the grand mean in the model of analysis of covariance

μ_{ij} the mean preference for i over j when presented in the order (i, j) in Scheffé pairwise comparison

π_{ij} the average preference for i over j in Scheffé pairwise comparison

$\hat{\pi}_{ij}$ the estimate of π_{ij} in Scheffé pairwise comparison

τ time lag

τ_i the effect of *i*th level of the categorical independent variable in analysis of covariance

Subscripts

L, R left and right channels of the ears

Style/Formatting

italic is used to indicate it a name of a database, or a name of a distribution method, evaluation term, or parameter used in experiments.

Quotation mark is used for emphasis and to indicate the first use of a term.

Abbreviations

ADM audio definition model

ANCOVA analysis of covariance

ANOVA analysis of variance

VR augmented reality

ASW apparent source width or auditory source width

BRIR binaural room impulse response

DAW digital audio workstation

HpTF	headphone transfer function
HRIR	head-related impulse responses
HRTF	head-related transfer function
IACC	interaural cross-correlation coefficient
IHL	in-head localization
ILD	interaural level difference
IPD	interaural phase difference
ITD	interaural time difference
OBA	object-based audio
STFT	short-time Fourier transform
VR	virtual reality
VST	Virtual Studio Technology

Chapter 1

Introduction

Binaural technique, which involves direct control of signals transferred into both ears of listeners, not only can solve the problem of spatial impression of headphone reproduction, but also has the ability to provide realistic auditory experiences, especially in the aspect of the 3D spatial acoustic reproduction [10]. Head-related transfer functions (HRTFs) encapsulate transmission properties from sound sources to a listener's ears and contain important perceptual spatial information. Thus, incorporating HRTFS into input signals sent directly into ear canals can achieve authentic reproduction of an acoustic space. Binaural synthesis, which virtualizes sound objects by convolving HRTFs, can provide more flexibility and possibility of interaction than traditional methods such as dummy-head or binaural recording and virtualization of loudspeakers setup for channel-based audio. Hence, it becomes increasingly important with the development of virtual reality (VR) and object-based audio (OBA).

An HRTF usually represents transmission data of a sound source from a single direction to the ear position in the free field. As a result, simply convolving monophonic audio signals with an anechoic HRTF can produce only a perceptually point-like sound image. However, in the real world sound sources are usually not simply point sources but have extent. Hence, source widening processing for monophonic sound objects is necessary to provide a more complex and realistic acoustic spatial impression in binaural synthesis.

The fast-evolving VR technique has drew attention to 3D audio, especially in headphone reproduction. Providing auditory information can assist recognition of a virtual space and connecting auditory perception to visual can make experience more realistic. Interactive audio and auralization of virtual sources are essential elements in VR system to achieve high degrees of realism. Hence, object-based audio rendering via binaural synthesis has became an important technique for application of VR and augmented reality (AR). The audio definition model (ADM) [34], which provides a standard of metadata for next generation audio content, for object-based format includes parameters related to source extent such as width, depth, and height. Thus, techniques for object-based audio rendering capable of expressing those parameters are demanded.

Among the attributes related to localization or the shape of sound objects, spatial extent is an important aspect of the overall spatial impression due to the strong ability of our auditory system to perceive the lateral size of the source. Auditory perception of spatial extent has been investigated in various fields such as room acoustics and audio production. For perceived source extent of a direct sound without reflections, research basically investigated the effect of frequency-dependent decorrelation or panning. Nevertheless, previous studies about source widening processing were mainly for loudspeaker reproduction. This study thus focuses on the perceived source width in headphone reproduction and source widening processing methods for binaural synthesis.

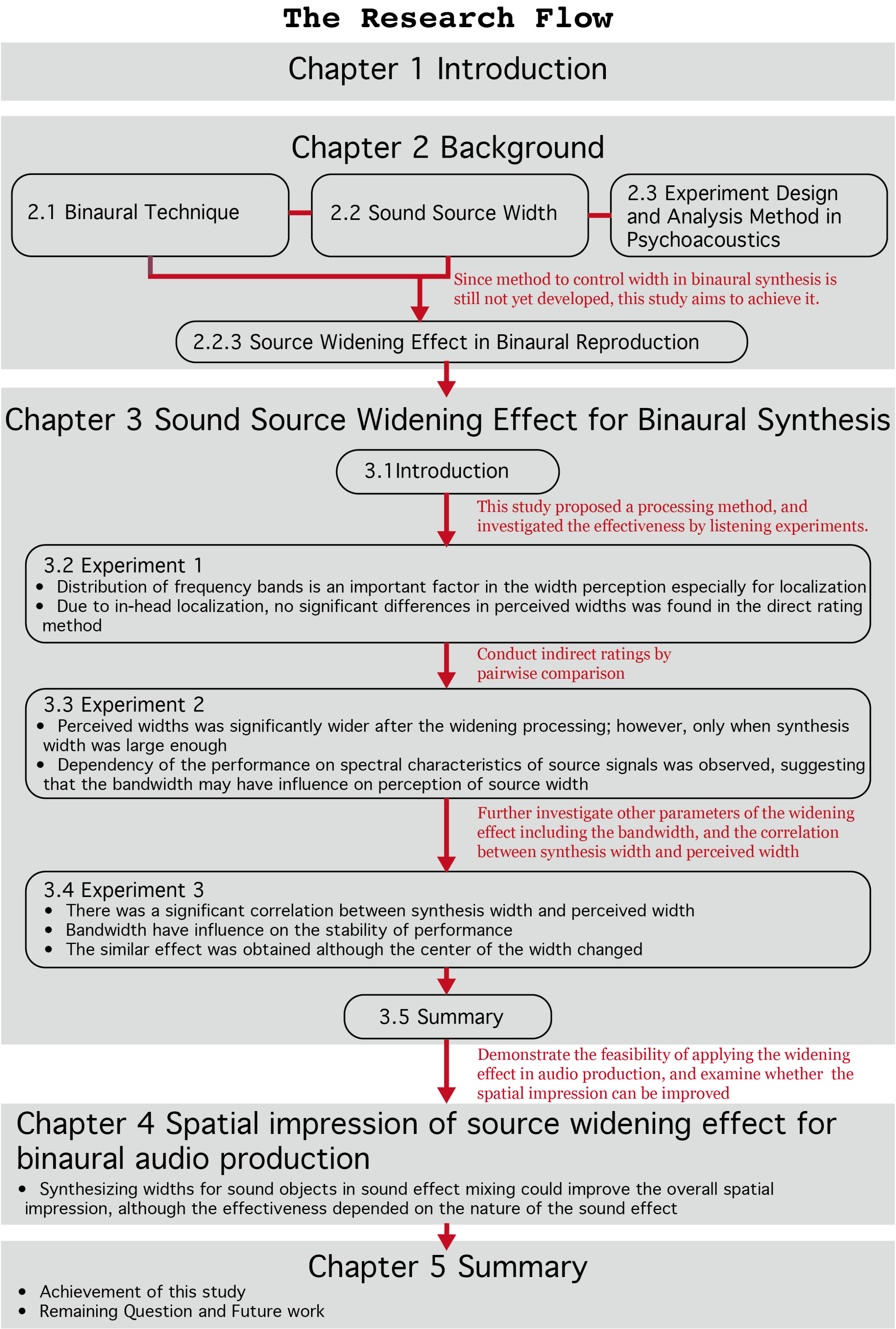
1.1 Aims and Scope of the Thesis

This dissertation aims to propose a method to control the source width in binaural synthesis. With the implementation of an approach of frequency band decomposition and spatial distribution to binaural synthesis, the effectiveness of widening processing was verified, and influences of the parameters of the processing on perceived source width were examined. Besides, the other aim is to investigate the effect of the widening processing on the spatial impression in the practical application of audio production.

In this dissertation the scope is in the synthesis and the perception of the width, which is the width produced by the direct sound, *i.e.* the sound source itself in the absence of room reflections. Also, only the lateral size of the sound is within the scope of this thesis.

1.2 Organization of the Thesis

The research flow and the organization of this thesis is displayed in Fig. 1.1. Chapter 2 provides backgrounds related to this dissertation, including an overview of binaural technique, reviews of studies about perceived source width, and basic concepts of analysis methods used in this thesis. In Chapter 3 there experiments investigating a source widening processing method and its effects on width perception are presented. In Chapter 4, a validation experiment for practical application of the widening processing to audio production is presented. The study is concluded in Chapter 5.



The Research Flow

Chapter 1 Introduction

Chapter 2 Background

2.1 Binaural Technique

2.2 Sound Source Width

2.3 Experiment Design and Analysis Method in Psychoacoustics

Since method to control width in binaural synthesis is still not yet developed, this study aims to achieve it.

2.2.3 Source Widening Effect in Binaural Reproduction

Chapter 3 Sound Source Widening Effect for Binaural Synthesis

3.1Introduction

This study proposed a processing method, and investigated the effectiveness by listening experiments.

3.2 Experiment 1

- Distribution of frequency bands is an important factor in the width perception especially for localization
- Due to in-head localization, no significant differences in perceived widths was found in the direct rating method

Conduct indirect ratings by pairwise comparison

3.3 Experiment 2

- Perceived widths was significantly wider after the widening processing; however, only when synthesis width was large enough
- Dependency of the performance on spectral characteristics of source signals was observed, suggesting that the bandwidth may have influence on perception of source width

Further investigate other parameters of the widening effect including the bandwidth, and the correlation between synthesis width and perceived width

3.4 Experiment 3

- There was a significant correlation between synthesis width and perceived width
- Bandwidth have influence on the stability of performance
- The similar effect was obtained although the center of the width changed

3.5 Summary

Demonstrate the feasibility of applying the widening effect in audio production, and examine whether the spatial impression can be improved

Chapter 4 Spatial impression of source widening effect for binaural audio production

- Synthesizing widths for sound objects in sound effect mixing could improve the overall spatial impression, although the effectiveness depended on the nature of the sound effect

Chapter 5 Summary

- Achievement of this study
- Remaining Question and Future work

Fig. 1.1 Research flow in this thesis

Chapter 2

Background

2.1 Binaural Technique

2.1.1 Introduction of Binaural Technique

The concept of binaural technique is to directly control the input signals to the auditory system, *i.e.* the acoustic signals at the eardrums of ears. It can be defined as all recording, analysis, synthesis, reproduction techniques which involve the signals directly replayed in the entrances of both ear canals of the listener [10].

Binaural hearing, as opposed to monaural hearing, can provide much more reliable acoustic cues for cognition of audio events, especially regarding aspects of spatial perception. This is because the auditory system can process information of the differences between input signals to two ears which are at different positions in the sound field [14]. If crucial binaural cues for auditory system are captured and provided properly, realistic sound fields can be reproduced. Hence, binaural technique is especially important for applications in the field of spatial audio.

To provide backgrounds of binaural techniques relevant to the studies in this thesis, this section firstly introduces head-related transfer function, which is the most crucial part in binaural technique. Then basic concepts for recording, reproduction, and synthesis techniques for binaural audio are given. Finally methods of binaural synthesis used for rendering various audio formats are overviewed.

2.1.2 Head-Related Transfer Function

Localization Cues

The ability of humans to recognize the direction of the sound source relies on the processing of information of the sound, which is called "localization cues," by auditory system [25]. Localization cues are classified as binaural cues and monaural cues. Binaural cues, or interaural cues, comes from comparing the differences between the two input signals at the two ears. Interaural level difference (ILD) and interaural time difference (ITD) are the two binaural cues and also considered the most important cues in horizontal plane. ITD is resulted from the difference in distances from the sound source to both ears, as illustrated in Fig. 2.1. The path to the contralateral ear (on the opposite side to the sound source) is longer than the path to the ipsilateral (on the same side to the sound source) ear, which leads to the difference in arrival times of the sound wave emitted from the sound source. Based on a simplified model assuming the geometry of the head is spherical, the ITD can be computed as following equation:

$$\tau = \frac{r(\theta + sin\theta)}{c} \tag{2.1}$$

where τ is the ITD, r is the radius of the head, θ is the azimuth of the source in radians, and c is the velocity of sound. ILD occurs due to the presence of the head of the listener, which serves as an obstacle causing attenuation of the incident sound wave. The sound pressure level at the contralateral ear is lower than the level at the ipsilateral ear. Both ITD and ILD vary depending on the incident direction of the sound, thus human auditory system can utilize the ITD and ILD to judge the direction of the sound. On the other hand, monaural cues are spatial information that can be resolved only by one ear, or the information is common to both ear. The most dominant monaural cues related to the localization of the direct sound is from the variation of spectrum of the signals depending on the incident direction of the sound, which is called "spectral cues". Spectral cues are mainly derived from the filtering effect of the head, which will be introduced in the next section.

Head-Related Transfer Function

A Head-related transfer function (HRTF) describes transmission properties from a point sound source to a position in the ear canal of a listener in free-field. The transfer function encapsulates filtering effect of head, pinna and torso of the listener corresponding to the particular direction. Thus, HRTF not only intuitively represents binaural directional cues of ILD and ITD, but also provides monaural spectral cues for localization, which have been known as primary cues for localization of elevation [5].

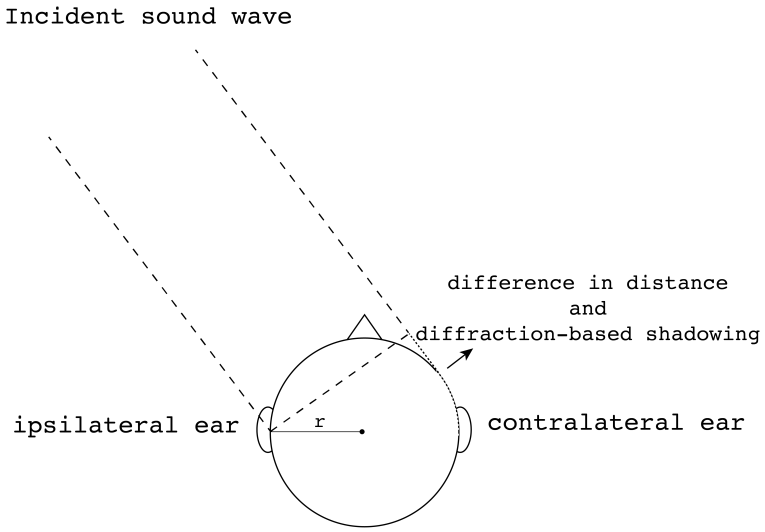


Fig. 2.1 The interaural differences

HRTF a commonly defined as the transfer function from the source to the ear divided by the transfer function from the source to the center of the head without the head being present, *i.e.*

$$\begin{aligned} H_R &= \frac{E_R}{c} \\ H_L &= \frac{E_L}{c} \end{aligned} \tag{2.2}$$

where H_R, H_L are the HRTFs of right and left ears, E_R, E_L are the transfer functions from the position of an emitted sound to the ear, c is velocity of sound, which is usually obtained by the transfer function from the source to the center of the head (Fig. 2.2). In this definition, the transfer functions of measurement apparatus, such as a microphone or loudspeaker, have been inherently compensated.

Encoding HRTFs into input signals sent directly to both ears can provide important perceptual spatial information and reproduce the acoustic space authentically. However, due to the anatomical differences of human, the HRTF varies largely between individuals. Using non-individual HRTF will cause problems such as localization inaccuracies, inside-head localization, front-back and up-down confusions, and timbral coloration [32].

Individualization of HRTF

Individual HRTFs can be acquired by direct acoustic measurements from human subjects [18]. Nevertheless, the measurement procedure is usually tedious and time-consuming, since specific equipment such as anechoic chamber, loudspeakers, and in-ear microphones are required, and the measurement and recalibration require a lot of time to obtain HRTFs

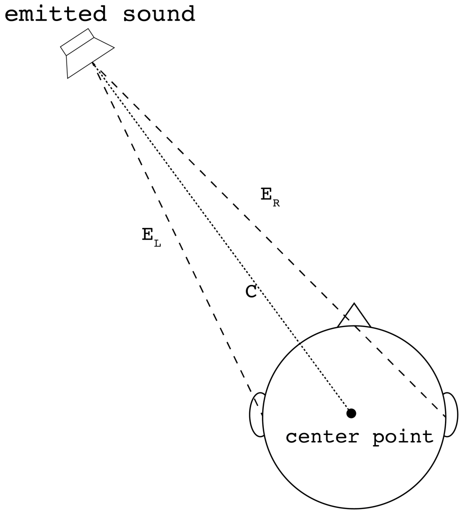


Fig. 2.2 The transfer function from the emitted position to the ear, and the transfer function from the source to the center of the head.

from different directions with sufficient resolution [8]. As the measurement is not always feasible for all researchers or engineers, a lot of HRTF databases of human subjects or artificial body measurements are provided by many organizations such as *KEMAR – the MIT Media Lab HRTF Database* [9], *LISTEN – the IRCAM HRTF Database*, [1] *the CIPIC Lab HRTF Database* [2], and *the RIEC HRTF Dataset* [37], which are all available online.

Apart from direct measurements, other methods to obtain individual HRTFs are also developed [38, 32], such as theoretical diffraction computation based on modeling the shape of head, ear, and torso with individual anthropometric data of the subject. With non-individual HRTFs from database, individualization can also be achieved through subjective selection, tuning, or matching according to similarity of anthropometric data.

Subjective selection is an easy and fast way to achieve individualization. Seeber and Fastl [30] presented a subjective selection method in which a two-step procedure was used to single out an optimal HRTF set. In the selection procedure, white noise pulses positioned at $-40°$, $-20°$, $0°$, $20°$, and $40°$ in the frontal horizontal plane by filtering with non-individual HRTFs were used. Multiple criteria relating to localization, externalization, and front-back confusion were provided for the selection. The results showed that the selection procedure could lower the variance of the localization responses, the number of inside-the-head localizations, localization error, and the number of front-back confusions.

However, if the size of HRTF catalogue is large, the selection procedure can become tedious. To obtain a suitable number of HRTF sets for subjective selection, Tama *et al.* [33] used *k*-means cluster analysis to obtain a subset of HRTF sets with maximal differences

[1] http://recherche.ircam.fr/equipes/salles/listen/index.html

as the best representative from a database. The impulse responses for both ears at two positions (0° and 180° azimuth, 0° elevation) from 62 subjects were used as data for analysis. The algorithm of k-means cluster analysis was as follows: First, k centers were tentatively determined for k clusters and each data point was assigned to the cluster with the nearest center. Then each center was recalculated as the average of the data points within the cluster, and reassignment of data and recalculation of centers were iterated. In their study, the index of the data nearest to each center was recored, and when $k = 5$ the results showed consistent indices after 1000 iterations. Thus, a subset of 5 HRTF sets was used for the subjective selection. The result of the listening test indicated that subjects who were offered a choice of HRTF had better front-back discrimination than subjects assigned an arbitrary HRTF. These studies suggested that by conducting subjective selections with a suitable size of HRTF catalogue, HRTF individualization can be achieved in some degree with good perceptual qualities while the time cost is low and the procedure is simple.

2.1.3 Binaural Recording, Synthesis, and Reproduction

Binaural Recording

Binaural recording is based on the concept that if the two input signals recorded in the ear canals of a listener are reproduced at the same positions, the listener can be provided with all information of the sound scene, the same as the listener would receive when in the real auditory experience [10]. Instead of a real listener, an artificial body, or a "dummy head" which simulates the human body shape with the relevant acoustical properties, is more often used for convenience. Usually a dummy head is equipped with two in-ear microphones, and the binaural signals can be directly recorded by placing it in the presumed listening position. However, for most cases the listener is not the same person used for recording or the dummy head usually just represents an "average listener." As a result, differences in HRTFs lead to inter-individual variances in the quality of binaural reproduction. Moreover, generally the recording is suitable only for headphone reproduction. For loudspeaker reproduction further processing is necessary.

Binaural Synthesis

HRTF characterizes the filtering of incident sound wave caused by the head and torso of the listener, which serves the same function of the human or the dummy head in the binaural recording. Hence, instead of recording with a physical head, binaural signals can also be generated by filtering an original signal with HRTFs. This method is called "binaural synthesis." The filtering is usually conducted by convolution with head-related impulse responses (HRIRs), which are the time-domain representation of HRTFs. The operation can be expressed by the following equations:

$$\begin{aligned}\mathrm{BinauralSignal}_{\mathrm{L}} &= s(t) * \mathrm{HRIR}_{\mathrm{L}} \\ \mathrm{BinauralSignal}_{\mathrm{R}} &= s(t) * \mathrm{HRIR}_{\mathrm{R}}\end{aligned} \tag{2.3}$$

where $*$ mark stands for the convolution operator, $s(t)$ is a monophonic source signal, and subscripts L and R represent left and right channels of the ears. As HRTFs of all possible directions are available, the directions of virtual sound sources can be controlled by using the corresponding HRIR [35].

Binaural synthesis has many advantages over binaural recording. For example, it provides more flexibility, since for different listeners different HRTFs can be used for synthesis such as individual HRTFs. Furthermore, with dynamic synthesis, updating the HRTF according to the change of direction of sound incidence due to the head movement is possible, which is closer to natural listening and can resolve front-back confusions [1]. Nowadays, binaural synthesis techniques have been widely developed and used in many applications, such as communication systems, virtual surround sound headphones, and virtual reality. Methods for generating binaural signals from common audio formats in audio engineering using binaural synthesis are introduced in the following section.

Binaural Reproduction

The binaural signals are usually reproduced via headphones, since the signals of the left and right channels can be replayed separately to the corresponding ears, and the listening environment can be isolated to achieve a better control of reproduction. However, non-flat headphone frequency response and acoustic coupling with listener's ears leads to spectral coloration. Thus, the equalization is necessary for authentic synthesis of binaural signals [10]. The headphone transfer function (HpTF) can be measured at the blocked ear canal and compensated for by inverse filtering. Nevertheless, acquisition of HpTF is tedious since it also varies with individual morphology and is very sensitive to headphone repositioning. Also, proper design of the equalization filters is also required to fulfill perfect compensation [28].

As opposed to headphone reproduction, if binaural signals are reproduced via loudspeakers, the left channel of the binaural signals transmits not only to the left ear but also to the right ear of the listener, and does as the right channel. This phenomenon, which is called "cross-talk," deteriorates authenticity of binaural reproduction and distorts spatial impression. In order to cancel cross-talk, an inverse filter to compensate for the transfer function between the contralateral loudspeaker and the respective ear can be applied [26].

This is called “transaural systems”. Recent development of transaural system is focused on solving problems such as small listening area, adaptation with head-movements, and the affect of reflections and reverberations of the room.

2.1.4 Binaural Auralization of Channel-based, Object-based, and Scene-based Audio

Binaural synthesis is an essential part of the technique of “auralization,” which is a term to describe the technique to create audible signals with numerical data, such as room acoustic modeling and virtual auditory display [36]. Auralization of various audio formats to binaural signals, *i.e.* “binauralization,” or binaural rendering, becomes increasingly demanded with the rapid development of 3D audio and virtual reality.

Binauralization of Channel-based Audio

Binaural synthesis was firstly widely applied for replaying the channel-based audio on headphone. Since stereophonic audio was originally created for loudspeaker reproduction, reproducing it directly with two channels of headphones distorts the sound scene and causes inside-head localization. To solve these problems, the loudspeaker setup, usually at -30° and 30° azimuth for stereophony, can be simulated by convolving channel signals with HRTFs of the directions of virtual loudspeakers. By doing this, the signal of each channel is reproduced as if from the corresponding virtual loudspeaker in the direction of the HRTF. Besides stereophony, virtualization of loudspeakers can also be applied to multichannel loudspeakers layout such as 5.1ch [22] or 7.1.4 [6] channel systems. This virtual surround technique can create 3D surround sound with just 2 channel headphone reproduction, and has attracted much attention and been commercialized for applications such as home theater system.

However, virtualization of loudspeakers to render channel-based audio inevitably suffers from problems occurred when using nonindividual HRTFs, which is usually the case for commercial audio. With timbre colorization and insufficient improvement in spatial impression such as lack of externalizaion, many studies report that listeners preferred original stereo or down-mixed stereo versions over binauralized versions. Using binaural room impulse responses (BRIRs) instead of HRTFs measured in the free-field is known to improve the timbre and provide better impression since the experience closer to that when listening to loudspeakers in a room [6]. BRIRs are HRIRs measured in a listening room, so early reflections and reverberation of the room are all so included in the impulse response, or these components can be added through room simulation.

Binauralization of Object-based Audio

Instead of reproducing channel-based signals by virtual speakers, recently developed object-based audio (OBA) provides an alternative approach to auralization sound scene into binaural signals. In OBA, as opposed to transmitting a set of channel signals which are specified for a particular loudspeaker setup, a set of sound objects and their metadata are transmitted. Those sound objects constitute a sound scene based on the metadata describing the attributes of sound objects such as spatial position and playback level. The rendering is done at the reproduction end according to the given reproduction setup to ensure that the spatial impression will not distort and can be optimized for different reproduction systems and listening environments. Hence, OBA format is independent of the reproduction platform [31]. Due to these benefits, OBA is considered an important format for future spatial audio distribution.

Binaural rendering of OBA can be done by directly convolving individual source signals of sound objects with HRTFs based on the position metadata. Directly virtualizing each sound object of a sound scene individually not only provides better spatial impression and more natural auditory experience, but also provides more flexibility and enables interaction. Hence, it is promising for applications in game audio, broadcast, and virtual reality, in which interactive experience is important.

Binauralization of Scene-based Audio

Ambisonics is a scene-based audio format which is also independent of reproduction systems, since its sound field is encoded by spherical harmonics and can be decoded according to loudspeaker setup [3]. With the development of ambisonics microphones, which can capture a sound scene in 3D representation, it has achieved popularity in 3D audio applications, especially in virtual reality. Binaural rendering of scene-based audio for headphone reproduction can be done by virtual loudspeakers as described above by firstly transforming Ambisonics format to loudspeakers feed and convolving those signals with HRTFs corresponding to loudspeakers positions.

2.2 Sound Source Width

In acoustic research many theories or models simplify sound sources as point sources. For example, an HRTF simply describes a transfer function from one source point to the point of an ear cannel of the listener. However, in real life sound sources are usually not point sources but have extent due to the physical size and radiation pattern of the source. In addition, the "perceived" source extent, *i.e.* the size of the sound image in auditory space of the listener as opposed to the size of the "real" source, also increases due to room reflections.

The *extent* is an important attribute and has a significant influence on the overall spatial impression [15]. Perception of the extent of a sound in auditory space has been studied in various areas from different aspects and in different definitions. In this thesis, only extent of one single source in the horizontal direction, that is, "sound source width" is discussed. This section firstly introduces the perception of sound source width, after which studies related to source widening effects are reviewed.

2.2.1 Auditory Perception of Extent

Although *extent* is usually thought as a localization-related attribute, the definition and perception of sound source width is more complicated than the "localization" attribute, which has absolute or specific values and definition and can be referenced to a real space with other perceptual dimensions such as vision. It has been well-established that auditory extent is related to frequency, intensity level, and temporal duration of a signal [20]. Perceived extent of a sound increases as the level or duration increases, and decreases as the frequency increases. It has been assumed to have connections with the experience in our daily life when encountering with naturally occurring acoustic sources, as sound sources with bigger size usually produce a higher-level, lower-frequency, longer-duration sounds. Thus, perception of extent is usually considered a "learned attribute," which means it is a relative and subjective impression of space like other spatial attributes, *e.g.* envelopment, spaciousness, reverberance and presence, which vary with individual apprehension.

In addition to frequency, level and duration, which are difficult to alter without changing other aspects of perception of sounds, the interaural difference in binaural hearing also has influence on the perceived width of the sound. There has been extensive research regarding the relationship between correlation, *i.e.* the similarity, between two input signals to the two ears and the perceived extent of the sound. Perrott and Buell [20] found that two uncorrelated noises replayed at two channels of headphones produced a sound image with size bigger than that of correlated noises. Kendall [13] proposed a method to create decorrelated signals and discussed its effects on spatial impression of the sound image. The term decorrelation means to process an audio source signal to lower the correlation with the original signal by transforming the waveforms, while maintaining certain aspects of the signal so that it still sounds the same as the original. By a pair of all-pass filters with random phase responses, a pair of decorrelated signals can be produced. One of the effects of decorrelation was stated that with two signals reproduced by stereo loudspeakers, the image width increases as the correlation decreases. The correlation can be statically described by the measure of interaural cross-correlation coefficient (IACC), which is the maximum absolute value of the normalized cross-correlation function between two ear signals:

$$\mathrm{IACC} = \max \left| \frac{\int_{t_1}^{t_2} y_l(t) y_r(t+\tau) dt}{\sqrt{\int_{t_1}^{t_2} y_l^2(t) dt \int_{t_1}^{t_2} y_r^2(t) dt}} \right| \quad \text{for} - 1\text{ms} < \tau < +1\text{ms} \tag{2.4}$$

where $y_l(t)$ and $y_r(t)$ represent signals measured at the entrance of the left and right ear canals, and τ is the time lag, which corresponds to the ITD when the maximum value is obtained [14]. Time lag between 1 ms and +1 ms is usually used based on the ITD of a completely lateral sound considering the size of the human head. It has been found that there is an inverse relationship between the IACC and perceived source width [16, 4].

In the field of concert hall acoustic, auditory perception of width has been extensively studied to develop a model to predict perceived source width based on IACC or other parameters [39]. In this context, usually a term of apparent source width (ASW, or auditory source width) is used instead, which describes the phenomenon that the perceived extent of the sound source is broadened to exceed its actual physical size due to the influence of early reflections. When a listener receives the direct sound and the reflections, which are mostly from directions different from the direct sound, the auditory system recognizes those sounds as one auditory event as long as the time delay is within certain thresholds. That is, sounds from different directions fuse together as one diffuse, or broadened, sound image. This "fusion" phenomenon also happens in the *precedence effect* [14].

However, even when the reflections of the room are absent, a sound source can still be perceived as having extent. Many sounds in the natural world, such as those made by leaves on a street blown by the wind, a piano, or the seashore sound, do not radiate like point sources and can be perceived as substantially extended. Due to the physical size of the source, multiple parts or positions of the source would radiate similar but not identical sounds. As long as those sounds share similar characteristics, they can be perceived as one sound source although they come from different directions.

2.2.2 Perceived Source Width in Audio Reproduction and Sound Source Widening Effect

If identical signals come from different directions, such as emitted from loudspeakers at different positions, those signals are summed up when arriving ear canals and would end up producing a "averaged" directional cue. In this situation, usually only a narrow sound image is produced at the center of gravity of those loudspeakers according to gain factors. This is how phantom source be generated in the amplitude panning of loudspeaker reproduction. It can also be interpreted as HRTFs of directions those identical signals from are averaged to rebuild a HRTF corresponding to the position of the phantom source, which is often called “summing localization” [3]. This is similar to the idea for HRTF interpolation using the method of computing a weighted average of two or more neighboring HRTFs [8]. On the other hand, if incoherent signals are emitted from loudspeakers at different positions, a spread sound image can be produced, until the coherence is too low that the auditory event disintegrates to separate sound images.

Based on this concept, Potard and Burnett [23, 24] used decorrelation to produce multiple uncorrelated point sources signals replayed by multichannel loudspeakers to control the perception of sound source extent. The results showed that sources with different extent could easily be perceived and discriminated by listeners. In addition, they proposed a method to alter the decorrelation in different frequency bands, which basically produces sources with frequency bands perceived in different positions with different spatial extents.

Deccorelation is also widely used in the traditional pseudo-stereo techniques to produce a widened sound image when replaying a monophonic signal via stereo loudspeakers. Zotter and Frank [40] proposed filter pairs to generate decorrelated signals and investigate such performance for phantom source widening in stereo loudspeaker reproduction. The filter algorithms are either phase or amplitude-based, introducing frequency-dependent differences in pairs of loudspeaker signals following a sine or cosine function. As opposed to random-phase Fourier-based FIR, such as the method proposed by Kendall [13] described in the previous section, these filters are deterministic, so can generate stable results to investigate the relationship between parameters, acoustic attributes, and perceived source width. By adjusting parameters these two filter implementation can control IACC, which has been shown to have correlation with perceived source width in previous results of listening tests [42]. This type of decorrelation method can also extend to multichannel filters and to application in Ambisonics format [41].

It should be noted that manipulating phase or amplitude spectrum is actually a kind of frequency-dependent panning to produce different IPD (interaural phase difference) or ILD in different frequency bands. However, the approach of decorrelation is usually known to suffer from the problem of spectral coloration.

Hirvonen and Pulkki [11] studied a different but similar approach by using bandpass noises in various frequency bands presented via different loudspeakers of a loudspeaker array from -22.5° to 22.5° azimuth on the horizontal plane to investigate the center of sound image and the perceived width. The results of listening test showed that the perceived width was less than half the actual width for all test cases, suggesting that frequency bands from different loudspeakers were perceived as fused together spatially. The stimuli used in their study can also be interpreted as broadband noises with directional cues, such as ITD and ILD, suggesting different localizations at different frequency bands. Signals with conflicting cues were found to produce diffuse, unsharp sound images [26].

To implement this concept as a method for synthesizing the perceived spatial extent for a monophonic input signal in auditory displays, Pihlajämaki *et al.* [21] revised previous explorations and established a algorithm which uses short-time Fourier transform (STFT) to decompose source signals into time-frequency bins and distribute them to loudspeakers from different directions. Different parameters related to spatial distribution and window size of STFT were examined to achieve an optimal quality of perception. Results indicated that the effect could depend on signal content and suggested that parameter tuning was required. Generally, this study demonstrated that distributing narrow frequency bands into space can create a spatially extended perception of sound source, and various distribution widths can be produced. The subjective preference and the naturalness were also investigated. The results of formal and informal listening tests indicated that this approach could maintain good timbral quality while achieving synthesis of spatial extent.

2.2.3 Source Widening Effect in Binaural Reproduction

Techniques able to create and control extent of sound sources without altering reverberation are necessary to produce complex auditory events as usually experienced in natural auditory environments, especially for 3D audio systems which aim to provide immersive and realistic spatial perception. However, methods described in the previous section are mainly for loudspeaker reproduction. To our knowledge there is still a lack of study about width perception or methods to control widths of sound objects in binaural synthesis, although it is necessary especially for VR applications when rendering object-based audio in binaural reproduction.

For implementation of source widening effect to binaural synthesis, an approach which distributes frequency components across different directions is intuitive and feasible. Instead of a large number of loudspeakers, which is impractical for general applications, what is required is only a set of HRTFs. The distribution can be easily done by convolving frequency components with HRTFs with proper spatial resolution, and the width and its localization can be easily controlled by utilizing HRTFs of various directions. Thus, this study implemented this approach to binaural synthesis and examined the effect on perceived source width.

The focus of this dissertation is on the widening effect and width perception of sound source itself in the absence of room reflections. Here “width” represents the spatial extent on the horizontal plane, as we focused only on the width in the frontal direction where humans are most sensitive regarding localization of sounds [17]. Since the localization mechanism of auditory system for elevation is quite different from that for azimuth, the spatial spread in the vertical direction is not within the scope of this study.

2.3 Experiment Design and Analysis Method in Psychoacoustics Research

Subjective listening experiment is the most common methodology in the field of psychoacoustic research, and also frequently used for audio quality evaluation in audio engineering. In subjective listening experiment, stimuli are prepared under different conditions according to the subject of investigation, and participants of the experiment are asked to perform evaluations according to the presented stimuli, then statistical analysis is conducted based on the evaluation data.

Analysis of variance (ANOVA) and student's t-test are frequently used analysis methods in psychological statistics to compare differences among means of two or multiple groups, such as responses under different treatments, to evaluate the effects of treatments.

In this dissertation, other analysis methods which are less commonly used are introduced.

2.3.1 Scheffé's Pairwise Comparison

For subjective evaluation the description can be in a direct, such as ratings with regard to a certain attribute, or indirect scaling through comparison of two stimuli. Pairwise comparison, which compares two out of all stimuli at a time for all possible pairs, can provide a easier way to judge, so is suitable for circumstances when perceptual differences between stimuli is small or the absolute quantity of the perception is difficult or impossible to evaluate.

Scheffé developed a method to analyze pairwise comparison experiments [29]. In such comparison, participants not only indicate which one of a pair they prefer with regard to the respective attribute, but also evaluate the preference on a scale which can be converted to a numerical score. An example of statements for a 7-point scale when comparing a fixed order pair of (i,j) is listed in the following table, and corresponding values of the scores are also shown.

Under the mathematical model Scheffé proposed, the corresponding value of the score for the kth judgement is x_{ijk}, the mean of x_{ijk} for all judgements can be used as the estimate of μ_{ij}, which represents the mean preference for i over j when presented in the order (i,j). Also, $-\mu_{ji}$ represents and the mean preference for i over j in the order (j,i). The average of

Table 2.1 Example of statements of 7-point scales

statements	numerical score
I prefer *i* to *j* strongly.	3
I prefer *i* to *j* moderately.	2
I prefer *i* to *j* slightly.	1
No preference.	0
I prefer *j* to *i* slightly.	-1
I prefer *j* to *i* moderately.	-2
I prefer *j* to *i* strongly.	-3

these two means is:

$$\pi_{ij} = \frac{1}{2}(\mu_{ij} - \mu_{ji}) \tag{2.5}$$

where π_{ij} denotes the average preference for i over j, which is

$$\pi_{ij} = \alpha_i - \alpha_j \tag{2.6}$$

where α_i and α_j can be considered the main effects. The estimate of α_i can be obtained by the means of $\hat{\pi}_{ij}$ for all j.

Based on this model, sums of squares for estimates can be computed, so ANOVA can be performed. The "yard-stick" Y_ε is then deduced based on the variance of the estimate $\hat{\alpha}_i - \hat{\alpha}_j$, and the confidence interval of $\alpha_i - \alpha_j$ under the confidence coefficient $1 - \varepsilon$ is:

$$\hat{\alpha}_i - \hat{\alpha}_j - Y_\varepsilon \leq \alpha_i - \alpha_j \leq \hat{\alpha}_i - \hat{\alpha}_j + Y_\varepsilon \tag{2.7}$$

If 0 is not included in this range, analysis suggests that the probability that there is difference between the main effect of i and j is $1 - \varepsilon$.

In Scheffé's method, each participant only judges only one time for a (i, j) pair in one order. This is not practical when the number of participants is small, which is the usual case for psychoacoustic experiments. For the experiment design that one participant judge all possible pairs in both orders, Ura's variation can be used [27]. In the modified model the term of individual difference is also included.

2.3.2 Analysis of Covariance

Analysis of covariance (ANCOVA) is an analysis method based on a model which combines regression and ANOVA. ANCOVA is used when there are not only categorical independent variables but also continuous independent variables, which are called "covariates." The dependent variable is assumed to have a linear relationship with the covariate, and the differences among levels of independent variables, *i.e.* the "effect" of different treatments, are to be analyzed. The ANCOVA model can be written as [7]:

$$Y_{it} = \mu + \tau_i + \beta x_{it} + \varepsilon_{it} \tag{2.8}$$

where τ_i is the effect of ith level of the categorical independent variable, x_{it} is the tth observation of covariate under ith level, Y_{it} is the tth observation of dependent variable under ith level, ε_{it} represents error, and μ represents the grand mean. The ANCOVA model is under the assumption that the slope of covariate β is the same for all levels of the categorical independent variable, which is called homogeneity of covariate regression coefficients, or "parallel lines model." A statistical test of this assumption can be conducted by testing the model that slope of covariate depends on categorical independent variable, *i.e.* β_i is used in the model instead. If the interaction term between the categorical independent variable and covariate is significantly different from zero, regression slopes are not the same and ANCOVA should not be performed.

Chapter 3

Sound Source Widening Effect for Binaural Synthesis

3.1 Introduction

As mentioned in the previous chapter, previous studies about perceived source width and source widening effect are mainly for loudspeaker reproduction. To propose a method to control source width in binaural synthesis, the approach which distributes frequency components across different directions as proposed in [11, 21] and described in the previous chapter, was implemented for binaural synthesis. The effects of source widening processing were investigated by conducting subjective listening experiments to investigate the perceived source width, naturalness, and spatial attributes. Parameters of the processing method were tested to investigate their influence on width perception. In this chapter, a series of three experiments conducted to achieve the above objectives is presented.

In this study, the intended source width of a monophonic sound to be synthesized, which can be controlled in the processing, is called "synthesis width." On the other hand, the source width perceived by listeners, which is the actual target to be controlled, is called "perceived width." In this study, the perceived width usually represents the ratings from the evaluation in subjective listening experiments.

In the first experiment, a processing method to implement frequency bands division and distribution was proposed, and the effectiveness of the processing method was investigated by directly evaluating the perceived width on the spatial coordinate. In the second experiment, an indirect ratings method was used to further examine the effect of the processing method. In the third experiment, the influence of other parameters of widening processing on the performance were also investigated, and the relationship between the synthesis width and the perceived width was examined to verify whether the processing method can control source width effectively.

3.2 Experiment 1: Virtual source width in binaural synthesis with frequency-dependent directions

In order to develop a method to control perceived width in binaural synthesis, the aim of this experiment is to verify whether the concept of distributing frequency bands of monophonic sources across different directions to achieve widened spatial extent can also be applied to binaural reproduction. Stimuli generated with different synthesis width and distribution methods were used, and subjective listening experiments were conducted to investigate the effects on the perception of source widths.[1]

3.2.1 Methods

Processing Algorithm of the Widening Effect for Binaural Synthesis

The processing algorithm of the widening effect for binaural synthesis is displayed in Fig. 3.1. To perform division and distribution of frequency components, source signals were firstly filtered by an FFT-based 1/3-octave filter bank. Each band was then convolved with HRTFs of directions within the intended source width range. For example, if the intended synthesis width is 60° with center at 0° azimuth on horizontal plane, the HRTFs of $-30°$, $-25°$, $-20°$,..., 25°, 30° azimuth and 0° elevation would be used for convolution. Finally, convolved signals were summed up for reconstruction of signals which were then used as stimuli in subjective listening tests. How frequency components are distributed spatially to different directions, *i.e.* how each bands was assigned to different HRTF to perform convolution, is determined by various distribution methods, which is described in the next section.

Stimuli

Three types of signals — including anechoic cello recording, anechoic xylophone recording, and Gaussian white noise — were used as source signals. The white noise was generated by random sampling from the standard normal distribution with a duration of 10 seconds with

[1] The result of this experiment was also published in *H. Su, A. Marui, T. Kamekawa, "Virtual Source Width in Binaural Synthesis with Frequency-Dependent Directions," presented at the Audio Engineering Society Convention 142 (2017)*

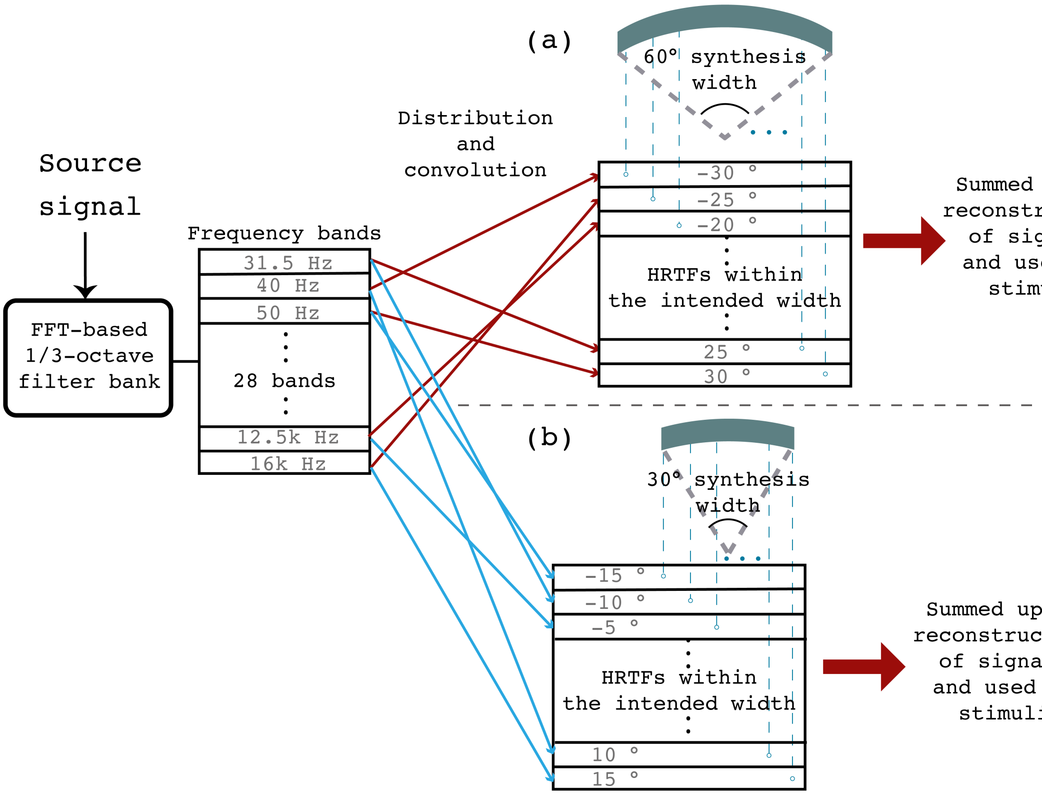

(a)
60° synthesis width
Distribution and convolution
Source signal
FFT-based 1/3-octave filter bank
Frequency bands
31.5 Hz
40 Hz
50 Hz
28 bands
12.5k Hz
16k Hz
-30 °
-25 °
-20 °
HRTFs within the intended width
25 °
30 °
Summed up for reconstruction of signals and used as stimuli
(b)
30° synthesis width
-15 °
-10 °
-5 °
HRTFs within the intended width
10 °
15 °
Summed up for reconstruction of signals and used as stimuli

Fig. 3.1 Widening processing algorithm to generate stimuli: (a) Stimulus with 60° synthesis width, (b) Stimulus with 30° synthesis width.

a 20 ms fade-in and fade-out, and normalization that the maximum absolute of samples was 0.999. The anechoic recordings were sampled from the Audio CD of Bang & Olufsen, Music for Archimedes [19] [2] in wave files. The duration of the cello recording was 21 seconds and that of the xylophone was 7 seconds. The 1/12 octave smoothed spectra of the three source signals are shown in Fig. 3.2.

Before processing, level alignment was performed to ensure that the loudness of three signals was perceptually the same. Three participants including the author adjusted the gain of three signals until they felt that all signals were at the same loudness. The gain could only be adjusted lower than units, *i.e.* lowering the level, to avoid clipping. The average values of the gains adjusted by the three participants were then used and all three participants agreed that there were no obvious misalignment of loudness among three signals. The average gain values were then applied to the three signals before processing.

HRTFs database of KEMAR dummy-head measured and provided by MIT Media Lab were used for synthesis [9]. Since this study focused only on source extent on the horizontal plane, only the HRTFs of 0° elevation were used. The synthesis widths under investigation

[2] http://pcfarina.eng.unipr.it/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm

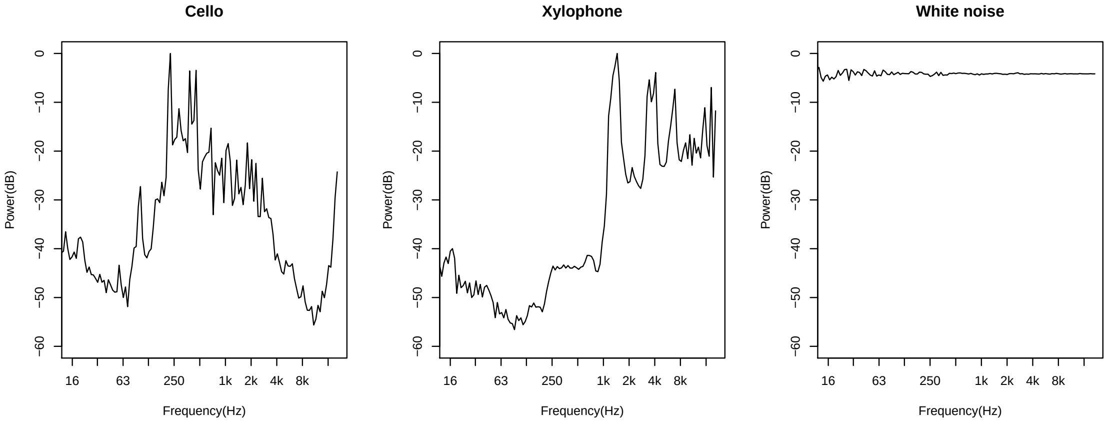


Fig. 3.2 1/12 octave smoothed spectrums of the three source signals. The power was normalized so that the maximum value of each signal equals to 0 dB.

were 10°, 20°, 40°, 60°, in azimuth angles. Since the HRTF database transfer functions were measured at intervals of 5° azimuth, numbers of HRTFs used were 3, 5, 9, 13, respectively. Centers of widths for all stimuli were set to 0° azimuth, *i.e.* the center of the front side.

To determine which HRTF each band was assigned to, four distribution methods were applied. Identical sets of HRTFs, *i.e.* numbers and directions of HRTFs, were used in the four methods, but only convolved with frequency bands in different orders. 28 frequency bands, 1/3-octave bands from 20 Hz to 22050 Hz, were evenly divided to HRTFs within the range of intended width, and the remainder were assigned to HRTFs which were nearest to the center. As an example, the distribution results of four distribution methods for synthesizing 60° width stimuli are given in Fig. 3.3. The four methods were:

1. *order1*: The bands from low frequency to high frequency were distributed from left to right in order. Until it reached the rightmost, the distribution was repeated again from left to right in ascending order of frequency, and the remain bands were distributed as close to center as possible.

2. *order2*: The bands were distributed from left to right in ascending order of frequency as in the *order1* method, but the higher band must be in the right position. Hence, for each HRTF there were multiple adjacent bands, and for HRTFs near center there were more bands than others.

3. *random*: The bands were randomly distributed to the same sets of HRTFS, but identical random patterns were used for all source types.

4. *powerorder*: The bands were distributed according to the spectrum characteristics of signals, which is explained in detail following.

The forth method *powerorder* was adapted since in a preliminary investigation of stimuli, we found that the centers of sound images shifted to rightward, especially for the *order1* and *order2* methods and for xylophone signal. It could be assumed as a consequence of that the frequency component of xylophone signal was almost entirely concentrated between 2 kHz to 10 kHz. The distribution methods like *order1* and *order2* would cause the localization of signals with narrow spectrum characteristics shifted from the center of the width (0°azimuth). Since in those methods the adjacent bands were distributed along azimuth, the bands with higher energy would focus on certain directions.

Therefore, a method that took spectral characteristics into account was also developed. The idea was that the band with higher mean power had more influence on perception, so the band with the highest power was firstly distributed to center, then the band with the second high power was distributed to $-5°$, and the third was distributed to $5°$, etc. That is, the bands were distributed with a descending order of power to HRTFs in order of proximity to center. When all HRTFs were used, the distribution started from 0° HRTF again until all bands were assigned. Consequently, the distribution results of three source types by *powerorder* method were different as shown in Fig.3.3, in which the gray gradient color represents the power of each band. The cello signal had rich components from 250 Hz to 700 Hz, and the xylophone signal was dominant from 2 k to 10 kHz, so those primary components were distributed near the center.

Stimuli with 0° synthesis width were also generated as references without widening processing. Source signals were convolved with 0° HRTF to synthesize sound at the 0° azimuth.

Besides stimuli described above, stimuli generated with multiple uncorrelated white noises located at different directions within intended source width ranges were also used as references. These references should be perceived with widest width, since the correlation between any pair among the signals is near 0; the perceived width of the noises ensemble should be as wide as the actual range of the directions of noise sources.

13 completely decorrelated white noises were generated by principal component analysis (PCA) of Gaussian white noises. Firstly, 13 randomly sampled Gaussian noises with 10 seconds duration were generated. PCA with 13 components were performed with orthogonal transformation of the 13 white noises to obtain a set of 13 linearly uncorrelated white noises. The correlation coefficients between any pair of the 13 noises after performing PCA were all on order of $10^{-15} \sim 10^{-17}$. The 13 noise sequences were then normalized so that the maximum amplitude equaled 0.999 as a safer value to avoid clipping. The spectra of 13

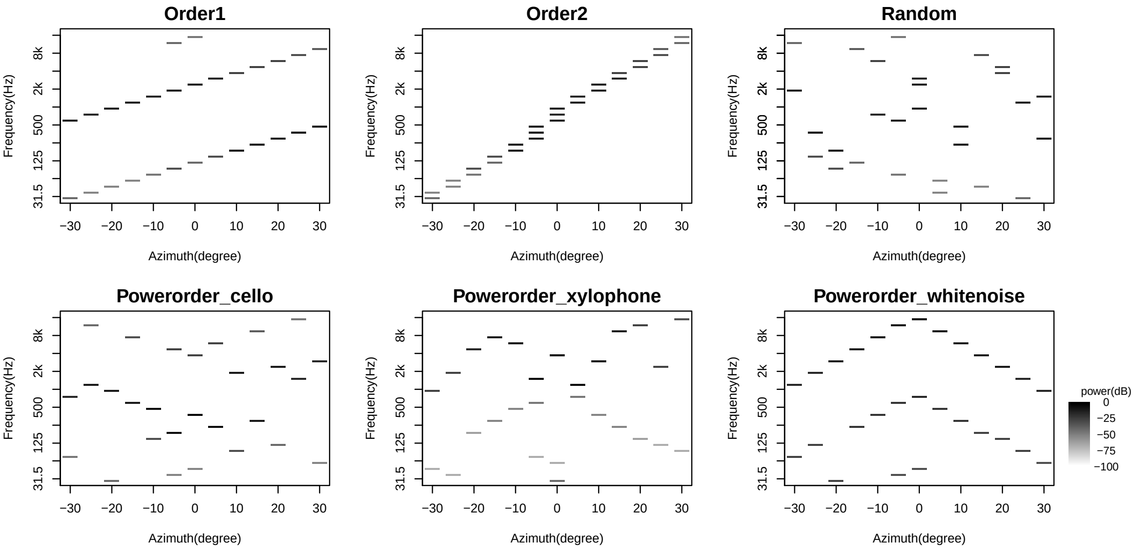


Fig. 3.3 Distribution result of different distribution method for stimuli with 60° synthesis width. The power of each band converted to dB respecting to the maximum value of each stimuli is represented by gray gradient color. For *order1*, *order2*, and *random* method, the power of stimuli of cello signals is represented.

signals were examined, and all of them retained the "white" spectral characteristic with similar average power. According to the number of HRTFs within intended source width, the same number of noises were randomly chosen, convolved with HRTFs respectively, mixed and divided by root of numbers of noises to normalize the level, since when uncorrelated signals are added up the level will be root of the sum of the level.

Environment

The test was conducted in a quiet studio room with proper acoustic treatment in Senju campus of Tokyo University of the Arts. The studio room was equipped with a computer, audio interface (Avid Rack 003), and five loudspeakers. Participants were asked to use a graphical user interface (GUI) constructed in Cycling'74 Max on a computer to answer questions and perform the experiment. Stimuli were replayed via GUI through an audio interface (Avid Rack 003) and emitted by headphone (SONY MDR-CD900ST). Since participants needed to estimate perceived widths in azimuth coordinates corresponding to virtual sound sources replayed via headphone, the task could be difficult to project perceived virtual sound images into real space. Especially headphone listening often suffers from the problem of in-head localization (IHL). Hence, to facilitate the task in which participants were asked to indicate where they perceived the width of virtual sound image by answering the coordinates, five loudspeakers located at −60°, −30°, 0°, 30°, 60° azimuth with respect to the position of the participant were used as directional references. In addition, a series of numbers to indicate azimuths in an interval of 5° was labeled on a black cloth between the loudspeakers. A photo of the setup of subjective listening experiment is shown in Fig. 3.4.

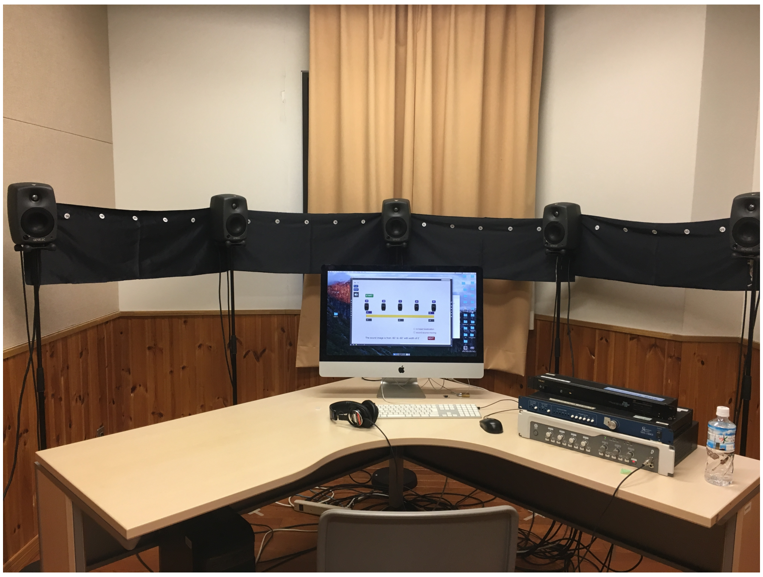
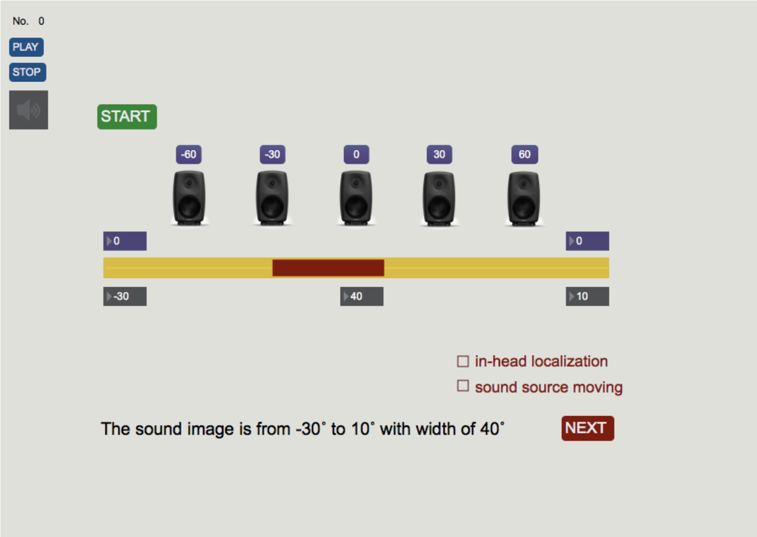


Fig. 3.4 Setup of subjective listening experiment

Procedure

The listening experiment was divided into two blocks to evaluate the *perceived width* and *naturalness* of stimuli respectively. In the block of *perceived width* evaluation, participants were asked to indicate sound source widths of stimuli in azimuth coordinate, *i.e.*, where they perceived extents of sound images. They used a bar on GUI to select the range of azimuths. The azimuth of leftmost and rightmost extent and the width of the range were shown in the GUI, as shown in Fig. 3.5. The selectable range was from $-60°$ to 60° in 5° intervals. In addition, if they perceived in-head localization and/or the localization of sound source moving while answering, they were asked to check the corresponding box to indicate that. Stimuli were played in loop so that participants were free to spend as much time as they needed for answering questions. In addition, in order to facilitate participants to answer the coordinates of virtual sound images, five reference sounds were provided. They were generated by a 500 ms white noise pulse convolved with HRTFs of $-60°$, $-30°$, 0°, 30°, 60° azimuth, *i.e.* the directions of the loudspeaker set in the experiment room. These reference sounds virtualized the white noises reproduced by the loudspeakers which they could utilize as visual assists. Participants were instructed to use them freely, and as they click the number of azimuth presented on the GUI, corresponding sound would be replayed.



□ in-head localization
□ sound source moving

The sound image is from -30° to 10° with width of 40°

Fig. 3.5 The GUI for *perceived width* evaluation.

In the block of *naturalness* evaluation, only stimuli synthesized from anechoic cello and xylophone sources were used, since naturalness was supposed to be only suitable for evaluating real sound sources but not for synthesized sound. Participants were asked to rate the stimuli according to the *naturalness of timbre* and the *naturalness of spatial attributes* respectively. A 7-point scale was used in the evaluations, in which 1 represented unnatural and 7 represented natural.

In each block, there were two sections in which participants repeated the same evaluation twice, *i.e.*, participants performed the same tasks on the identical set of stimuli but in different random orders. The order of the two evaluation blocks were also randomized for each participant.

Participants

A total of 8 participants participated the experiment. All participants were students in the Research Area of Creativity of Music and Sound of Tokyo University of the Arts, majoring in recording and psychoacoustics. All had previous experiences in listening experiments, and none reported any hearing disorder.

3.2.2 Results

Histogram of Width Response

Histograms of azimuth angles of *perceived source widths* obtained from the listening experiment are given in Fig. 3.6, under different synthesis widths, source types, and distribution methods.The X-axis represents the azimuth angles, which is within the answerable range from $-60°$ to $60°$ in discrete intervals of $5°$. Y-axis represents the frequency, which was computed by the number of times that the azimuth angle was included in the range of answered sound source width. Fig. 3.7 shows how the histogram was plotted. The upper part showed an example of the perceived widths as reported by each participant, and based on how many times each azimuth angle was included in the answered widths the histogram in the lower part could be obtained.

The total number of times that each stimulus was rated was 16, as 8 participants repeated the test twice, so frequency = 16 meant that in both sections all participants answered source widths including that azimuth angle. As the case of, for example, 60° decorrelated noise, frequency reaches 16 from $-25°$ to $25°$, which indicates that the sound image in that range was clearly perceived, so all subjects reached an unanimous consensus. On the other hand, if the frequency was low along all azimuth angles, as the case of xylophone source of 60° and *order2* method, it may suggest that the localization was ambiguous, so there was a lack of consensus between participants and repeated sections.

Since 0° stimuli were not processed by frequency bands division and distribution, for the same source type the results from the same stimuli are shown in the leftmost column in Fig. 3.6, so they are identical in the four panels of distribution methods. In addition, the results of 0° decorrelated noise and 0° white noise were actually from the same stimulus (a white noise convolved with 0° HRTF), identical data were shown in these two conditions.

Perceived Width

For decorrelated noises, there is a clear trend that the range of frequency = 16 increases as synthesis width increases. However, for other distribution methods and source types, there is no evidence showing that source widths increase accordingly with increase of synthesis width. Instead, shapes of histograms become flat, which may suggest that localization becomes diffuse as synthesis width increases. It can be assumed that the localization of source became unclear so sound source widths reported by participants did not fall to particular directions. Another reason could be that individual differences due to the non-individual HRTF increased as number of HRTFs used for synthesis increases with synthesis width.

Fig. 3.12 displays mean source widths by distribution methods and source types. The trend that *perceived width* increases with synthesis width can be observed only in the results of decorrelated noises.

To investigate if there were significant differences of the perceived source width among different synthesis widths, analysis of variance was performed on the *perceived width* using the data excluding deccorelated noises with three factors: synthesis width, source type,

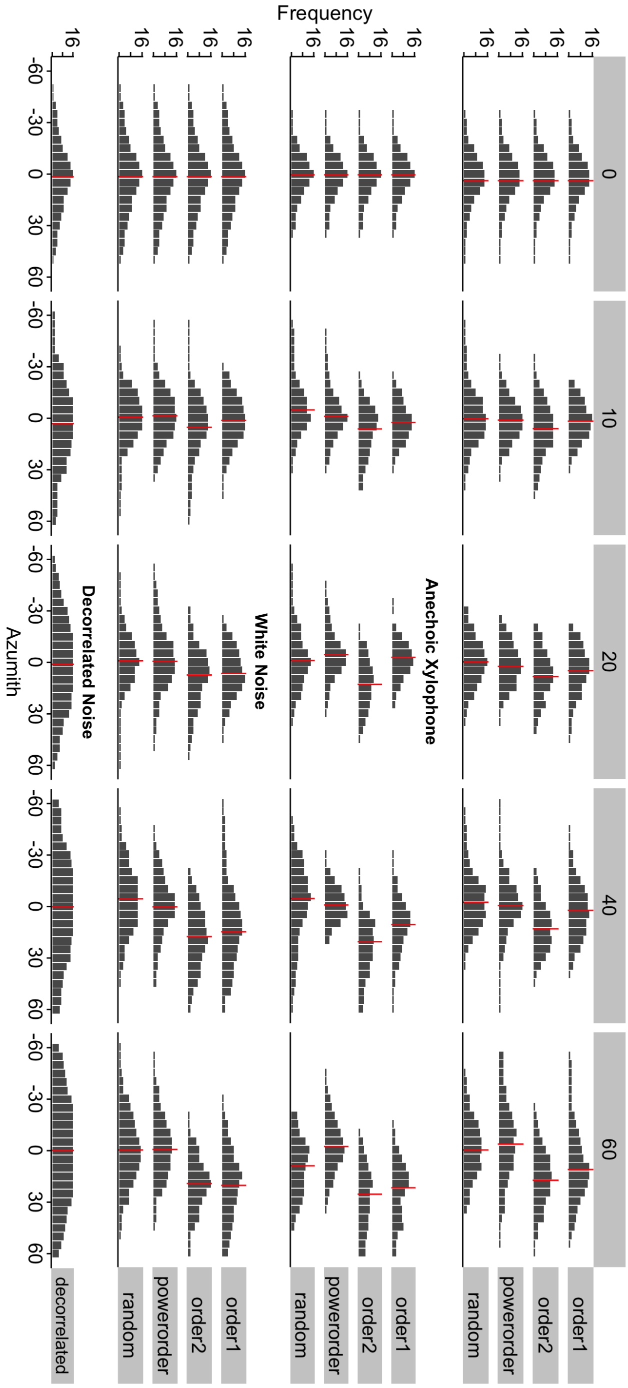


Fig. 3.6 Histogram of azimuth angles of sound source width and azimuth angles of centers. 4 subplots from top to bottom represents different source types, panels in horizontal directions represents different synthesis widths, and panels in vertical directions represents different distribution methods. The vertical lines (red) indicate mean azimuth angles of centers of source widths.

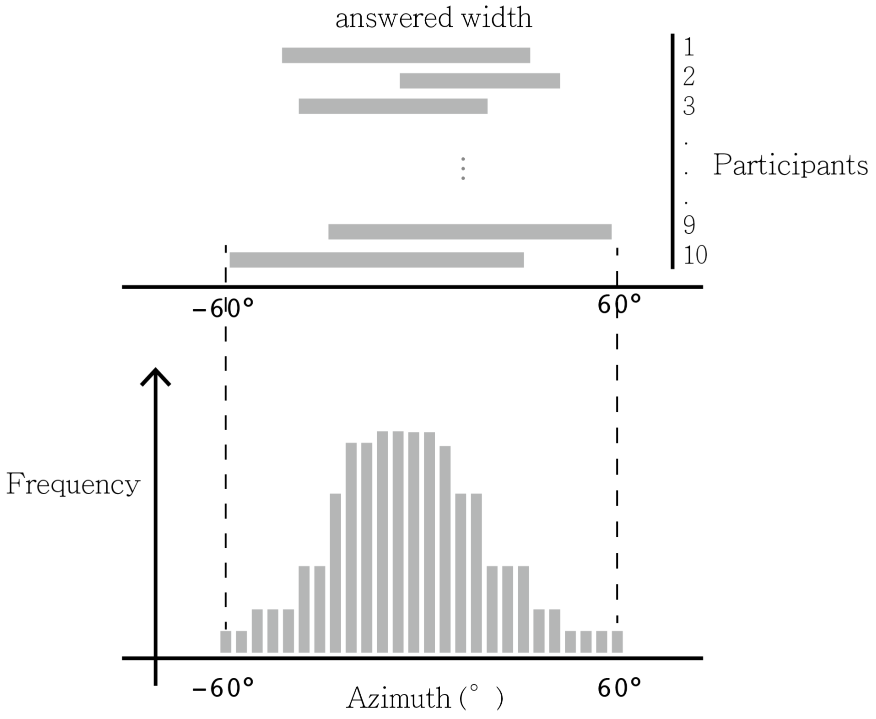


Fig. 3.7 How the histogram was plotted based on the reported perceived widths

and distribution method. The results indicated only factors of synthesis width (p=0.011) and source types (p=0.0078) showing significant effects. Therefore, a pairwise t-test was performed with respect to these two factors with p-value adjustment method of Holm. However, among different synthesis widths from 0° to 60°, there was no significant difference between any pair. On the other hand, there were significant differences between all pairs among three source types (p<.006). The average *perceived width* was largest for the white noise source (36.1°) and smallest for xylophone source (25.6°).

Since the processing method used in this study is based on the approach of distributing frequency bands across different directions, it can be assumed that source signals with broad spectral characteristics would be perceived wider than narrow ones such as xylophone source. However, it is noteworthy that the result of 0° stimuli still shows the same tendency. Therefore, it can be concluded that perceived source width can be influenced by nature of source types, for example instruments or noise, and by frequency characteristics or the duration of signals. It has been observed in previous studies that signals with high frequency and short duration would be perceived narrower in source width. This is the case for the xylophone source due to its percussive timbre with short duration and mainly high frequency spectral components.

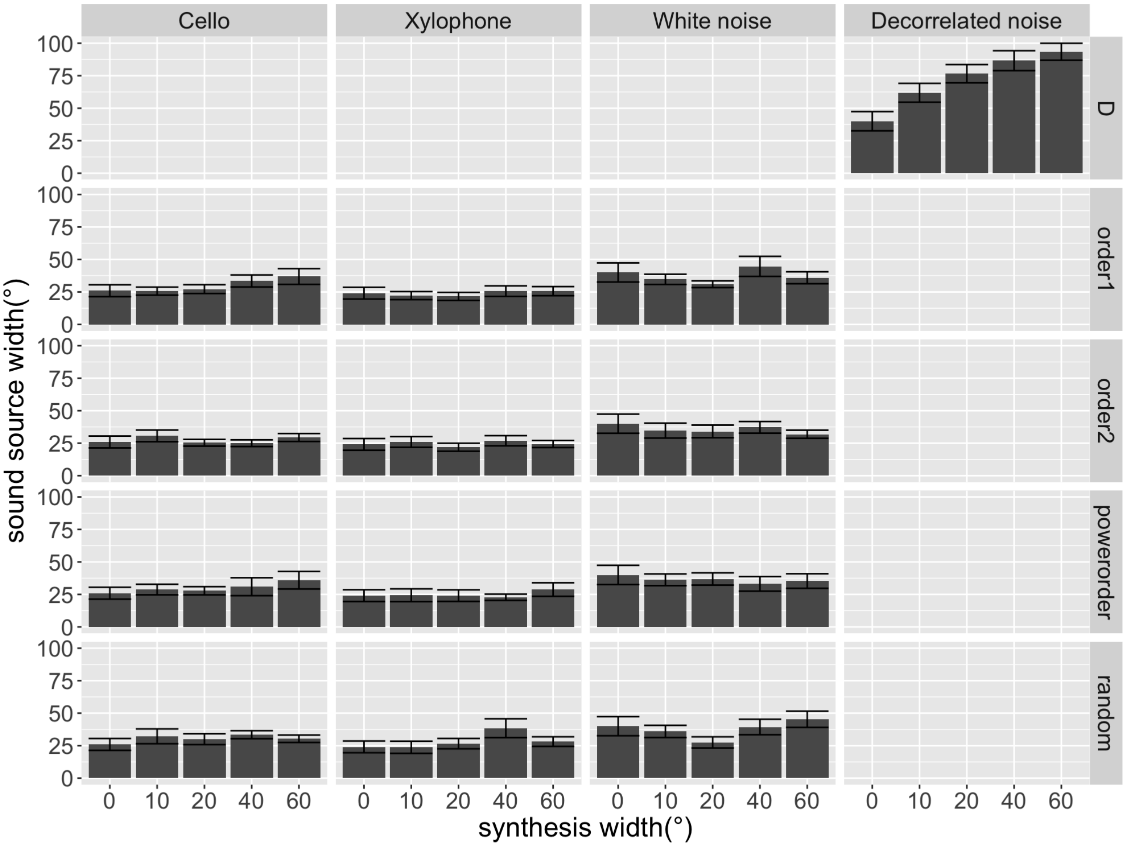


Fig. 3.8 Mean source widths of different synthesis width across responses of all participants, with error bars representing standard errors. Panels in horizontal directions represent different source types, and panels in vertical directions represent different distribution methods.

Correlation coefficients between synthesis width and *perceived width* under different distribution methods and source types are lower than 0.3 except for decorrelated noise (r = 0.52).

Localization of Perceived Source Widths

The localization of perceived width was evaluated by computing the center of the *perceived width* participants responded, *i.e.*, the means of the azimuths of the left end and the right end of the *perceived width*. Mean azimuth of centers of perceived source widths across all participants are indicated by vertical lines (red) in each panel in Fig. 3.6.

Since the centers of synthesis widths were all at 0° azimuth, the centers of perceived width should ideally be 0°. It is apparent that centers of sound source width shifted to the right for *order1* and *order2* methods, as can be seen in Fig. 3.6. The shifts become more obvious as synthesis width increases. In these two methods, bands were distributed from low to high frequency in ascending order of azimuth, so the bands with higher frequency, which may have a dominant influence on localization, were distributed to right. As a result, participants may perceive the sound source width shifted to the right side. Shifts are larger in *order2* than in *order1*, since for *order2* method the higher frequency band was always distributed to the direction of larger azimuth angle.

An ANOVA was performed to analyze centers of sound width respect to factors including synthesis widths, distribution methods, and source types. Effects of synthesis widths ($p<$ .0001), distribution methods ($p<$.0001), and all interaction effects ($p<$ .03) between two factors were found to be statistically significant. Pair-wise t-test shows that both for *order1* and *order2* methods there were significant differences from any other methods. There were no significant differences between *powerorder* and *random* methods, and between both these two methods and stimuli without widening processing (0° width stimuli).

Regarding the reason why higher frequency bands could have a dominant influence on localization, one assumption is related to the power of frequency bands, which is discussed in the previous stimuli section. Thus, the power of each band were computed, and the mean center weighted by the power of each band was computed as follows:

$$\text{mean center} = \frac{\sum_{i=1}^{28} p_i \times a_i}{\sum_{i=1}^{28} p_i}$$

where p_i is the power of each band, a_i is the azimuth angle in degrees of the band distributed to. Besides, A-weighting applied power was also used for computation to take the relative loudness perception of the human ear into account. The results are shown in Fig. 3.9, in which black points represent mean center of the experimental result, red circles represent weighted mean center, and blue triangles represent the weighted mean center computed by A-weighting applied power. Although the trends of some results from weighted means are in accordance with the experiment results, some inconsistency can still be observed. It suggests that the importance on localization perception of frequency bands may vary with frequency, so a more complicated model is necessary. In addition, in the *order1* and *order2* distribution methods, adjacent frequency bands were also distributed to adjacent directions, which could also have influence on localization perception of source width.

To investigate whether the mean azimuth of center was significantly different from 0°, a t-test was performed. The results are given in Fig. 3.9, in which mean centers significantly differ from 0° are denoted with asterisks. As can be seen, localizations of centers of width in *order1* and *order2* method shifted from 0° significantly.

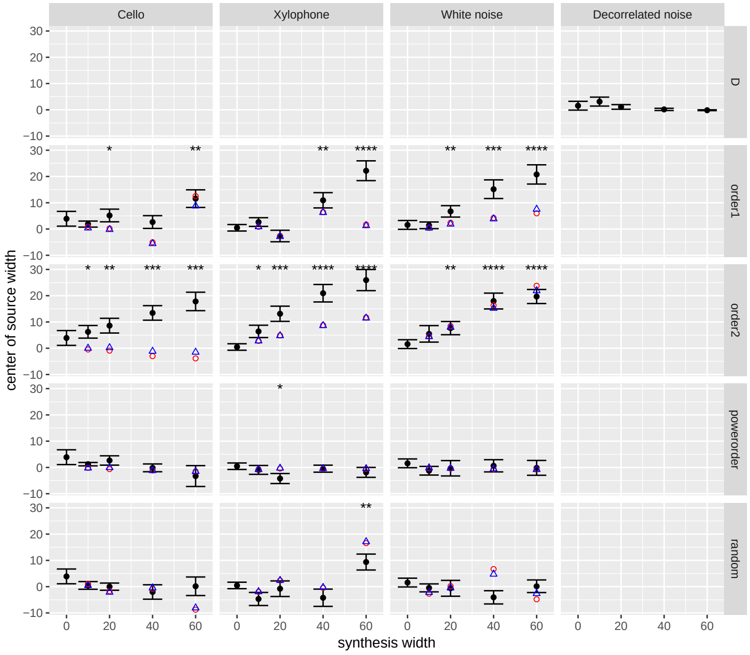


Fig. 3.9 Mean center computed from experiment results (black) with error bars representing standard error, and mean center weighted by power (red circle) and A-weighting applied power (blue triangle). The significances of *t*-test analyzing whether the mean = 0 were marked by asterisks.

In-head localization and sound source moving

The frequencies at which participants perceived *in-head localization* and *sound source moving* during listening test are shown in Fig. 3.10. There is no apparent tendency of in-head localization. On the other hand, sound source motion was barely perceived except for cases of the xylophone source. This could be due to the temporal variation of spectral characteristics of the xylophone signal. This suggests that a processing method in which analysis and synthesis adapt to spectral characteristic over time, such as using short-time Fourier transform (STFT), are necessary to avoid this kind of problem.

Naturalness Evaluation

Ratings of naturalness respect to timbre and spatial attributes were analyzed by ANOVA. The results show no significant differences both for ratings of *naturalness of timbre* and *naturalness of spatial attributes* among factors of source types, synthesis widths, and distribution methods.

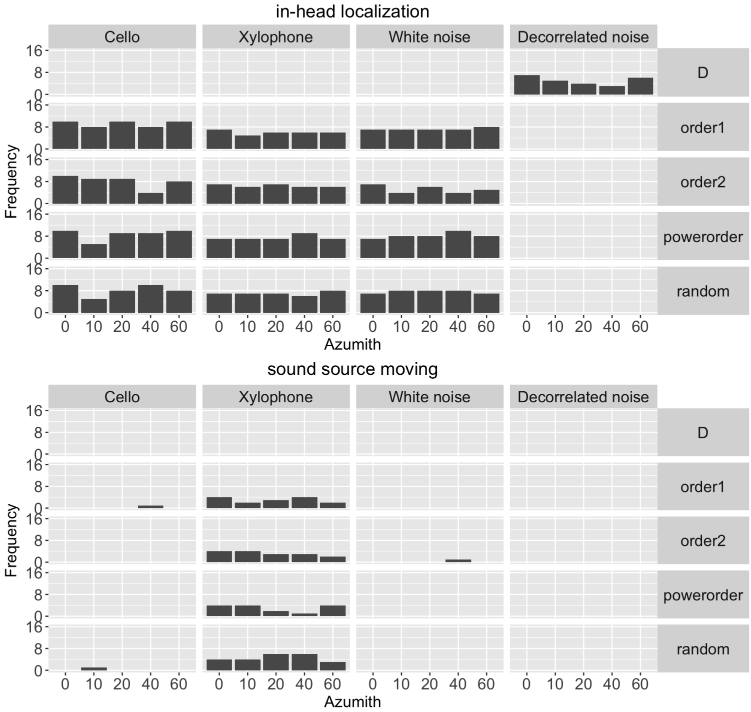


Fig. 3.10 The frequencies of *in-head localization* and *sound source moving*.

3.2.3 Discussion

According to results of perceived source width evaluation obtained from the listening experiment, there was no significant evidence showing that the *perceived width* increased after processed by the widening method proposed in this study. Results inconsistent with those of earlier studies [11, 21] could be due to differences between loudspeaker and headphone reproduction. Reproduction by headphone is known to suffer from the problem of in-head localization. Besides, HRTFs used in this experiment were non-individual, which may lead to localization error and front-back confusion. All of these problems could make it difficult for participants to project the extent of virtual sound source in real space and to describe it in the coordinates accurately, consequently enlarging the error and individual differences of results. However, even evaluated by the same method, results of decorrelation noise still showed clear tendency for source width to increase with synthesis width. When convolving with non-individual HRTFs, the difference between broadband decorrelation noise signals and frequency bands with only 1/3-octave bandwidth could lead to different degrees of degradation on the spatial perception. Still, it is noteworthy that for the decorrelation noise of 60° synthesis width, the average *perceived width* was over 90°, which suggests that problems of in-head localization and non-individual HRTFs could also exist.

On the other hand, stimuli of 0° of all source types, which were only convolved with the 0° HRTF, were perceived averagely as over 24°. Non-individual HRTFs may cause localizations not sharply defined, and in-head localization also could lead to inconsistent response when projecting the sound image into real space. In addition, in earlier studies [11, 21] which used a similar processing method but replayed stimuli via loudspeakers, sound images tended to fuse together so perceived widths were only less than half the real widths of the speaker ensemble. As a result, synthesis widths used in the listening experiment, which was only up to 60°, may not be large enough to differentiate from 0° stimuli.

Shifts of centers of source widths indicate that higher frequency bands may have dominant influence on localization, which could be due to the spectral characteristics of source signals, or because weightings (or importance) of each band to the perception of localizations may depend on frequency. In addition, since adjacent bands were also distributed adjacently for *order1* and *order2* method, this may also have influence on localizations of source widths. The problem of shift is minor in the *powerorder* and *random* methods, which suggests that methods to distribute frequency bands would have a significant effect on width perception.

Ratings of *naturalness* did not show significant differences under different conditions. According to the feedback of participants, the evaluations were difficult when there was no comparison or reference. Consequently, average ratings of *naturalness* were almost around the middle of the 7-point scale.

3.2.4 Conclusion

In this experiment, the effectiveness of the proposed method aiming to control the source width in binaural synthesis was examined. The method was derived from the idea proposed by earlier studies [11, 21] to distribute frequency components to different directions, which produces directional cues varying with frequency. The influence of distribution methods and source types on *perceived width* were investigated by conducting a subjective listening experiment.

According to the results, there was no significant differences in *perceived width* after processed by the widening method. Due to the evaluation method used in this experiment, in-head localization could enlarge the error which caused the effect insignificant. Still, the results showed that the distribution of frequency bands is an important factor in width perception. It also suggests that distribution should take spectral characteristics into account to avoid shift in localization.

3.3 Experiment 2: Frequency bands distribution for virtual source widening in binaural synthesis

In experiment 1, the developed widening method was examined by a subjective experiment. In the listening experiment, participants were asked to indicate where they perceived extents of sound images by using a bar on GUI to select the range of azimuths, *i.e.*, a direct rating task was performed. However, according to the results, there was no significant difference in *perceived width* among different synthesis widths. It was concluded that in-head localization owing to the non-individual HRTFs could make the task difficult to project the extent of virtual sound source to real space and to describe it in coordinates accurately.

Instead of the direct ratings of *perceived width*, an indirect ratings by comparison between stimuli can be more sensitive to differences. Also, the error produced when projecting the virtual source in headphone reproduction into real space can be avoided. Therefore, in this experiment, we performed a pairwise comparison in the listening experiment to investigate the performance of source widening in the binaural synthesis using the same processing method. The purpose of this experiment was to verify effectiveness of the widening processing in binaural synthesis, by investigating differences of perceived widths of sound sources and the naturalness between stimuli with different synthesis width and stimuli without processing.[3]

3.3.1 Methods

Stimuli

The method to generate stimuli was the same as the one used in the previous experiment, as illustrated by Fig. 3.1. Three types of source signals were used: anechoic cello recording, anechoic xylophone recording, and Gaussian white noise. Each recording was firstly trimmed to one phrase of about 3 seconds duration. White noise was also generated with 3 seconds length. Since in pairwise comparison, two stimuli are presented one after another and the evaluation is conducted based on the differences between them, the duration of stimuli should not be too long to prevent difficulty in comparison. A 100 ms fade-in and fade-out was then applied to all signals. After that, the three source signals were filtered into 1/3-octave bands by an FFT-based filter bank and convolved with HRTFs to distribute these frequency bands into different directions within intended synthesis widths. Finally, convolved signals were

[3] The result of this experiment was also published in *H. Su, A. Marui, T. Kamekawa, "Frequency Bands Distribution for Virtual Source Widening in Binaural Synthesis," presented at the Audio Engineering Society Convention 143 (2017).*

summed up for reconstruction of signals which were then used as stimuli in the listening experiment.

HRTFs used in this experiment were from a database of a KEMAR dummy head measured and provided by the MIT Media Laboratory [9]. Synthesis widths under investigation were 10°, 20°, 40°, and 60° in azimuth angle on horizontal plane. The centers of widths were all set to 0° azimuth, *i.e.*, directly in front of the listener. Thus, the HRTF set used was from $-30°$ to 30° azimuth in a 5° interval at 0° elevation.

To determine which HRTF each frequency bands were convolved with, four distribution methods were investigated in experiment 1, as shown in Fig. 3.3. In this experiment, only the two distribution methods which did not have the problem of shifts of localization were adopted, including the *random* method, with a random assignment, and the *powerorder* method, which assigned each frequency band according to its average power. In experiment 1, the *powerorder* method was found to be effective to avoid the shift of localization of sound source. For this method, the band with a higher power is presumed to have more influence on perception, thus sub-bands were distributed with a descending order of power to HRTFs in order of proximity to the center of source width. When all HRTFs of directions within the synthesis width were used, the distribution started from 0° HRTF again until all bands were assigned.

To investigate the effectiveness of widening process, stimuli without frequency band division and distribution were also included in the listening experiment as references. The source signals were simply convolved with the HRTF of 0° azimuth and 0° elevation, creating stimuli localized in the 0° azimuth, as same as other stimuli, but with a 0° synthesis width.

Procedure

The evaluation method was based on Scheffe's pairwise comparison [29]. A pair of stimuli, *A* and *B*, was presented in order. Participants were asked to compare the two stimuli and to judge which one had wider source width and evaluate the difference on a 7-point scale. In addition, for stimuli from anechoic recordings, differences in the naturalness of timbre and the naturalness of spatial impression were also evaluated to examine if there were any degradations of sound quality after the widening processing. The statements of 7-point scales for three evaluation items are listed in Table 4.1. These statements were provided on the GUI and on the instruction (with Japanese translation) of the listening experiment, but the verbal instruction was in Japanese. The statements in Japanese are provided in Appendix 2.

The experiment was divided into 3 sections for 3 types of source signals. In each section, the stimuli from the same source signals but with different synthesis widths or/and different distribution methods were compared. There were 4 synthesis widths with 2 distribution

Table 3.1 Statements of 7-point scales for the width evaluation and the naturalness evaluations.

items	statements
Sound source width	*A* is highly wider
	A is moderately wider
	A is slightly wider
	A and *B* are equal
	B is slightly wider
	B is moderately wider
	B is highly wider
Naturalness of spatial impression/ Naturalness of Timbre	*A* is highly more natural
	A is moderately more natural
	A is slightly more natural
	A and *B* are equal
	B is slightly more natural
	B is moderately more natural
	B is highly more natural

methods and a stimulus without widening processing, resulting in 9 stimuli for total of 72 pairwise comparisons for each section. Participants were asked to adjust the volume to a comfortable level before starting each section. The stimuli pairs were presented in random orders. The presentation of each stimuli pair could be replayed as many times as participants wanted.

The listening experiment was conducted in a quiet classroom in Senju campus of Tokyo University of the Arts. Stimuli were reproduced via a laptop (MacBook Pro) and a headphone set (SONY MDR-CD900ST). Participants were instructed to use graphical user interface (GUI) to answer the questions, which is shown in Fig. 3.11. The average duration of the experiment was about 1 hour.

Participants

A total 12 participants from Tokyo University of the Arts took part in the experiment. All were students major in the Research Area of Creativity of Music and Sound and had previous experience in listening experiments

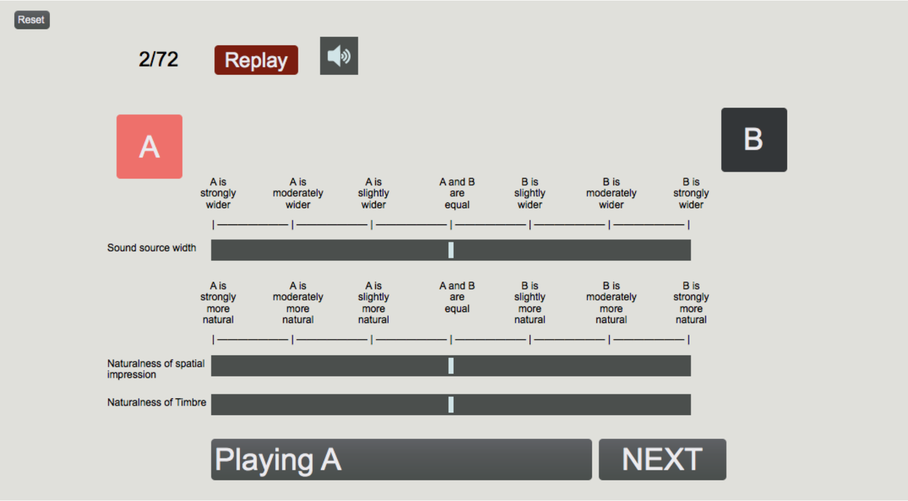


Fig. 3.11 GUI constructed in Cycling'74 Max for pairwise comparison in the listening experiment.

3.3.2 Results

Results of Comparison for Perceived Width

The average relative ratings of perceived widths were obtained by analysis of the data from the subjective listening experiment based on the model of Scheffé's Pairwise comparison [29], *i.e.*, computing the main effects by the method described in Section 2.3.1. Fig. 3.12 shows the relative perceived source width of nine stimuli as a function of synthesis width in degrees for each type of source signal. A higher index of relative width means a wider perceived source width. Analysis of variance was performed based on Scheffé's model. There were significant main effects (p<0.001) for all source types (see also Tables 3.2, 3.4, and 3.6), indicating that there were significant differences of perceived widths between stimuli. Multiple comparisons were conducted by computing the confidence interval of differences between all pairs. For cello and white noise sources, the perceived widths of stimuli of 40°, 60° synthesis width in the *random* distribution and 60° synthesis width in the *powerorder* distribution were significantly wider than the unprocessed 0° stimulus at the 0.01 significant level (Table 3.3, 3.7). For the xylophone source, only the 60°-*random* stimuli were perceived significantly wider than the 0° stimulus (Table 3.5). The results indicated that after processing by the widening method proposed in this study, the perceived source width of a monophonic sound could be widened in binaural synthesis. However, the effectiveness was significant only when the synthesis width was large enough, and could vary depending on source signals.

The effect of distribution method was investigated by examining the differences between pairs with the same synthesis width but different distribution methods. For 60° synthesis

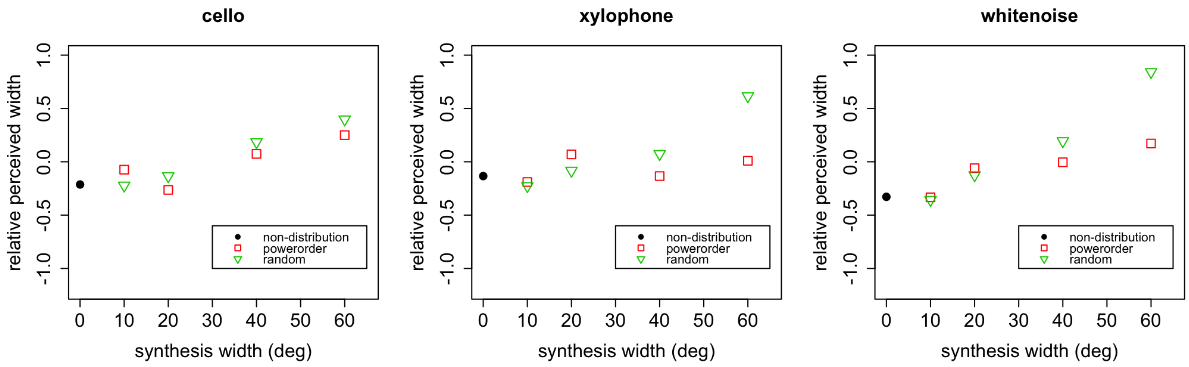


Fig. 3.12 Average relative ratings of perceived widths as a function of synthesis width for three types of source signals. Symbols denotes different distribution method as described in the legend.

Table 3.2 ANOVA table of width ratings for cello source

ANOVA Table					
Source	SS	df	MS	F	p
Main	96.9167	8	12.1146	11.9536	< 0.0001
Main × Indiv	470.4167	88	5.3456	5.2746	< 0.0001
Combi	32.9167	28	1.1756	1.1600	0.2609
Order	10.6667	1	10.6667	10.5249	0.0012
Order × Indiv	49.2778	11	4.4798	4.4203	< 0.0001
Error	737.8056	728	1.0135		
Total	1398.0000	864			
	Y(0.05)=0.3014, Y(0.01)=0.3493				

width, the stimulus of *random* distribution was significantly wider than the stimulus of *powerorder* distribution for xylophone and white noise source. Since in *powerorder* method, the frequency bands with higher power were distributed in order of proximity to the center, most of energy of the signal could be distributed into a narrower range compared to the random method, which may produce a narrower perceived source width. To investigate this assumption, the distribution results of all stimuli, including the power of each band and which direction it was distributed to, were examined. However, there was no clear evidence indicating that bands with higher power were distributed to a wider range for *random* distribution.

Since IACC is considered to be related to perceived source width, as described in Section 2.2.1, the IACC of each stimulus was computed. The relation of relative source width with $1 - \text{IACC}$ is displayed in Fig. 3.13, together with Pearson's correlation coefficients for each source respectively. Correlations from 0.78 to 0.88 indicated fairly high correlations

Table 3.3 Stimuli pairs with significant differences for cello source. * represents a 0.05 and ** represents a 0.01 significance level.

stimuli pair	significance
0° non-distribution - 40° random	**
0° non-distribution - 60° powerorder	**
0° non-distribution - 60° random	**
10° powerorder - 60° powerorder	*
10° powerorder - 60° random	**
10° random - 40° random	**
10° random - 60° powerorder	**
10° random - 60° random	**
20° powerorder - 40° powerorder	*
20° powerorder - 40° random	**
20° powerorder - 60° powerorder	**
20° powerorder - 60° random	**
20° random - 40° random	*
20° random - 60° powerorder	**
20° random - 60° random	**
40° powerorder - 60° random	*

Table 3.4 ANOVA table of width ratings for xylophone source

ANOVA Table					
Source	SS	df	MS	F	p
Main	112.3241	8	14.0405	14.6180	< 0.0001
Main x Indiv	209.5648	88	2.3814	2.4794	< 0.0001
Combi	33.3009	28	1.1893	1.2382	0.1858
Order	95.3345	1	95.3345	99.2555	< 0.0001
Order x Indiv	101.2350	11	9.2032	9.5817	< 0.0001
Error	699.2407	728	0.9605		
Total	1251.0000	864			
	Y(0.05)=0.2934, Y(0.01)=0.3401				

Table 3.5 Stimuli pairs with significant differences for xylophone source. * represents a 0.05 and ** represents a 0.01 significance level.

stimuli pair	significance
0° non-distribution - 60° random	**
10° powerorder - 60° random	**
10° random - 40° random	*
10° random - 60° random	**
20° powerorder - 10° random	*
20° powerorder - 60° random	**
20° random - 60° random	**
40° powerorder - 60° random	**
40° random - 60° random	**
60° powerorder - 60° random	**

Table 3.6 ANOVA table of width ratings for white noise source

ANOVA Table					
Source	SS	df	MS	F	p
Main	246.8056	8	30.8507	25.8186	< 0.0001
Main x Indiv	912.8611	88	10.3734	8.6814	< 0.0001
Combi	27.8611	28	0.9950	0.8327	0.7149
Order	47.2269	1	47.2269	39.5236	< 0.0001
Order x Indiv	25.3565	11	2.3051	1.9291	0.0329
Error	869.8889	728	1.1949		
Total	2130.0000	864			
	Y(0.05)=0.3272, Y(0.01)=0.3793				

Table 3.7 Stimuli pairs with significant differences for white noise source. * represents a 0.05 and ** represents a 0.01 significance level.

stimuli pair	significance
0° non-distribution - 40° random	**
0° non-distribution - 60° powerorder	**
0° non-distribution - 60° random	**
10° powerorder - 40° powerorder	*
10° powerorder - 40° random	**
10° powerorder - 60° powerorder	*
10° powerorder - 60° random	**
10° random - 40° powerorder	**
10° random - 60° powerorder	**
10° random - 40° random	**
10° random - 60° random	**
20° powerorder - 60° random	**
20° random - 60° random	**
40° powerorder - 60° random	**
40° random - 60° random	**
60° random - 60° powerorder	**

between source width and IACC, in accordance with previous findings reported by other studies [16, 4, 39].

Results of Comparison for Naturalness

The evaluations of naturalness were analyzed by the same method as the analysis of perceived widths. The relative ratings of naturalness are given in Fig. 3.14 as a function of synthesis widths. As can be seen in the figure, the differences between stimuli are smaller compared to the results of the evaluation of source width, indicating that on average participants did not perceive much difference between stimuli pairs so they tended to use a smaller range of scales to rate the differences. As a result, there is no clear relation with the naturalness and synthesis width. Results of analysis of variance indicated that there was no significant difference of naturalness of spatial attributes between stimuli for both sources, but there are significant differences of naturalness of timbre between stimuli for cello (p<0.01) and xylophone (p=0.02). However, no clear relation between naturalness of timbre and synthesis width or distribution method could be found after examining stimuli pairs with significant difference.

The results of multiple comparisons indicated that for cello source the 60°-*powerorder* stimulus was significantly more natural in timbre than 20°-*powerorder*, 20°-*random*, 40°-

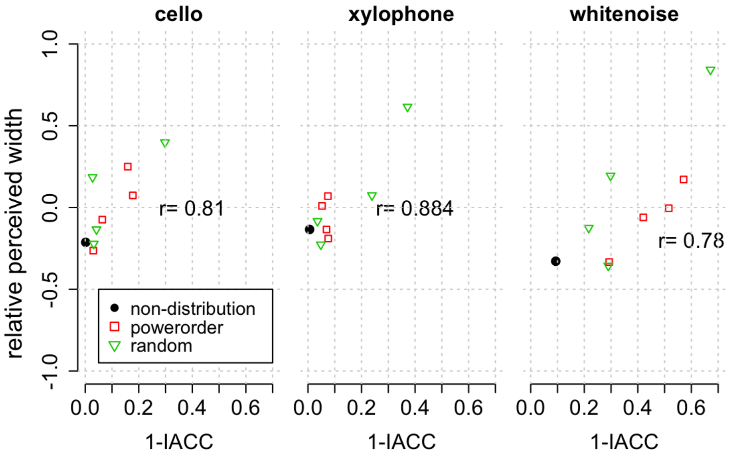


Fig. 3.13 Relative perceived width as a function of 1 − IACC and Pearson's correlation coefficients between them for each source signal

random, 60°-*random* stimuli, and for xylophone 20°-*random* was significantly more natural in timbre than 40°-*random*.

Cluster Analysis

Individual differences between participants were found when analyzing and examining responses of each participant individually. The results of ANOVA showed agreement as *p*-values of interactions of main effects and individual effects were lower than the 0.05 significance level. Therefore, hierarchical cluster analysis was performed to group participants with similar response patterns. With respect to each evaluation item, distances between participants were computed using responses of all source signals to obtain the distance matrix. According to the dendrogram generated based on the distance matrix (see Fig. 3.15), the participants were divided into two groups for the *width* evaluation, three groups for *naturalness of spatial impression* evaluation, and three groups for *naturalness of timbre* evaluation.

After clustering, the analysis of pairwise comparison was performed again with respect to each group. The results are shown in Figs. 3.16–3.18 for each evaluation item respectively, which present relative ratings as a function of synthesis width for each group and each source. The number in parentheses in the titles represents the number of participants in each group.

For the width evaluation, there was a clear difference between two groups in relations of perceived width with synthesis width. As seen in Fig. 3.16, participants in Group 1 rated the source width wider for stimuli with wider synthesis widths; however, in Group 2 participants rated the source width in an opposite way. Considering the approach and the purpose of

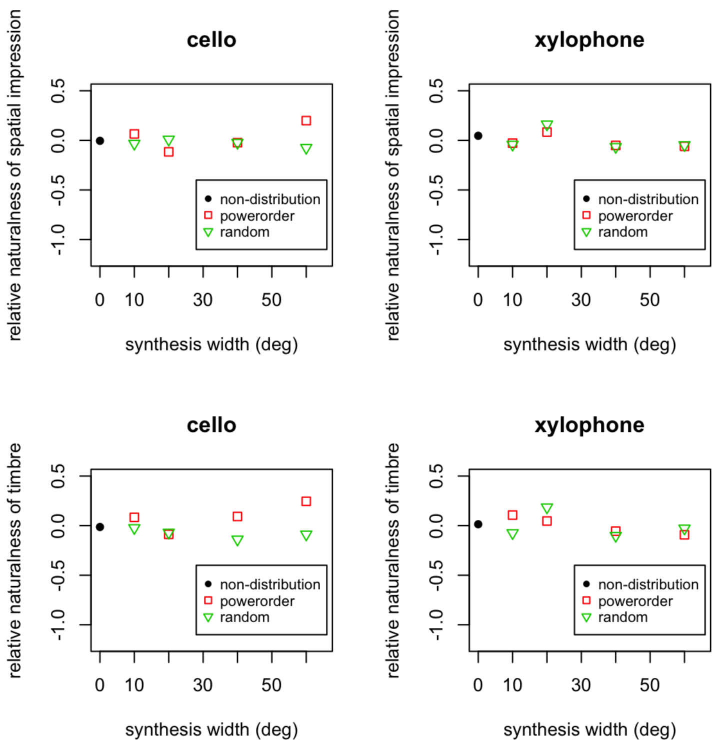


Fig. 3.14 Relative ratings of naturalness from subjective listening experiments. The upper panels are the relative naturalness of spatial attributes for cello and xylophone sources, and the lower panels represent naturalness of timbre.

the processing method, this result was surprising, so is further discussed in the Discussion section.

For the naturalness of spatial impression, there are three patterns seen in Fig. 3.17. In Group 1, the naturalness seems to increase with synthesis width slightly, but significant differences could only be found between the 60°-random stimulus and stimuli with 0° or 10° synthesis width for cello source. On the other hand, naturalness ratings in Group 2 decrease as synthesis width increases. In Group 3, which includes only one participant, the tendency seems to depend on source signals. As synthesis width increases, naturalness was rated higher for cello source but lower for xylophone source.

The differences in the tendency of ratings between groups also can be observed in the results of the naturalness of timbre (Fig. 3.18). In Group 1, there seems no relation between naturalness and synthesis width, whereas naturalness tends to decreases with synthesis width in Group 2. In Group 3, opposite tendency can be observed for different source types. After comparing the results of hierarchical cluster analyses of two naturalness evaluations, it is worth noting that the participants in Group 1 and Group 2 are somehow different, with few

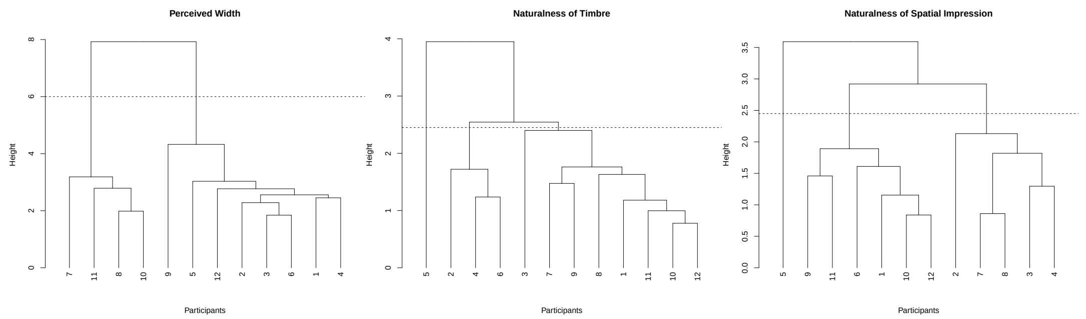


Fig. 3.15 Dendrogram of hierarchical cluster analysis. Height represents intergroup dissimilarity between two groups, the number of each individual observation represent each participant. Dotted lines show how groups were divided.

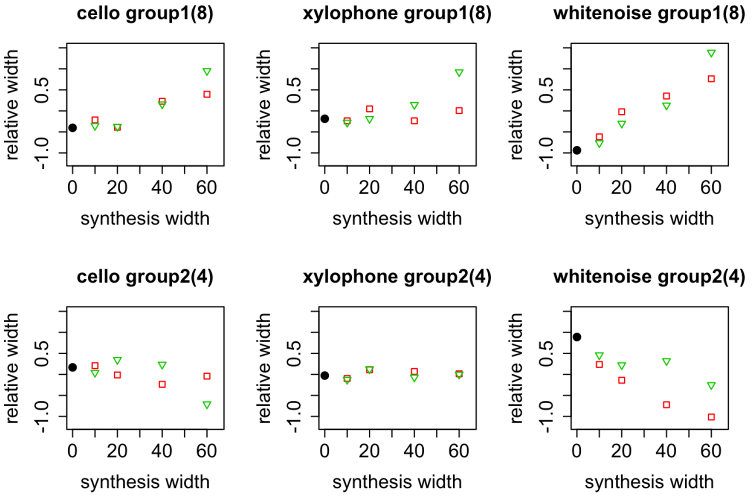


Fig. 3.16 Average relative perceived width for each group.

participants exchanged between the groups for different evaluations of naturalness, and only Group 3 includes the same participant (#5).

3.3.3 Discussion

Perceived source width

The result of the width evaluation indicated that the perceived source width could widen after the frequency bands division and distribution. However, significant differences could only be found when synthesis width was larger than 40°. The non-individual HRTFs used in this experiment can explain why minor difference in perceived width for stimuli with synthesis width narrower than 40° could not be perceived significantly. In Experiment 1, described

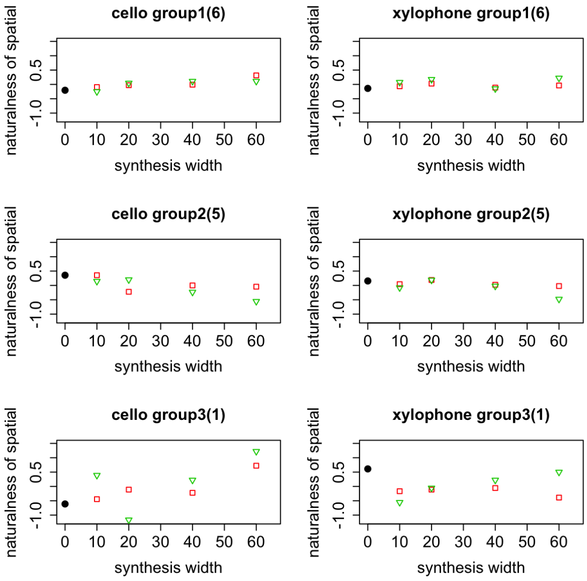


Fig. 3.17 Average relative naturalness of spatial impression for each group

in Section 3.2, we found that the source widths of stimuli only processed by convolution with HRTF of 0° azimuth, *i.e.* the stimuli with 0° synthesis width, were perceived with average over 24°. It was concluded as a result of non-individual HRTFs which could cause localizations to be not sharply defined. In earlier studies [11, 21] using a similar approach but replaying stimuli via a loudspeaker array, sound images tended to fuse together, so perceived widths were less than half of real widths of loudspeaker ensemble. Therefore, it may be reasonable to suppose that synthesis widths lower than 40° could be only perceived with a width lower than 20°, which was too narrow to be differentiated from 0° stimuli.

The two distribution methods adopted in this experiment were methods that did not have significant problems of shift of localization in Experiment 1. However, it would seem that random distribution sometimes “randomly” distributed bands unevenly, so the localization of the stimuli could shift to the side where bands with higher power concentrated. There was a similar finding in this experiment. According to feedback from participants after the listening experiment, some stimuli were perceived shifted from the center. After examining the stimuli and the distribution results, a stimulus of *random* distribution from xylophone source indeed had this problem. Nevertheless, it appears that *random* distribution may be more effective than *powerorder* for source widening according to the result of width ratings. Thus, refining the distribution method that can strike a balance between random and deterministic distribution will be future work.

The effectiveness was found to depend on source signals. It is reasonable to suppose that spectral characteristics would affect width perception since the approach in this study

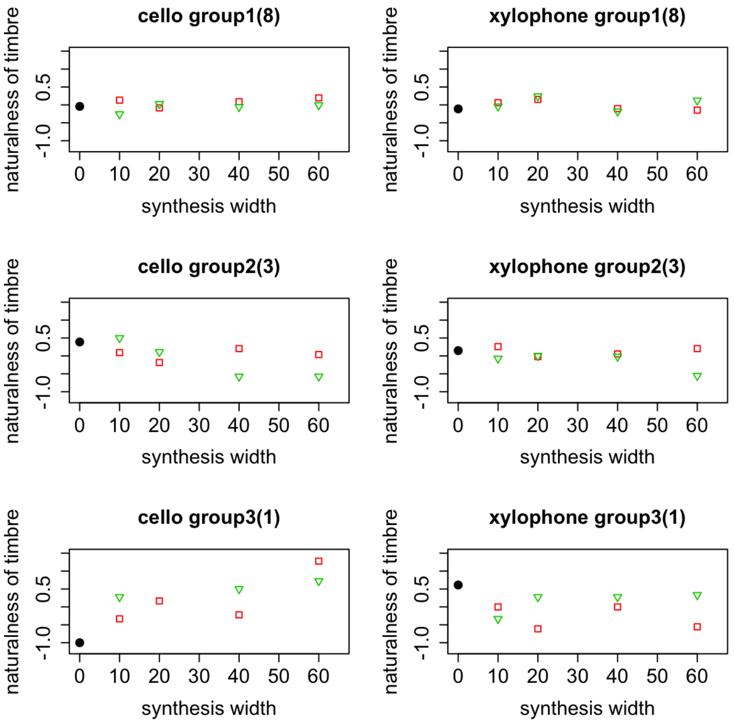


Fig. 3.18 Average relative naturalness of timbre for each group

involved frequency bands division and distribution. According to the results, it seems harder to achieve a wider extent for xylophone source, while for white noise sources the tendency for perceived width to increase with synthesis width was obvious. After the experiment, most of the participants also reported that the comparison task was harder for the section of the xylophone source. This was consistent with the result in Experiment 1 which showed significant differences in perceived widths among three source types. It can be assumed that spectral characteristics would lead to this difference. Due to the nature of this processing approach, energy may be distributed more evenly for broadband signals, so a better perceptual quality of source width could be obtained. This conclusion suggests that dividing the frequency bands more finely may achieve more stable effectiveness in source widening, since the spectral components can be distributed more evenly.

Individual differences

Individual differences were found in results of all evaluation items. For ratings of perceived width, the inverse relationship with synthesis width observed in some participants was unexpected and contradicted the aim of this study. To investigate the reason, an interview was conducted with participants whose responses showed the inverse relation between perceived width and synthesis width. They reported that when they compared two stimuli, rather than the difference of source width, the difference of timbre, pitch, or frequency characteristic was more obvious. It can be assumed that for these participants, their own individual HRTF were relatively dissimilar to the non-individual HRTF used in this experiment. As a result, instead of changes of width, which they could hardly perceive, they used other attributes of sounds to link to ratings of the source width.

For example, one participant reported that differences among stimuli were perceived as filtered or equalized in different frequency bands. Stimuli rich in high frequency content produced more spaciousness, especially for vertical direction. In this case, the source width was rated wider. However, due to the frequency responses of HRTFs and the distribution results of frequency bands, stimuli with more high frequency components were actually stimuli with narrower synthesis width in this experiment.

The other participant reported that for some stimuli the sound images were not localized in the center front but in the left or right side, or the images were even split into two parts. If HRTF of this participant matched the non-individual HRTF in only some frequency bands, it could be assumed that only those bands could be localized clearly. Localization of other bands may be ambiguous, so some parts of the sound image composed by these bands might be "missing." In this case, the source width was rated narrower than 0° stimuli, since the localizations without widening processing were more clearly defined.

Therefore, Group 1 and Group 2 of the cluster analysis showing different ratings tendencies could be interpreted as whether the HRTF of the participant matched the non-individual HRTF in an acceptable way or not, although this assumption needs further investigation. This result also suggests that using individual HRTFs may improve the performance of this approach.

Individual differences could also be found in evaluations of naturalness. It suggests that participants had different criterion for judging the naturalness. However, although different patterns could be found in different groups after clustering, differences were not significant. Thus, it could be concluded that there was no significant degradation of naturalness after widening processing.

3.3.4 Conclusion

In this experiment, the performance of a source widening processing method proposed in this study was further investigated by an indirect ratings method. A listening experiment using pairwise comparison was conducted to investigate the difference of perceived width and naturalness after widening processing. The results showed that for synthesis widths above 40°, perceived widths were significantly wider than stimuli without widening processing. In addition, IACC decreased after widening processing, and fairly high correlations between 1 – IACC and perceived width were found. No significant degradation of naturalness after processing was found. However, the performance seems to depend on the source signals. In order to improve the processing method to be applicable to various source type, revising the processing method by optimizing other parameters to divide frequency components more finely and to distribute sub-bands more uniformly is needed.

Individual differences were found, and the reverse relation between perceived width and synthesis width observed in results of some participants was assumed to be associated with non-individual HRTFs. This suggested that there would be a limitation for this method when using non-individual HRTFs. Individualization of HRTFs may improve the performance of this approach.

3.4 Experiment 3: The Source Widening Effect in Binaural Synthesis with Spatial Distribution of Frequency Bands

In Experiment 2, a pairwise comparison, which could be more sensitive to differences, was performed in the listening experiment to investigate the performance of source widening in the binaural synthesis using the same processing method and the same source signals. The results showed that for synthesis widths above 40° perceived widths were significantly wider than stimuli without widening processing. However, a dependency of the performance on characteristics of source signals was observed. For the xylophone source, perceived widths were significantly wider only for 60° synthesis width. It was reasonable to suppose that due to the spectral characteristic of xylophone, with most energy in narrow high frequency region, the distribution of energy could be less uniform, so the source width could be perceived narrower. It therefore suggested that the bandwidth, which was 1/3-octave in Experiment 1 and Experiment 2, of the frequency components may have influence on perception of source width, since it can be assumed that more finely the frequency components were divided the more uniform spatial distribution of frequency components could be obtained. In addition, high frequency characteristics and short duration such as percussive timbre of xylophone were already found to be associated with narrower width perceptions [12, 20], which could make it more difficult to achieve a widened perceived width for source signal such as the xylophone recording.

Furthermore, individual differences were found in the results of evaluation. For some participants, inverse tendencies of the perceived width evaluation were found. Especially for stimuli of white noise source signals, they rated the perceived width narrower with the increase of synthesis width. It was assumed that for these participants, their individual HRTFs could differ more from the the non-individual HRTF of a dummy-head used in Experiment 2. Thus, instead of changes of perceived width which they could have difficulty to discriminate due to the mismatch of HRTFs, they could only discriminate stimuli by other attributes such as timbre or spectral characteristics. It suggested that individualization of HRTF should be done to some degree to avoid non-individual HRTF mismatching in an unacceptable way.

In Experiment 3, the aims are to further investigate other parameters of the widening effect, and to investigate the relation between synthesis width and perceived width, *i.e.* to verify whether the widening effect can control perceived source width, while excluding factors that may cause problem in width perception such as non-individual HRTFs. The influence of bandwidth of divided frequency bands was examined, with an assumption that dividing source signals more finely could produce more homogeneous distribution. The HRTF individualization was performed by subjective selection of HRTF to reduce the potential problems of non-individual HRTF. Furthermore, the widening effect should be applicable to any directions of sound sources. Hence, two center positions of source width: 0° and 15° azimuth, were investigated. Pink noise and a cello anechoic recording were chosen as monophonic source signals in this experiment. Since pink noise has the attribute that the energy is equal for each octave band, and bandwidths used for frequency division in this study were also 1/n-octave, it was assumed to be appropriate as a representation for synthesis sounds and for sounds with broadband characteristics. The xylophone recording, which we used in the previous studies, could suffer essential problems on narrow extent perception as described previously. In addition, the previous study [21] also suggested that for signals with mainly impulsive content other algorithms are advised. Therefore, only the cello recording was used as a representation for recording materials.

The subjective listening experiment was conducted using a multi-stimulus test. In this method, multiple stimuli were presented simultaneously, and thus direct comparisons between stimuli could be done in aspects of *perceived width*, *degradation in spatial quality*, and *degradation in timbre quality*. Besides the stimuli processed by widening effect, the stimuli without widening processing were also evaluated as an unprocessed reference, and stimuli generated by a decorrelation filter proposed by previous studies were evaluated to compare with another widening processing approach.[4]

[4] The result of this experiment is submitted for publication in *H. Su, A. Marui, T. Kamekawa, "The Source Widening Effect in Binaural Synthesis with Spatial Distribution of Frequency Bands," Engineering Report, the Audio Engineering Society.*

3.4.1 Methods

Subjective Selection of HRTF

Before generating stimuli for the listening experiment, subjective selection of non-individual HRTFs was performed individually to decide which HRTF to be used for the binaural widening processing for each participant. Instead of the MIT KEMAR database used in previous experiments, in which only one set of HRTFs from KEMAR dummy head was provided, other databases with multiple sets of HRTFs from different subjects were used to provide multiple candidate HRTFs which were necessary for subjective selection. The candidate HRTFs were from two databases, CIPIC and RIEC database [2, 37]. For each database, k-means cluster analysis was performed to divide HRTFs from different subjects to k clusters. The original HRIR dataset from the database was transformed to HRTF in the frequency domain by FFT, and the amplitude spectrum was converted to decibel scale, which is related more to the auditory perception. Since only the HRTFs of 0° elevation from -45° to 45° azimuth would be used to generate stimuli, amplitude spectrum data of these directions of both ears were concatenated as one dataset for each subject, which was then used for the cluster analysis. $k = 4$ was adopted considering the reproducibility of clustering and the adequate number of candidates for subjective selection. The most representative HRTF of each cluster was selected according to its distance to the center of cluster, as the previous study [33] described in the previous section. Thus, four representatives from two database resulted eight candidate HRTFs. Test signals for the subjective selection were generated using these eight HRTF datasets respectively.

Test signals for the subjective selection of HRTFs were white noise pulses with 30 ms duration and 5 ms fade-in/out, convolved with -45° to 45° azimuth and 0° elevation HRTFs. The pulses moved from -45° to 45° in a 5° interval subsequently with 100 ms pause between each pulse, then returned to -45° repeatedly but in inverse direction, resulting in a sound pulse moving from the left to the right and then returning to the left.

Subjective selection of HRTF was performed before the main experiment. Participants were asked to evaluate the spatial quality of eight test signals generated with eight different HRTFs and score them in a 0–100 scale based on the following criteria:

- If the sound is perceived as coming from the front side, the stimulus should be rated high. On the other hand, if the sound perceived as inside the head or behind the head, it should be rated low.

- The stimulus should be rated high if localization of each pulse could be clearly perceived. If necessary, the loudspeakers placed on the -45°, 0°, 45° azimuth in the experiment room can be used as a visual aid for evaluating the accuracy of sound localization.

- The stimulus should be rated high if the movement of pulses is uniform, *i.e.*, sound pulses are perceived as with equal intervals and constant elevation.

These three criteria were chosen to ensure that the selected HRTF could produce clear sound imagery in the frontal direction, considering the method and purpose of this study. Participants used a GUI to replay and switch between any of the eight test signals, which provided a full paired comparison between HRTF candidates. The test signals were replayed via a AKG K240 MKII headphone set, the same as that used in the main experiment. The task was conducted in the Studio B in Senju Campus of Tokyo University of the Arts. It took about 10 minutes for participants to complete the selection task. After the evaluation, the HRTF set used for the test signal rated with the highest score was selected as the individualized HRTF. The stimuli for the main experiment were then generated by that HRTF individually.

Source Signals

Anechoic cello recording and pink noise were used as source signals. They were monophonic audio at 44.1 kHz sampling rate and 16 bits per sample. Since the RIEC HRTF database is provided at 48 kHz sampling rate, the 48 kHz versions of source signals were also generated by upsampling for stimuli using HRTF from RIEC database. For stimuli using HRTF from CIPIC database, the original 44.1 kHz version of source signals were used since the database is provided in 44.1 kHz. Before widening processing, source signals were trimmed or generated with 10-second duration and padded with a 100-millisecond fade-in and fade-out included in the 10-second duration.

Widening Processed Stimuli and the Processing Method

The algorithm of widening processing method was basically the same as the method used in our previous studies, as illustrated in Fig. 3.19. First, source signals were divided into multiple frequency bands by an FFT-based filter bank. Each band was then convolved with HRTF to distribute these frequency bands into different directions which were in the range within intended localizations and widths. For example, to synthesize a stimulus with 30° width with center located in 15° azimuth, the HRTFs of 0°, 5°, 10°, . . . , 30° were used. Finally, convolved signals were summed up for reconstruction of signals, which were then used as stimuli in the listening experiment.

Since the bandwidth of the frequency bands was considered a factor in how evenly the frequency components of the source signals were distributed, stimuli were generated with

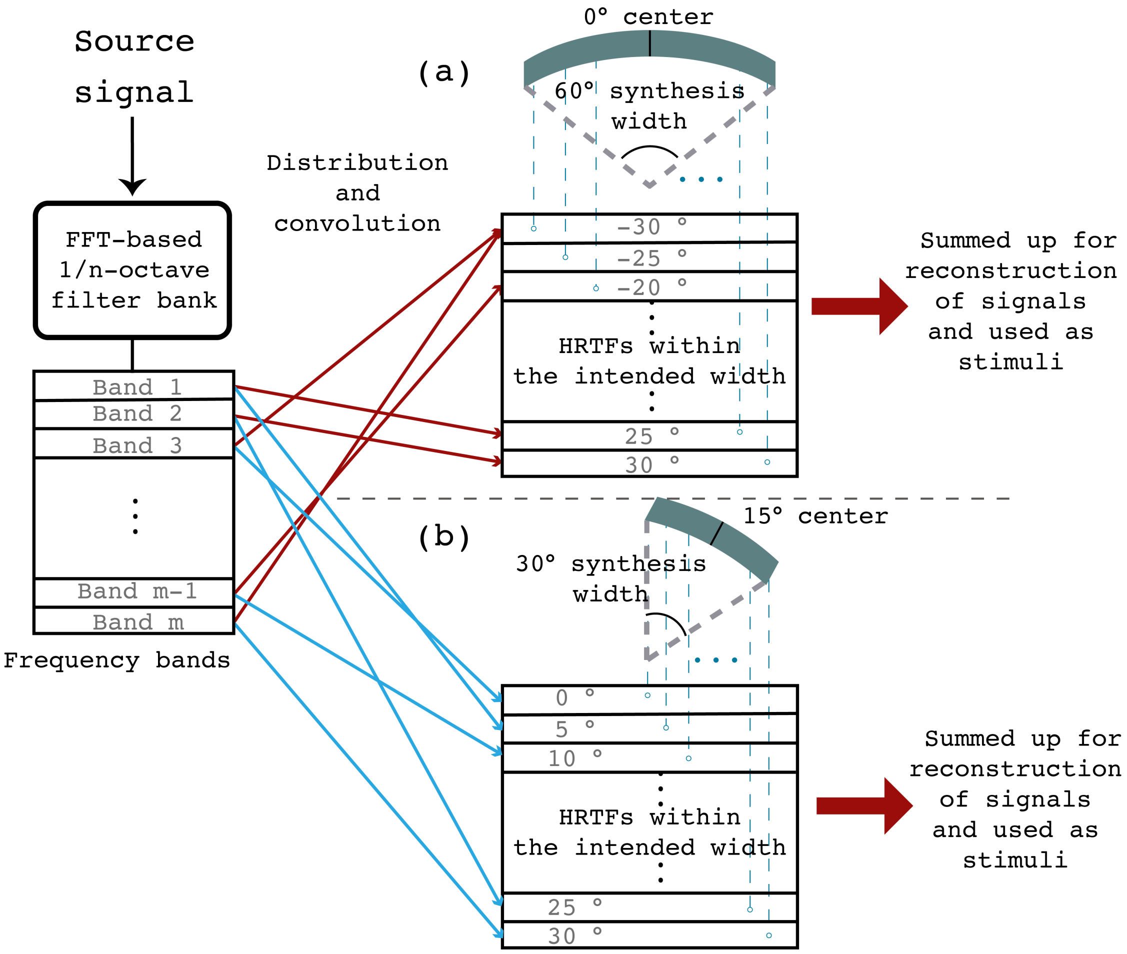

Source signal
FFT-based 1/n-octave filter bank
Band 1
Band 2
Band 3
Band m-1
Band m
Frequency bands
Distribution and convolution
(a)
0° center
60° synthesis width
-30 °
-25 °
-20 °
HRTFs within the intended width
25 °
30 °
Summed up for reconstruction of signals and used as stimuli
(b)
15° center
30° synthesis width
0 °
5 °
10 °
HRTFs within the intended width
25 °
30 °
Summed up for reconstruction of signals and used as stimuli

Fig. 3.19 The widening processing method. (a) Stimuli with 60° synthesis width with center at 0° azimuth. (b) Stimuli with 30° synthesis width with center at 15° azimuth.

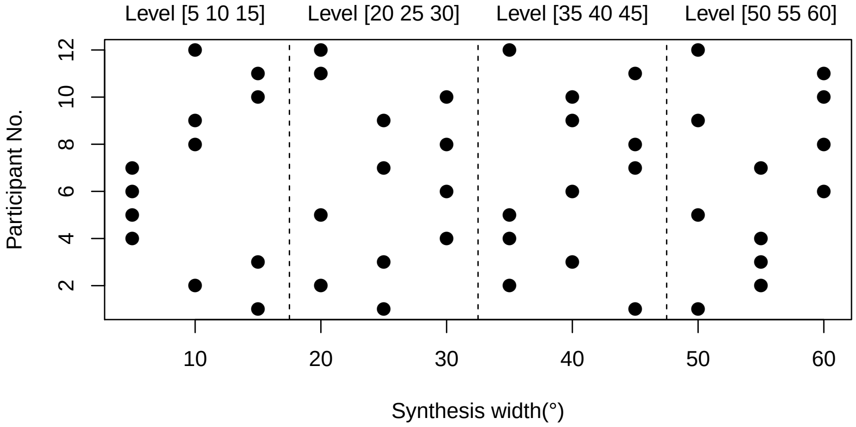


Fig. 3.20 Synthesis widths of each participants tested. Each dot indicates the synthesis width on horizontal axis was tested by the participant on vertical axis.

1/3, 1/6, and 1/12-octave filter banks respectively to investigate the influence of fineness of frequency components division on widening performance and sound quality.

To investigate the relationship between intended synthesis widths and perceived widths, 12 synthesis widths from 5° to 60° in a 5° interval were tested. Considering the experiment time, each participant evaluated only four synthesis widths. The four synthesis widths were chosen from four levels of synthesis widths respectively to produce stimuli with sufficient differences which could be evaluated in the listening experiment. 16 synthesis widths under investigation were divided into four levels, which were [5°, 10°, 15°], [20°, 25°, 30°], [35°, 40°, 45°], and [50°, 55°, 60°]. In order to reduce biases, from each level one synthesis width was randomly chosen, resulting four synthesis widths for each participant, but with a restriction that all widths were tested for the same number of participants, as shown in Fig. 3.20, where the synthesis widths each participants tested were given. All widths were on the horizontal plane (i.e, 0° elevation) in the frontal direction. The centers of synthesis width under investigation were 0° and 15° azimuths.

To distribute frequency components of source signals to different directions in the range of the intended width, the synthesis width was divided equidistantly by the number of frequency bands of source signals after processing by the filter bank. For example, if a source signal was processed by the 1/3-octave filter bank, there were 28 bands from 31.5 Hz to 16 kHz. For stimuli of 60° synthesis width with 0° center, the −30° to 30° azimuth would be divided to 28 directions equidistantly. Since HRTFs used in this study were at intervals of 5° azimuth, the equidistant azimuths were rounded to the azimuths of the nearest measurement points. Each band was then assigned randomly to HRTFs of different azimuths. Different randomized patterns were used for every participant to evaluate the overall effect of the random distribution, but for each participant the stimuli from the two source signals were applied with the same randomized pattern.

Reference Stimuli

Besides stimuli described above, stimuli without widening processed (*non-widen*) and stimuli processed by decorrelation filter (*decor*) were also included in the multi-stimulus test. Stimuli simply convolved with the HRTF of the center of width, *i.e.* 0° or 15° azimuth, without frequency band division and distribution were generated. These stimuli were supposed to have the narrowest perceived width, while could be thought of as references for *timbre degradation* and *spatial quality degradation* evaluations to examine the degradation after widening processing. Besides, in order to compare the widening processing with other methods, the phantom source widening effect for a standard stereo loudspeaker setup by a decorrelation filter proposed in previous studies [42, 40] was used. Sinusoidal phase all-pass filter pair as described in equation 3.1 was applied to source signals, and the decorrelated pair signals were convolved with HRTFs of $-30°$, $30°$ azimuth for comparison with stimuli with center on 0° azimuth, or HRTFs of $-15°$, $45°$ azimuth for stimuli with center on 15° azimuth.

$$H_{1,2}(\omega) = \frac{1}{\sqrt{2}} e^{\pm i\hat{\phi}\sin(\omega T)} \tag{3.1}$$

To align playback levels of decorrelated signals to other stimuli, the levels were adjusted to have the same root-mean-square (rms) power with average rms of all other stimuli with the same center from the same source signal.

Experiment

The experiment was conducted in a quiet studio room in Senju Campus of Tokyo University of the Arts. Participants were instructed to use the GUI constructed in Cycling'74 Max to answer the questions and control the playback of stimuli. Stimuli were reproduced via a laptop (MacBook Pro) and a headphone set (AKG K240 MKII). 12 participants from Tokyo University of the Arts took part in the experiment. All were students majoring in Research Area of Creativity of Music and Sound and had prior experience in listening experiments.

Multi-stimulus test was conducted to evaluate the *perceived source width* of stimuli. 12 stimuli with four synthesis widths and three types of bandwidths were randomly divided into two parts as consecutive comparison trials. One comparison contained six stimuli and two reference stimuli (*non-widen* and *decor*), thus total of eight stimuli were presented at a time.

The presentation order of the eight stimuli on the GUI was randomized. In addition, four conditions with combinations of two source types and two center positions were tested in randomized orders, resulting in a total of eight comparisons in one session.

Besides the evaluation of *perceived width*, the evaluation of *degradation in spatial quality* and *degradation in timbre quality* were also investigated. In these two evaluation, stimuli simply convolved with the HRTF of center of synthesis width, *i.e.* the *non-widen* stimuli, were used as references, thus the evaluated degradation could be considered as resulting from the widening process of frequency bands division and distribution. The examples of degradation in spatial quality were given as shift of localization, non-uniformity of sound image distribution, and separation of sound image. For evaluations of degradation, only stimuli from cello recording were used, thus there were four comparisons in one session.

Three evaluation attributes resulted in three sessions of the experiment. The orders of sessions were randomized for every participant. Participants were instructed to score the stimulus with the highest rating 100 points and the stimulus with the lowest rating 0 point. For other stimuli, they should rate them in respect of the highest and lowest stimuli in the 0–100 scale according to the evaluation attribute. For the sessions of *degradation in spatial quality* and *degradation in timbre quality* evaluations, references (*non-widen*) were presented in GUI, as shown in Fig. 3.21, along with eight stimuli to be evaluated including hidden references, thus participants could directly compare the differences to the reference. In theses two sections, participants were instructed to rate the stimulus which was the most similar to the reference among the eight stimuli as 100 points, and to rate the stimulus degraded the most as 0 point. For the session of *perceived width* evaluation, since there was no presented reference stimuli, participants were instructed to rate the stimulus which they perceived as the widest among the eight stimuli as 100 points, and to rate the stimulus perceived as the narrowest as 0 point. Participants could switch between stimuli seamlessly to compare any pair among stimuli freely. They could take as long as they needed to finish the ratings of all stimuli. The average duration for the experiment was about 50 minutes.

3.4.2 Results

Evaluation of Perceived Width

Mean scores of perceived widths are shown in Fig. 3.22 for cello source and and Fig. 3.23 for pink noise source, with two center positions respectively. The ratings for three different bandwidths of the filter bank, 1/3-, 1/6-, and 1/12-octave were displayed in three different panels just for better visualization, although they were all evaluated in the same two consecutive comparison parts with the same reference stimuli (*non-widen* and *decor*). Hence, their

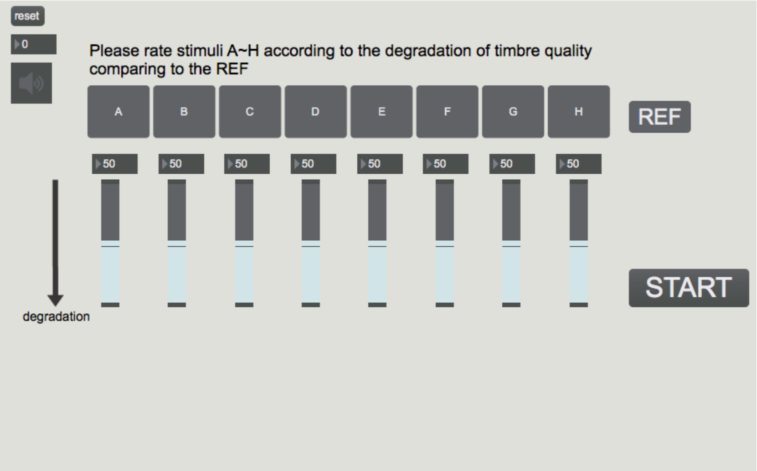

Please rate stimuli A~H according to the degradation of timbre quality comparing to the REF

Fig. 3.21 The GUI for the timbre degradation evaluation.

ratings can be compared directly, and the scores of *non-widen* and *decor* were identical in the three panels.

To verify whether the widening effect can control the perceived source width, the linear relationships between synthesis width and perceived width were investigated, while effects of the bandwidth on the perceived source width were also evaluated. Therefore, analysis of covariance (ANCOVA) was performed. The synthesis width was a covariate, *i.e.*continuous independent variable, and the bandwidth was a categorical independent variable with 3 levels. ANCOVA was performed with respect to ratings of the stimuli which were processed with the widening method, *i.e.*excluding the *non-widen* and *decor* stimuli, for 4 conditions (combinations of 2 source types and 2 center positions) respectively.

Before the analysis, a statistical test of the assumption of homogeneity of regression coefficients was conducted by testing the model including the interaction term. The results showed no significant interaction between synthesis width and bandwidth for all 4 conditions. Thus, ANCOVA could be performed.

For the cello source type, the results of ANCOVA are shown in Tables 3.8 and 3.9 for center positions of 0° and 15° respectively. For both cases, there was no significant difference in the effect of the bandwidth. However, there was a significant linear trend in mean perceived width as a function of synthesis width. The least squares estimate for the coefficient of covariate was positive (0.3695 for 0° and 0.3369 for 15°), meaning that the perceived width increased statistically significantly with the increasing synthesis width.

For the pink noise source type, however, there was no significant difference for either effects of bandwidth or synthesis width. These results indicated that the effectiveness of the widening processing depended on the characteristics of source signals and could also be associated with individual differences, which will be discussed in Section 3.4.3.

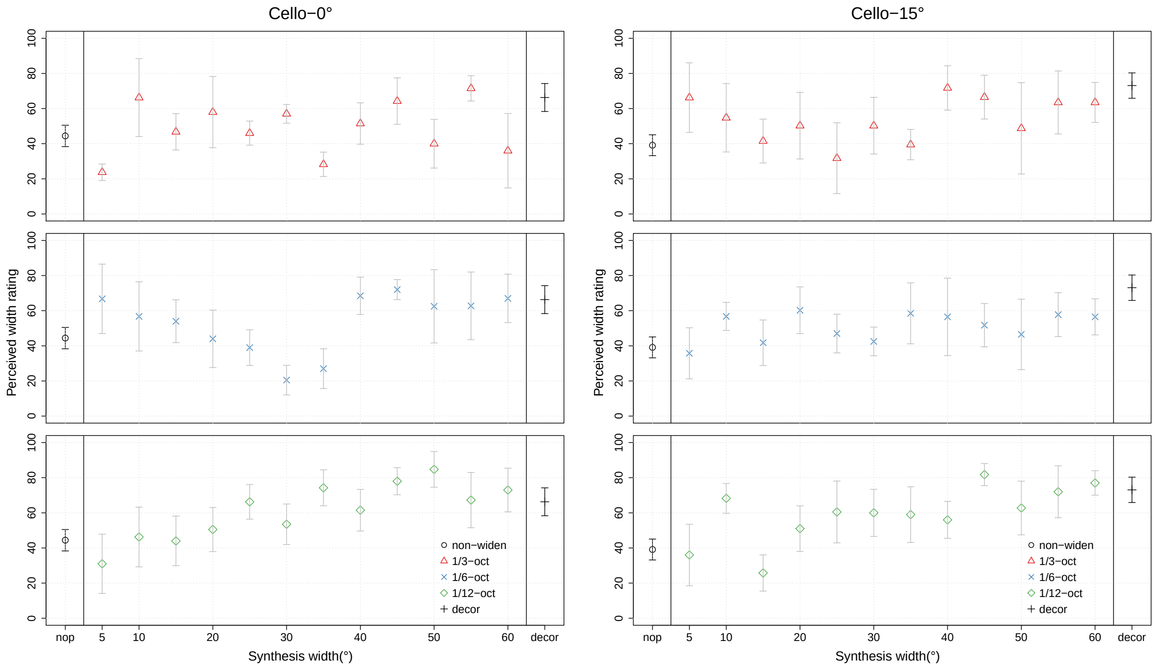


Fig. 3.22 Mean scores for *perceived width* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively.

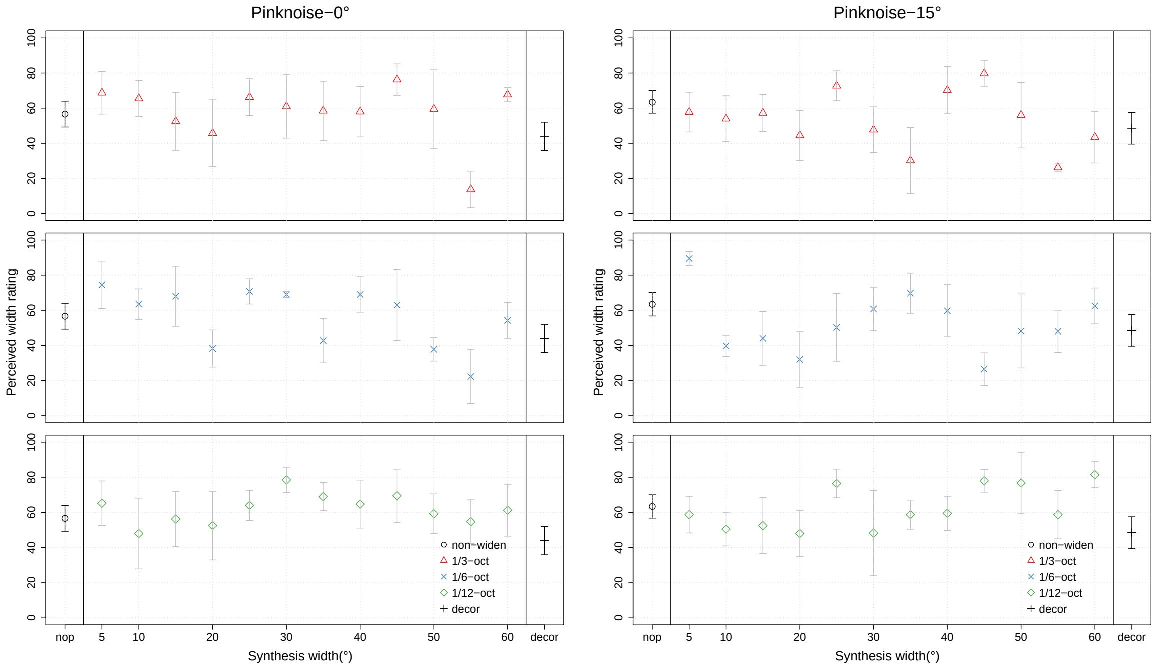


Fig. 3.23 Mean scores for *perceived width* ratings of stimuli of the pink noise source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively.

Table 3.8 Analysis of Covariance Table for cello source with 0° center position

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	5858	5857.8	7.491	0.007**
bandwidth	2	3394	1696.9	2.170	0.118
Residuals	140	109480	782.0		

Table 3.9 Analysis of Covariance Table for cello source with 15° center position

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	4869	4868.9	6.130	0.014*
bandwidth	2	1652	825.9	1.040	0.356
Residuals	140	111191	794.2		

Since there was no significant effect of bandwidth on perceived width for all conditions, linear regression with perceived width as the dependent variable and synthesis width as independent variable was performed separately for each bandwidth to further investigate the influence of bandwidth. The regression coefficients, R-squared, and the p-value are summarized in the Table 3.10. For cello sources, positive regression coefficients indicates that the perceived width increases with the increase in synthesis width, since the linear regression model was based on the relationship of $width\ ratings = coef \times synthesis\ width(°)$. However, the linear correlation was only significant for stimuli with the 1/12-octave bandwidth. For pink noise sources, regression coefficients were minus for 1/3- and 1/6-octave bandwidth, whereas for 1/12-octave bandwidth the coefficients were positive, although the correlations were significant only for 1/6-octave, 0° center, and for 1/12-octave, 15° center. Generally, the results suggest that as bandwidth decreases, the correlation between perceived width and synthesis width increases.

Interaural Cross-correlation Coefficient and Its Correlation with Perceived Width

Interaural cross-correlation coefficients (IACC) of stimuli were computed to investigate its relationship with parameters of processing and ratings of perceived widths. Since the stimuli were binaural signals, IACC was computed directly from the maximum absolute value of the cross-correlation function between two channel signals of each stimulus. Fig. 3.24 shows IACC of each stimulus as a function of the synthesis width. As can be seen, IACC decreased as synthesis width increased. This indicates that this widening processing reduced the IACC, which is generally assumed to have relationship with perception of source width [39].

However, the Pearson correlation coefficients between $1 - \mathrm{IACC}$ and ratings of perceived width computed across all stimuli in each condition were low ($< \pm 0.4$). For further investiga-

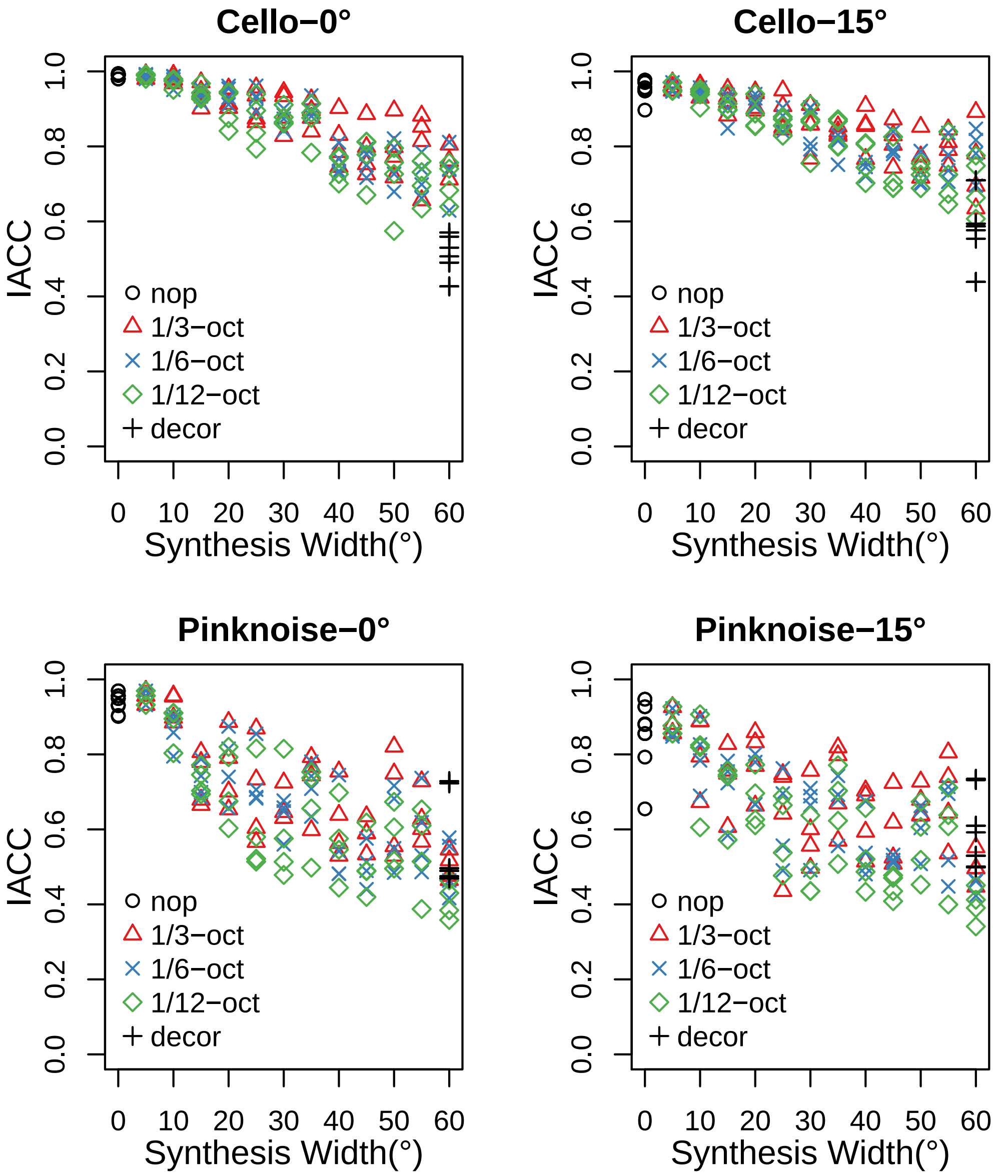


Fig. 3.24 IACC of stimuli from each condition as a function of synthesis width

Table 3.10 The results of regression analysis.

Condition	Bandwidth	coef	R^2	p-value
Cello-0°	1/3-oct	0.108	0.005	0.647
	1/6-oct	0.246	0.019	0.345
	1/12-oct	0.755	0.234	<0.001*
Cello-15°	1/3-oct	0.203	0.012	0.457
	1/6-oct	0.191	0.016	0.387
	1/12-oct	0.618	0.152	<0.01*
Pink noise-0°	1/3-oct	-0.212	0.015	0.400
	1/6-oct	-0.499	0.106	0.024*
	1/12-oct	0.081	0.003	0.710
Pink noise-15°	1/3-oct	-0.185	0.014	0.428
	1/6-oct	-0.128	0.006	0.602
	1/12-oct	0.422	0.081	0.049*

tion, the correlation coefficient between $1-\mathrm{IACC}$ and ratings of perceived width of stimuli tested by each participant was computed separately and listed in Table 3.11. As can be seen, for some participants the $1-\mathrm{IACC}$ and ratings were highly correlated (> 0.5). However, for other participants the correlations were low or even negative. This suggests that the perception of source width under this processing method differs among individuals.

Evaluation of Timbre Degradation and Spatial Degradation

Figs. 3.25 and 3.26 present mean scores for degradations of timbre and spatial quality respectively. It is evident that the ratings of both timbre and spatial quality degraded as synthesis width increased, and the decorrelated stimuli were rated averagely lowest.

ANCOVA was performed with the same procedure as performed on *perceived width* evaluation. Both results of evaluations of *timbre degradation* and *spatial quality degradation* for both center positions satisfied the assumption of homogeneity of regression coefficients. The results showed that for all evaluations, there were significant correlations between ratings and synthesis width. The least squares estimates for the coefficient of covariate were all negative (-0.80 for center position of 0° and -0.59 for 15° in *timbre degradation*, -0.60 for 0° and -0.83 for 15° in *spatial quality degradation*), indicating that the timbre or spatial quality degraded more as synthesis width increased. The ANCOVA tables for results of both evaluations are given in Tables 3.12 to 3.15.

Table 3.11 Pearson correlation coefficient between $1 - \text{IACC}$ and ratings of perceived widths for each participant under four conditions. The coefficients >0.5 are emboldened.

Participant	Cello 0°	Cello 15°	Pink noise 0°	Pink noise 15°
1	**0.703**	**0.647**	0.301	**0.659**
2	**0.545**	**0.648**	0.308	**0.882**
3	-0.021	0.301	-0.604	-0.616
4	**0.726**	-0.458	-0.073	-0.394
5	0.279	0.098	0.021	-0.248
6	-0.272	0.237	0.348	-0.026
7	**0.787**	**0.822**	-0.627	-0.291
8	-0.308	-0.151	-0.096	-0.177
9	0.098	**0.507**	-0.568	0.002
10	**0.518**	**0.912**	-0.149	0.314
11	**0.682**	**0.828**	0.348	**0.593**
12	-0.745	-0.527	-0.481	-0.548

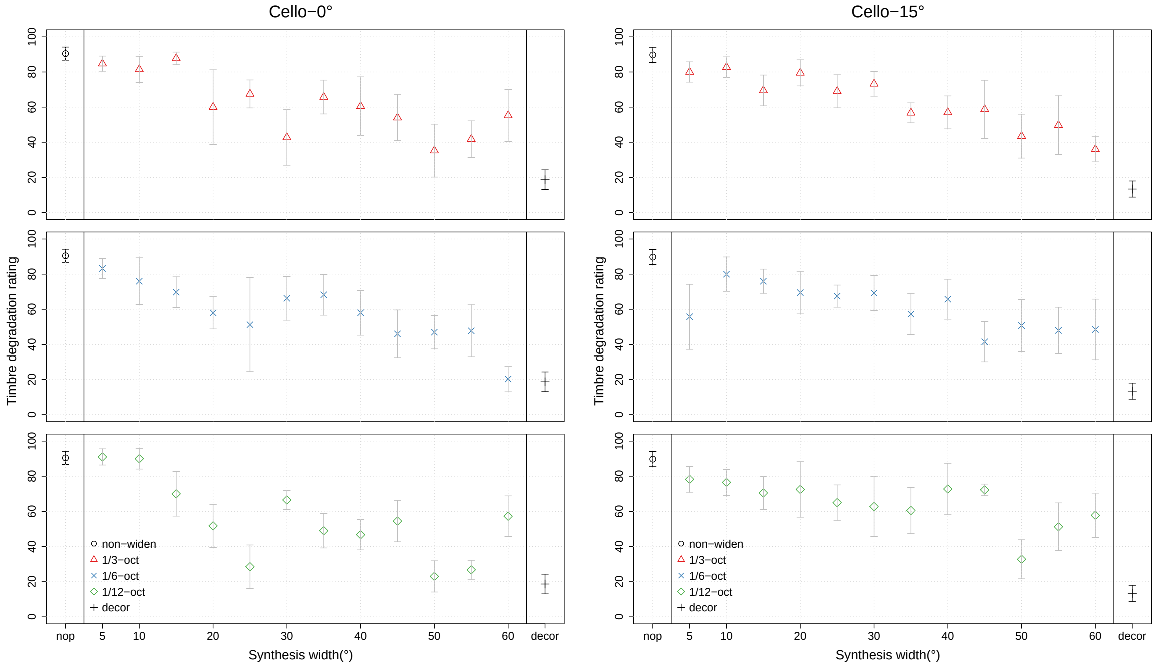


Fig. 3.25 Mean scores for *timbre degradation* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively.

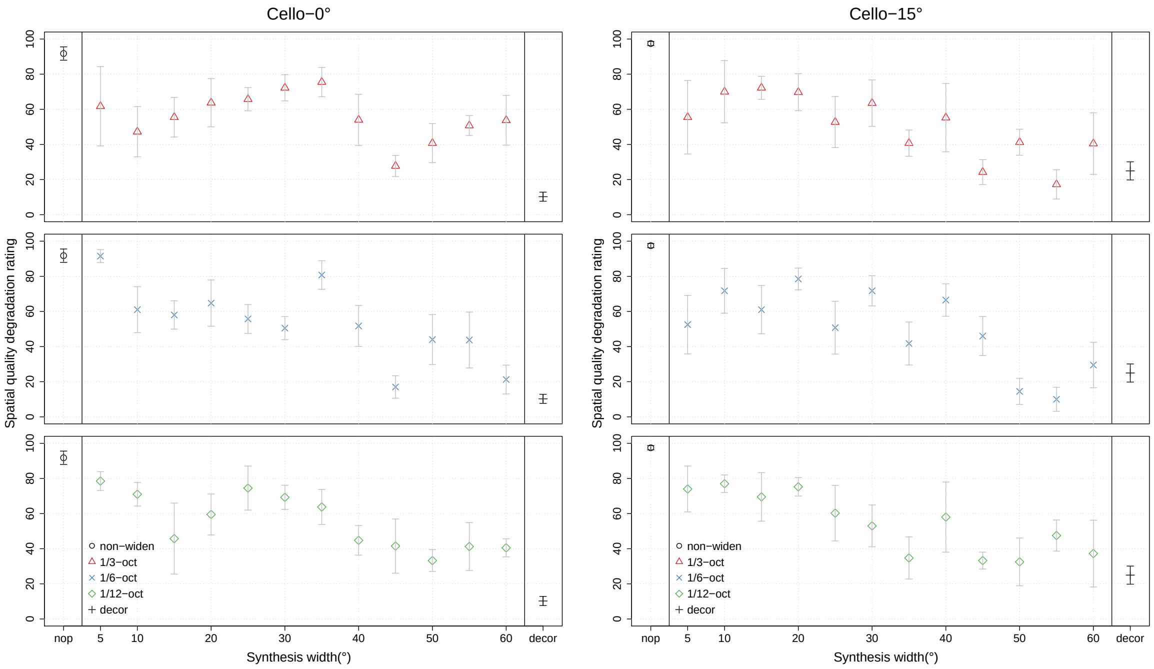


Fig. 3.26 Mean scores for *spatial quality degradation* ratings of stimuli of the cello source, along with the error bars representing standard deviations. Results for center position of 0° (left) and 15° (right) are shown respectively.

Table 3.12 Analysis of Covariance Table of timbre degradation ratings for cello source at 0° center position

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	27592	27592.2	47.4337	<0.0001***
bandwidth	2	1118	558.8	0.9607	0.3851
residuals	140	81438	581.7		

Table 3.13 Analysis of Covariance Table of timbre degradation ratings for cello source at 15° center position

Response: ratings	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	14782	14782.2	31.1460	<0.0001***
bandwidth	2	313	156.3	0.3294	0.7199
residuals	140	66446	474.6		

Table 3.14 Analysis of Covariance Table of spatial quality degradation ratings for cello source at 0° center position

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	15471	15471.0	27.7028	<0.0001***
bandwidth	2	156	78.1	0.1399	0.8696
residuals	140	78185	558.5		

Table 3.15 Analysis of Covariance Table of spatial quality degradation ratings for cello source at 15° center position

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
synthesis width	1	29257	29256.6	45.7444	<0.0001***
bandwidth	2	648	324.0	0.5067	0.6036
residuals	140	89539	639.6		

Comparison with Reference Stimuli

To evaluate the performance of widening processing, comparisons with reference stimuli were conducted by t-test.

The ratings of *perceived width* of stimuli with synthesis widths in the widest width level, *i.e.* [50°, 55°, 60°], were used to compare to the mean width ratings of decorrelated stimuli. The results under all four conditions with two source types and two center positions showed that there was no significant difference in ratings of *perceived width*. On the other hand, for *timbre degradation*, mean rating of stimuli with synthesis widths in the widest level was significantly higher than the mean of decorrelated stimuli for both 0° (t = 3.01, df = 44.54, p-value = 0.004) and 15° center position (t = 5.35, df = 52.9, p-value $<$ 0.0001). For evaluation of *spatial quality degradation*, the mean rating of stimuli with synthesis widths in the widest level was significantly higher than the mean of decorrelated stimuli for 0° (t = 6.93, df = 57.2, p-value $<$ 0.0001) but not significant for 15° center position (t = 0.77, df = 48.8, p-value = 0.447).

To compare to the stimuli without widening processing, the ratings of *perceived width* of stimuli with synthesis width from each width level were used respectively in the t-test. For stimuli of cello source with 0° center position, the mean rating of *perceived width* was significantly higher than stimuli without widening processing only for stimuli with synthesis widths in the widest width level (t = 2.27, df = 58, p-value = 0.027). For stimuli of cello source with 15° center position, the mean ratings of stimuli with synthesis widths in both level [50°, 55°, 60°] (t = 2.81, df = 58, p-value = 0.0067) and [35°, 40°, 45°] (t = 2.86, df

= 58, *p*-value = 0.0058) were significantly higher than the mean rating of stimuli without widening processing.

3.4.3 Discussion

The results of ANCOVA of *perceived width* evaluations showed that the perceived width increased with increasing synthesis width, suggesting that the widening processing was an effective method to synthesize width for monophonic sound objects in binaural reproduction. The similar results obtained from stimuli with different center positions implied that this method could also be applicable for sound objects located at various directions besides 0°. There was no significant effect of the bandwidth on the perceived width. However, the results of regression analysis indicated that dividing frequency components more finely could achieve a more stable widening effect. It can be assumed that with narrower bandwidth the components would be distributed more evenly into the region of intended source width. This influence could be more remarkable if the frequency bands are randomly distributed, as was done in this study, and if the source signals have a narrower frequency characteristic. Nevertheless, bandwidth which is too narrow could cause more serious timbre degradation.

The effectiveness was highly dependent on source characteristics, which was also observed in a previous study [21] and Experiment 1, as described in Section 3.2.2. The effect of synthesis width on perceived width was significant only for the cello source. In the case of pink noise, many participants reported that the changes or equalization of frequency components were more obvious so it was difficult to evaluate the difference in the perceived width. Pihlajämaki *et al.* [21] reported similar observations in which the resulting overall quality of pink noise was bad due to the comb-filtered sound. This result suggested that further modifications of parameters for certain source signals are necessary, or other widening processing approaches would be more suitable. Nevertheless, since width perception is known to be related to various attributes of sound sources such as durations and frequency characteristics, limitations of widening effect might be inevitable for certain types of sound sources.

The results of *t*-test comparing to the stimuli without widening processing showed that the difference in perceived width was significant only when the synthesis widths were above width levels of [35°, 40°, 45°]. This result was in accordance with the result of Experiment 2, in which the significant differences were found only when synthesis widths were higher than 40°. In previous studies [11, 21], it was found that perceived widths were less than half of real widths of the loudspeaker ensemble, which indicated that sound images fused together. In addition, due to the non-individual HRTFs, the localizations of stimuli without widening processing may not be sharply defined even though they were convolved only with HRTFs from one direction. For these reasons, the synthesis width may have to be large enough to produce significant wider sound image than sound without widening processing.

This widening processing increased the $1-\mathrm{IACC}$ of signals since the distribution algorithm basically created localization cues suggesting different directions in each frequency band, which resulted in less coherent binaural signals. However, the correlation between $1-\mathrm{IACC}$ and perceived widths varied between participants. It could be assumed that the difference depended on whether the attributes they used to evaluate perceived width related to the correlation between the left and right ear signals or not.

Individual differences were observed even though HRTF individualization was done. Although the effectiveness of individualization was not examined, it can be assumed that the individual differences could arise not only from non-individual HRTFs but also from different subjective criterions on the width evaluation, since most participants evaluated the widths with similar tendency as they have done in Experiment 2. According to the interview after the listening experiment, frequency characteristics such as the amount of high frequency components were mostly reported as different subjective criterions used to evaluate perceived widths other than spatial-related attributes. The changes in frequency characteristics were more obvious for the broadband pink noise source, which could be assumed to be related to the insignificant effect of the widening processing.

According to the result of t-test, the processing method proposed in this study showed comparable results in source widening with the conventional decorrelation method, which were implemented in binaural reproduction by locating the decorrelated pair with a 60° interval as in a conventional stereo loudspeaker setup. Still, it had better performance averagely in timbre and spatial quality since decorrelation methods usually manifest the “phasing” problem.

Degradations both in timbre and spatial quality were found after processing and the extent of degradation increased as synthesis width increased. However, since the degradation was evaluated by directly comparing to the unprocessed reference signal in the experiment, the degradation could be unnecessary to be interpreted as an unpleasant deterioration but rather just a change in timbre or spatial attributes. Furthermore, the result of Experiment 2 showed that when conducting pairwise comparisons between stimuli, there was no significant difference in preferences in timbre and spatial attributes between unprocessed stimuli and processed ones.

Random distribution of frequency bands used in this experiment could increase the variance of the effectiveness. Future work will therefore investigate the influence of the distribution of frequency components on perceived widths and develop an appropriate distribution method.

3.4.4 Conclusion

We have verified the effectiveness of the widening processing proposed in this experiment. By conducting listening experiments, the correlations between synthesis width and perceived width were investigated. The results showed that for the cello source signal with 1/12-octave bandwidth, perceived width increased with increasing synthesis width, suggesting that under appropriate conditions this method could effectively control the perceived width of a monophonic source in binaural synthesis. Analysis by comparing with stimuli without widening processing and decorrelated references suggested that after processing by the widening method the perceived width was significantly wider than unprocessed one and comparable to the decorrelated stimulus, while the degradations of timbre and spatial quality were less notable than the decorrelation method. However, some limitations were found such as source signal dependency, degradation, and subjective variation. This suggests modifications of parameters, including adjustment regarding source signal characteristics.

The parameter of bandwidth did not show significant effect on the perceived width. Nevertheless, it could be assumed that how finely frequency components were divided may have influence on the stability of performance and timbre degradation. For both center positions similar results were shown, suggesting that this method could be used on sources from different directions. Further studies are necessary for investigations of the effect of the source direction on widening performance including directions not on the horizontal plane.

3.5 Summary

In this chapter, the presented three experiments investigated a proposed method which aims to produce widths for monophonic sources in binaural synthesis. By dividing monophonic sources into octave bands and convolving frequency bands with HRTFs from different directions in a random or deterministic way, stimuli with different synthesis widths were generated. The effect of the widening processing method, and influences of parameters on width perception, were investigated through subjective listening experiments. According to the results, the processing method can successfully produce spatial widened sources in binaural synthesis. It has been found that several parameters of the processing have important influences on the perceived source width.

First, the distribution methods, which determine what direction each frequency band is distributed to, have a major influence on the localization of source width. Since usually there is a large variation in energy among different frequency bands of source signals, how to distribute the energy of the signal uniformly to the range of intended synthesis width is a major challenge to be solved for this processing method. If it can be achieved, not only can localization shifted from the intended direction be avoided, but the spatially quality of the source width may also be improved.

This brings up the idea that how finely frequency bands were divided, *i.e.*, the bandwidth, is also a crucial factor to the performance of source widening. A narrower bandwidth was found to able to ensure a more stable performance of source widening effect. The narrowest bandwidth investigated in this study is 1/12 octave bands. It may be reasonable to suppose that dividing the frequency bands more finely could further improve the performance, but timbre degradation may also occur.

Different directions of the centers of the source width were examined to investigate the effectiveness of widening processing on source at various directions. The results show similar effect of widening process, although only centers of 0° and 15° azimuth were investigated.

The effectiveness of source widening was only significant with sufficiently large synthesis width (generally larger than 40°). In addition, individual differences and a dependency of performance on characteristics of source signals were found, suggesting that further investigation and improvement of the widening processing are needed.

The purpose of developing a source widening effect for binaural synthesis is to provide a method to create and control the width of sound objects in binaural reproduction, with an assumption that producing sound objects with spatial extent, which is closer to the complex auditory events we usually experience in natural auditory environments, can achieve better spatial impression. The widening effect can be used in a wide range of applications such as binaural reproduction of object-based audio, binaural mixing consoles, and virtual reality audio. Therefore, to demonstrate the feasibility of applying the widening effect in audio production, an experiment was conducted and is presented in the next chapter.

Chapter 4

Spatial impression of source widening effect for binaural audio production

4.1 Introduction

In Chapter 3, three experiments were conducted to investigate the effect of the proposed source widening processing method. The results of subjective listening experiments showed that source signals could have wider perceived widths after processed by widening effect. In this study, to investigate the effect of the widening processing on the spatial impression in the practical application of audio production, the widening process method was implemented as a VST plugin for real-time processing in binaural audio reproduction. VST (Virtual Studio Technology) is an audio plugin software interface which can accommodate digital audio signal processing of synthesizers and effects in digital audio workstation (DAW). The VST plugin was used for sound effects mixing with widening effect applied to source signals of sound effects. Instead of using noises and anechoic recordings as in experiments described in Chapter 3, the source signals were chosen considering the practical audio production situation. The aim of this experiment was to investigate whether the synthesis of sound source widths could provide a better spatial impression of the work.[1]

[1] The results of this experiment were also published in *H. Su, A. Marui, and T. Kamekawa, "Spatial Impression of Source Widening Effect for Binaural Audio Production," presented at the Audio Engineering Society Conference: 2018 AES Int. Conf. on Spatial Reproduction-Aesthetics and Science (2018).*

4.2 Method

4.2.1 Plugin Algorithm

A VST plugin was designed and generated using the audioPlugin object in Audio System Toolbox of MATLAB. The processing algorithm was basically the same as that used in previous experiments. The source signal was first divided into multiple sub-bands by a 1/3-octave filter bank, then each band was randomly assigned to a direction and convolved with the HRTF of that direction. Finally, convolved signals were summed up for reconstruction of the original signals. However, considering the computational efficiency of the real-time processing, reduction of the number of times of convolution processing is necessary. Thus, HRTFs were divided into 1/3 octave bands and saved in the frequency domain beforehand. The realtime processing in the plugin was simply performing Fourier transform of the input signal, combination of HRTF sub-bands corresponding to the center and the width parameters, and then multiplication of the FFT-ed input signals and the combined HRTF, as illustrated in Fig. 4.1.

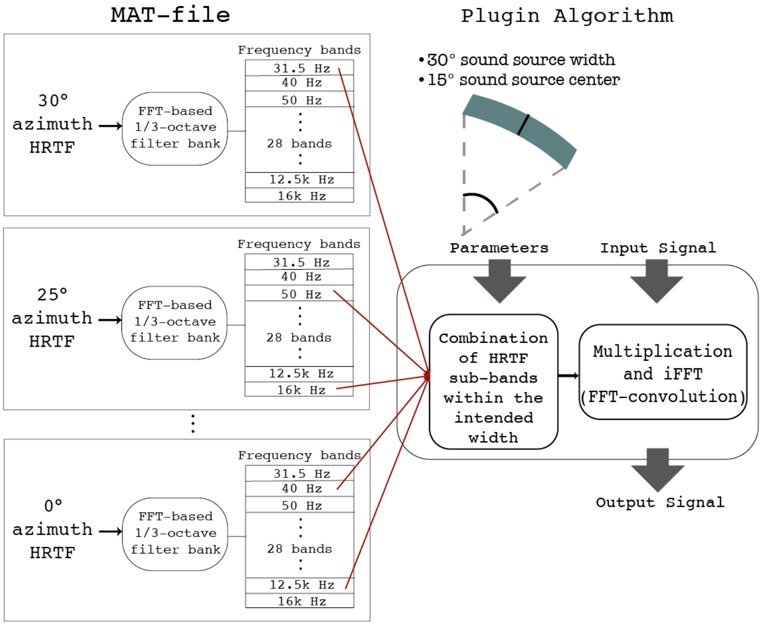
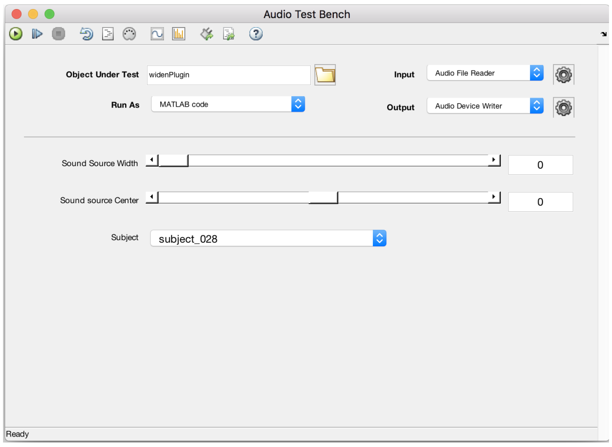


Fig. 4.1 The processing algorithm of the widening effect plugin. The example is for the parameter of *sound source width* set to 30° and the *sound source center* set to 15°.

The HRTF datasets were from the CIPIC database [2], in which HRIRs of 45 subjects were provided. HRTF sets of four subjects among them were included in the plugin, which could be selected in the plugin interface according to preferences of listeners. The HRTFs of the four subjects were chosen as representatives based on k-means cluster analysis, which was the same as the HRTFs used in the Experiment 3 as described in Section 3.4.1.

There were two other parameters could be adjusted in the plugin interface: the *sound source width* and the *sound source center*, as shown in Fig. 4.2. The *sound source width* defined the range of azimuths of HRTFs used for distributing frequency bands of input signals. The allowable values were from 0° to 60° azimuth. The *sound source center* defined the azimuth of the center of the sound source width, thereby directly influencing the localization of the sound source. Since the processing algorithm distributed frequency bands to HRTFs with 5° spatial resolution, which were only available between $-45°$ to 45° for the CIPIC database, the allowable values for center positions were from $-15°$ to 15° considering that the maximum source width extended to the range of ±30° from the center. For example, if the *sound source width* was set to 30° and the *sound source center* was set to 15°, HRTFs from 0° to 30° azimuth were used. Sub-bands of HRTFs were then picked and combined based on the random decided distribution order, which is identical to the random method of experiment 3 described in Section 3.4.1. For example, if the sub-band of 31.5 Hz was randomly distributed to 30° HRTF, the sub-band of 31.5 Hz of 30° HRTF was picked to be used in the combination of HRTFs.



Object Under Test: widenPlugin

Run As: MATLAB code

Input: Audio File Reader

Output: Audio Device Writer

Sound Source Width: 0

Sound source Center: 0

Subject: subject_028

Ready

Fig. 4.2 The interface of source widening plugin.

4.2.2 Mixing Experiment

Two engineers participated the mixing task. Both of them were Master's degree students in Research Area of Creativity of Music and Sound, Tokyo University of the Arts, and had received training for recording and mixing engineering for more than four years. They also had sufficient experience of sound design for movie, animation, and games.

The experiment was conducted in Studio B in Senju Campus of Tokyo University of the Arts. A monitor was connected to a laptop used for mixing to display the video content. Ableton Live 10 was used as a digital audio workstation (DAW) to host the VST plugin and for automation editing of parameters of the plugin. The audio was reproduced via an audio interface (RME Fireface UFX) and a headphone (AKG K240 MKII). Fig. 4.3 shows a photo of the actual condition of the mixing experiment.

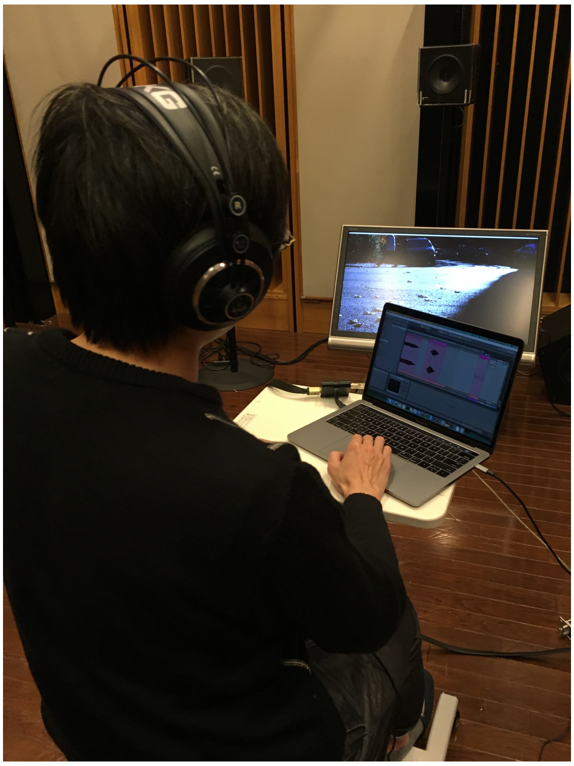


Fig. 4.3 The setting for the mixing experiment.

A 20 seconds long video with a street view, leaves blowing, a passing car, a bus turning, and a woman walking through was used as a material. Sound effects for leaves, the car, the bus, footsteps, and a street ambience were edited, and the volume of the mix was adjusted by the author beforehand. Participants were instructed to adjust and edit automations of parameters of the *sound source width* and the *sound source center* in the plugin for the four sound effects except ambience respectively. They could only control these two parameters, no other editing or mixing performance was allowed.

Before the mixing task, the subjective selection of HRTF described in Section 3.4.1 of non-individual HRTF was performed. The HRTF set used for the stimulus rated with the highest score in the subjective selection was selected in the widen plugin in all tracks. The participants could perform mixing until they were satisfied with the work. It took about 30

Table 4.1 Values or ranges of values of the *sound source width* parameter of each sound effect used by the two participants.

	Participant A	Participant B
foot steps	0°–50°	21°
bus	8°–54°	17°
car	7°–57°	13°
leaves	0°–6°	48°

minutes to 40 minutes for the experiment, including the subjective selection and the mixing task.

4.2.3 Mixing Results

The automations of the sound source width parameter edited by two participants were examined. It shows that participant A used more automation to change sound source width with time. On the other hand, participant B used rather steady values for each sound effect. Table 4.1 lists the range of values of the *sound source width* of each sound effect used by the two participants respectively.

4.2.4 Stimuli

After mixing, three versions of audio were exported along with the video for two mixes by the two participants respectively, resulting totally six stimuli. One version was the original mix (denoted as A and B as for two mixing participants respectively), the other was a mix in which the automation of the parameter of *sound source width* was disabled and set to 0 all the time (A0 and B0), and the third version was a mix in which all the breakpoints of automation envelopes of the *sound source width* parameter were adjusted to the half values (Am and Bm), *i.e.*, only half of the original source width. These six mix versions were exported with selecting each of the 4 HRTF sets respectively as stimuli for the subjective listening experiment.

4.2.5 Listening Experiment

The subjective listening experiment was conducted in the same room with the same equipment and setup as mixing experiment. 10 participants from Tokyo University of the Arts took part in the experiment. All were students major in Research Area of Creativity of Music and Sound and had experience in listening experiments previously. Before the experiment, the subjective selection of HRTFs described in Section 3.4.1 was conducted individually for each participant. Stimuli generated by the HRTF set rated highest by the participant were then used in the experiment.

The stimuli were replayed by GUI with video routed to the monitor and sound routed to audio interface and headphone. The 6 stimuli were randomly ordered and presented as stimulus A to F, as shown in Fig 4.4. Participants could click the button to replay each stimulus freely and use a bar on GUI to control the playback. Participants were asked to use the GUI to evaluate each stimulus on a scale of 0–100 according to the overall performance of spatial impression of the mix and were encouraged to use the whole scale. It took about 15–20 minutes for the experiment, including the instruction, subjective selection of HRTFs, and the main experiment.

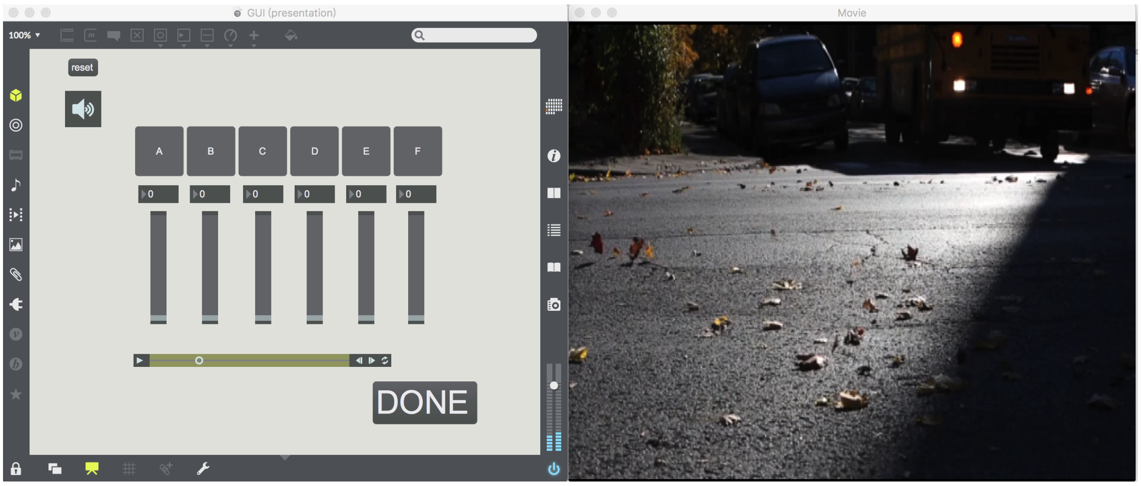


Fig. 4.4 GUI constructed in Cycling'74 Max for the subjective listening experiment.

4.3 Result

Fig. 4.5 shows a box plot of ratings for each stimulus. The result suggests that ratings depend largely on individual preferences since ranges of ratings for most stimuli were almost across the whole scale. However, on average the original versions were rated slightly higher than versions with 0° width.

4.4 Discussion

Since the evaluation was based on the overall spatial impression of sound effects, different criteria of different participants would lead to various tendencies of ratings. According to

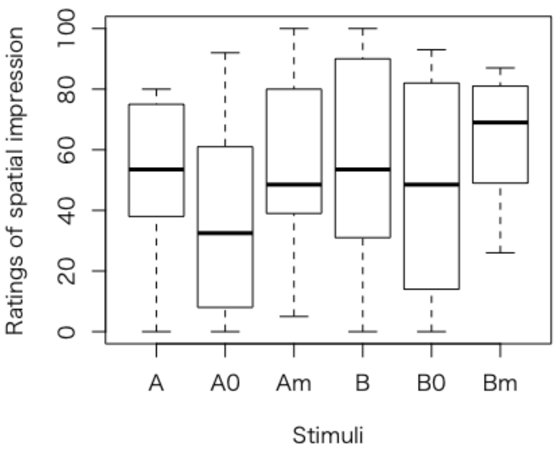


Fig. 4.5 Box plot of the ratings for each stimulus

interviews after the experiment, most participants rated mainly based on performances of movements and localizations of sound effects. Participants could prefer stimuli with 0° width since the localization and movement could be perceived more clearly without widening processing. Nevertheless, some participants still showed preferences for original versions and described them with more extent of environmental sounds.

One thing should be noted is that the random distribution of frequency bands performed in the plugin could be a cause for concern. As the plugin parameters would be reset every time the DAW was launched, the pattern of random distribution which the mixers listened to while mixing would differ from the one participants (or listeners) listened to. Since the distribution was random, the performance of widening processing also varied, since it has been found to be influenced by the distribution result in the previous experiments, as described in Chapter 3, such as the uniformity of energy distribution. This randomly varied performance would also significantly influence the spatial impression. Thus, an appropriate deterministic distribution method should be proposed in future work.

Since this experiment was just a preliminary attempt to accommodate the widening processing into audio production, the plugin algorithm used a bandwidth of 1/3-octave and a random method for simplicity. Using a narrower bandwidth may obtain a more stable widening performance according to the result of Experiment 3, although the processing load could become heavy since there are more bands to process for the narrower bandwidth. In addition, if a deterministic distribution method according to the spectral characteristics of the input signals, such as the *powerorder* method in Experiment 2, is used, analysis of the input signal would be necessary. Therefore, deterministic distribution methods independent to the input signal may be needed for an efficient plugin processing.

To conclude, synthesis of width for sound objects in a binaural reproduction could improve the spatial impression, but the performance may depend on nature of the source, such as whether it was moving or static, *i.e.*, whether the sharpness of localization was important for spatial impression or not. There could be a trade-off between the perception of spatial extent and localization.

4.5 Summary

In this experiment, we presented a tool to control sound source widths in binaural reproduction as a VST plugin which could perform real-time widening processing. To investigate the source widening effect when applied to audio production, sound effects mixing for a video clip using the plugin was performed. A subjective listening experiment was conducted to evaluate the overall spatial impression. The results show that even though there were different preferences in sound effect mixes due to individual criteria, synthesizing widths for sound objects could improve the overall spatial impression.

Chapter 5

Summary

This study aims to develop a source widening effect to create and control the source width in binaural synthesis. The approach of distributing frequency components across different directions to create a sound image with localization cues varying with frequency, which was proposed in previous studies for loudspeaker reproduction, was implemented in binaural synthesis. A processing method was proposed, and three experiments were conducted to examine different aspects and different parameters of the method. In addition, the widening processing was implemented in a VST plugin which can be used as a widening effect for audio mixing, and experiments including sound effects mixing and subjective evaluations were conducted to verify the feasibility of the source widening effect. The results demonstrated that the widening processing could successfully create and control the source width in binaural synthesis, and could actually be applied to audio production. However, the effectiveness was only significant when the synthesis width was large enough, suggesting that the processing method still needs improvement. Furthermore, some questions remain unsolved and further work is needed.

First, individual differences could be a crucial issue when considering the effectiveness of the processing method. Different tendency in evaluations of perceived source width by participants were found in the results of the listening experiments. One possible reason for the problem could be the individual subjective criterions of participants for the evaluation related to width perception. Another reason may be the non-individual HRTFs used in this study. However, although the individualization of HRTFs by subjective selection was performed, the problem of individual differences still existed. The effectiveness of individualization should be further investigated in the future work. On the other hand, head-tracking has been found to be a more effective way than HRTF individualization to resolve problems such as inside-head localization, which could be an essential issue for the width perception. In addition, the most promising application of the proposed widening effect should be in virtual reality. Therefore, incorporating the processing method into a VR system with head-tracking to investigate the effect on width perception will be worthwhile for future work.

Second, the effect of the widening processing varied depending on the source signals. This was not surprising since the processing method involved frequency band distribution, so it is reasonable to assume that the effect would depend on spectral characteristics of the source signals. In addition, the width perception has been found to depend on acoustical attributes such as level, duration, and frequency. Although limitations would still exist due to the fact that width perception is fundamentally affected by other acoustic features of signals, with the further improvement of the processing method by optimizing the parameters, it can be assumed that the effect could still be improved to some extent. For example, a deterministic distribution method, which can distribute the energy of the signal uniformly according to the spectral characteristic of the source signal, should be proposed. Further work can also investigate the influence of dividing the frequency bands further finely, since the results of Experiment 3 suggest that narrower bandwidth could ensure the stability of the performance. However, there may be a trade-off between timbre quality and widening effect.

Finally, only synthesis widths with centers at 0° and 15° azimuth were investigated. Since in the application of this processing method, synthesis source width in various directions would be necessary, the influence of centers at various directions, such as directions other than the front side, on the widening effect should be examined if sufficient spatial resolution of the HRTF database is available.

References

[1] V. R. Algazi and R. O. Duda. Headphone-based spatial sound. *IEEE Signal Processing Magazine*, 28(1):33–42, 2011.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102. IEEE, 2001.

[3] J. Blauert and R. Rabenstein. Providing surround sound with loudspeakers: A synopsis of current methods. *Archives of Acoustics*, 37(1), 2013.

[4] J. I. Bömer, S. Oode, and A. Ando. Effect of frequency bandwidth on interaural cross-correlation in relation to sound image width of reproduced sounds of a violin. *Applied Acoustics*, 72(9):623–631, 2011.

[5] C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *J. Audio Eng. Soc*, 49(4):231–249, 2001.

[6] G. Davidson, D. Darcy, L. Fielder, Z. Schuang, R. Graff, J. Breebaart, and P. Crum. Design and subjective evaluation of a perceptually-optimized headphone virtualizer. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[7] A. Dean, D. Voss, and D. Draguljić. Analysis of covariance. In *Design and Analysis of Experiments*, pages 285–304. Springer International Publishing, Cham, 2017.

[8] G. Enzner, C. Antweiler, and S. Spors. Trends in acquisition of individual head-related transfer functions. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 57–92. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[9] W. G. Gardner and K. D. Martin. HRTF measurements of a kemar. *J. Acoust. Soc. Am*, 97(6):3907–3908, 1995.

[10] D. Hammershøi and H. Møller. Binaural technique – basic methods for recording, synthesis, and reproduction. In *Communication Acoustics*, pages 223–254. Springer, 2005.

[11] T. Hirvonen and V. Pulkki. Center and spatial extent of auditory events as caused by multiple sound sources in frequency-dependent directions. *Acta Acustica united with Acustica*, 92(2):320–330, 2006.

[12] T. Hirvonen and V. Pulkki. Perceived spatial distribution and width of horizontal ensemble of independent noise signals as function of waveform and sample length. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[13] G. S. Kendall. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4):71–87, 1995.

[14] A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert. An introduction to binaural processing. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 1–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[15] R. Mason. How important is accurate localization in reproduced sound? In *Audio Engineering Society Convention 142*, May 2017.

[16] R. Mason, T. Brookes, and F. Rumsey. Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *J. Acoust. Soc. Am*, 117(3):1337–1350, 2005.

[17] A. W. Mills. On the minimum audible angle. *J. Acoust. Soc. Am*, 30(4):237–246, 1958.

[18] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J. Audio Eng. Soc*, 43(5):300–321, 1995.

[19] B. Olufsen. Music for archimedes. *Compact disc CD B&O 101*, 1992.

[20] D. R. Perrott and T. N. Buell. Judgments of sound volume: Effects of signal duration, level, and interaural characteristics on the perceived extensity of broadband noise. *J. Acoust. Soc. Am*, 72(5):1413–1417, 1982.

[21] T. Pihlajamäki, O. Santala, and V. Pulkki. Synthesis of spatially extended virtual source with time-frequency decomposition of mono signals. *J. Audio Eng. Soc.*, 62(7/8):467–484, 2014.

[22] C. Pike and F. Melchior. An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[23] G. Potard and I. Burnett. A study on sound source apparent shape and wideness. In *Proc. of the 2003 Int. Conf. on Auditory Display*, 2003.

[24] G. Potard and I. Burnett. Decorrelation techniques for the rendering of apparent sound source width in 3D audio displays. In *Proc. Int. Conf. on Digital Audio Effects (DAFx'04)*, 2004.

[25] V. Pulkki and M. Karjalainen. Spatial hearing. In *Communication Acoustics: an introduction to speech, audio and psychoacoustics*, pages 219–247. John Wiley & Sons, 2015.

[26] V. Pulkki, T. Lokki, and D. Rocchesso. Spatial effects. In *DAFX: Digital Audio Effects*, chapter 5, pages 139–183. Wiley-Blackwell, 2011.

[27] S. Satoh. *Statistical Methods in Sensory Tests (in Japanese)*. Number 19. Nikkagiren Publishing, 1985.

[28] Z. Schärer and A. Lindau. Evaluation of equalization methods for binaural signals. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[29] H. Scheffe. An analysis of variance for paired comparisons. *J. Am. Stat. Assoc.*, 47(259):381–400, 1952.

[30] B. U. Seeber and H. Fastl. Subjective selection of non-individual head-related transfer functions. Georgia Institute of Technology, 2003.

[31] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke. Platform independent audio. In *Media Production, Delivery and Interaction for Platform Independent Systems*, chapter 4, pages 130–165. Wiley-Blackwell, 2013.

[32] K. Sunder, J. He, E. L. Tan, and W.-S. Gan. Natural sound rendering for headphones: integration of signal processing techniques. *IEEE Signal Processing Magazine*, 32(2):100–113, 2015.

[33] R. P. Tame, D. Barchiese, and A. Klapuri. Headphone virtualization: Improved localization and externalization of non-individualized HRTFs by cluster analysis. In *Audio Engineering Society Convention 133*, Oct 2012.

[34] I. T. Union. Recommendation itu-r bs.2076-1: Audio definition model, 2017.

[35] M. Vorländer. Convolution and sound synthesis. In *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, pages 137–146. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[36] M. Vorländer. Signal processing for auralization. In *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, pages 103–122. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[37] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato. Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoust. Sci. & Tech*, 35(3):159–165, 2014.

[38] S. Xu, Z. Li, and G. Salvendy. Individualization of head-related transfer function for three-dimensional virtual auditory display: A review. In R. Shumaker, editor, *Virtual Reality*, pages 397–407, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[39] T. Ziemer. Source width in music production. methods in stereo, ambisonics, and wave field synthesis. In *Studies in Musical Acoustics and Psychoacoustics*, pages 299–340. Springer, 2017.

[40] F. Zotter and M. Frank. Efficient phantom source widening. *Archives of Acoustics*, 38(1):27–37, 2013.

[41] F. Zotter and M. Frank. Phantom source widening by filtered sound objects. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[42] F. Zotter, M. Frank, G. Marentakis, and A. Sontacchi. Phantom source widening with deterministic frequency dependent time delays. In *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2011.

Appendix A

List of Publications

Journal Articles

- Hengwei Su, Atsushi Marui, and Toru Kamekawa. “The Auditory Source Widening Effect in Binaural Synthesis with Spatial Distribution of Frequency Bands,” Journal of the Audio Engineering Society, accepted.

Presentations

- Hengwei Su, Atsushi Marui, and Toru Kamekawa, “Virtual Source Width in Binaural Synthesis with Frequency-Dependent Directions,” presented at the Audio Engineering Society Convention 142. Engineering Brief 327, Audio Engineering Society. Berlin, Germany. May 2017.
- Hengwei Su, Atsushi Marui, and Toru Kamekawa, “Frequency Bands Distribution for Virtual Source Widening in Binaural Synthesis,” presented at the Audio Engineering Society Convention 143. Convention Paper 9867, Audio Engineering Society. New York, NY, USA. October 2017.
- Hengwei Su, Atsushi Marui, and Toru Kamekawa, “Spatial Impression of Source Widening Effect for Binaural Audio Production,” presented at the Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science. Engineering Brief 76, Audio Engineering Society. Tokyo, Japan. August 2018.
- Hengwei Su, Atsushi Marui, and Toru Kamekawa, “The Effect of HRTF Individualization and Head-Tracking on Localization and Source Width Perception in VR,”

presented at the Audio Engineering Society Convention 146. Engineering Brief 520, Audio Engineering Society. Dublin, Ireland. March 2019.

Appendix B

Experiment Instructions

Experiment 1 (English translation)

Introduction

Thank you for participating in this experiment. This study aims to investigate the perceived source width of binaural synthesis for headphone reproduction. Please use the mouse and keyboard to answer questions on the GUI on the computer. The estimated time for the experiment is about 1 hour.

Procedure

There are 4 sections in this experiment. In sections 1 and 2, please answer how the perceived source width distributes on the horizontal plane in azimuths for each stimulus. In addition, if you perceive the in-head localization (the sound image is inside your head), and/or the sound image is moving while the stimuli is replayed, please check the corresponding box. In sections 3 and 4, please answer the degree of naturalness in a 7-point scale for the naturalness of spatial impression and the naturalness of timbre respect to each stimulus. The order of the 4 sections is random and different for each participant.

1. Sections 1 and 2: perceived source width

    - Click the "START" on the GUI to start the experiment. The stimulus for question No. 1 will be replayed. Please use the bar on the GUI to select the range of azimuths of the perceived width. The range is from $-60°$ to $60°$ in $5°$ intervals.
    - You can use the 5 loudspeakers behind the computer and the numbers on the black clothes between the loudspeakers as the references for the azimuths of the perceived source width. On the GUI, the 5 images of loudspeakers above the bar correspond to the $-60°$, $-30°$, $0°$, $30°$, and $60°$ azimuth respectively. You can

click the bottom above the image, and the reference sound from that direction will be replayed. Please use them as references.

- If you perceive the in-head localization, and/or the sound image is moving (the localization changes) while replay, please check the corresponding box. When you finish the question, please click "NEXT" to answer the next question.
- There are 55 questions in 1 section. There will be a message indicating the end of the section if all the questions were finished.

2. Section 3 and 4: naturalness

   - Click the "START" on the GUI to start the experiment. The stimulus for question No. 1 will be replayed. please answer the naturalness of spatial impression and the naturalness of timbre respect to each stimulus in a 7-point scale. The 7 point indicates natural and the 1 point unnatural. Please use the radio button to chose the 7-point scale.
   - When you finish the question, please click "NEXT" to answer the next question.
   - There are 34 questions in 1 section. There will be a message indicating the end of the section if all the questions were finished.

3. You can take a rest between sections.

4. Please adjust gain on the interface to a suitable replay level. Please don't adjust the level after the experiment begins.

5. There are total 178 questions. The estimated time for the experiment is about 1 hour.

バイノーラルシンセシスにおける音像幅について

実験機関名：国立大学法人東京藝術大学

実験責任者：蘇恒緯

東京藝術大学　大学院音楽研究科　音楽文化学専攻　音楽音響創造研究分野　博士１年

1　実験概要

この度は実験に参加していただき、誠にありがとうございます。本研究では、ヘッドホン聴取におけるバイノーラルシンセシスの音像幅について考察します。実験参加者の皆様にはマウスとキーボードを使用して、パソコン上の指定のアプリケーションでの回答を行って頂きます。実験時間は 60 分前後を予定しております。実験参加者の心身の安全には十分注意をして実験を行いますが、万が一実験中に苦痛や不快感などを感じて、体調が悪くなる等の症状が出た場合はすぐに実験責任者にお知らせ下さい。途中で実験を中断しても構いません。尚、実験の中断により実験参加者に何ら不利益を被ることはございません。またこの実験で収集した情報は、この研究に関してのみ使用いたします。

2　実験の流れ

本実験は 4 つのセッションがあります。セッション 1、2 では、各設問の刺激に対して、知覚した音像幅が空間の水平面における分布を方位角で回答してください。それに、音像が頭内定位 (音を頭の中に感じる) かどうかと、刺激が再生中に動いているかどうかを回答してください。セッション 3、4 では、音の空間の自然さと音色の自然さにおいて、各刺激がどの程度自然なのかを判断して、七段階で評価してください。4 つのセッションの前後順序は実験参加者によって違います。

1. セッション 1、2：音像幅
   - 画面上の「START」を押すと実験が始まり、問題 1 の刺激が流れます。刺激の音像幅を、画面上のバーで角度の範囲を選択してください。範囲は-60 度から 60 度まで、5 度間隔で選択できます。
   - パソコンの後ろに置いた５つのスピーカーと黒幕に表示された数字は、音像幅角度の判断の基準点として参考しください。回答用のバーの上に、その５つのスピーカーを意味する画像があり、それぞれ方位角-60 度、-30 度、0 度、30 度、60 度と対応しています。それらの上のボタンを押すと、その角度の基準音が再生されるので、それを基準として利用ください。
   - そして、頭内定位や再生中に音像が動いた（定位が変わった）と知覚すれば、チェックボックスでそれぞれに対応する項目を選択してください。回答が終われば、「NEXT」を押して、次の問題を回答してください。
   - 1 セッションは 55 問があります。全部の問題が終わると、画面上に実験完了を意味するメッセージが出ます。

2. セッション 3、4：自然さ
   - 画面上の「START」を押すと実験が始まり、問題 1 の刺激が流れます。刺激の空間印象に関する自然さと、音色に関する自然さを、7 段階で評価してください。7 は、自然と意味し、1 は不自然と意味します。オプションボタンで段階を選択してください。
   - 回答終われば、「NEXT」を押して、次の問題を回答してください。
   - 1 セッションは 34 問があります。全部の問題が終わると、画面上に実験完了を意味するメッセージが出ます。

3. セッションの間に、ヘッドホンを外して休憩しても構いません。すぐ次のセッションに入っても構いません。

4. 実験が始まる前に、お好きな音量をインターフェースによって調整してください。始まってから音量を変わらないようにお願いします。

5. 問題は全部で 178 問あり、実験時間は 60 分程度です。実験中にいつでも中断や中止をすることができます。

Experiment 2 (English translation)

Introduction

Thank you for participating in this experiment. This study aims to investigate the perceived source width of binaural synthesis for headphone reproduction. Please use the mouse to answer questions on the GUI on the computer. The estimated time for the experiment is about 1 hour.

Procedure

In this experiment, please compare the presented stimuli pair according to the perceived source width and the naturalness. In each question, two sound clips of A a B will be replayed in order. Please judge which one is wider than the other and evaluate the degrees of difference. For the stimuli of instruments recording, please also evaluate the differences in naturalness of spatial impression and the naturalness of timbre. Please use the scales described in Table 1 to perform the 7-point scale evaluations. There are three sections, and there are 72 questions for each section. The estimated time for the experiment is about one hour.

- Click the “START” on the GUI to start the experiment. The stimulus for question No. 1 will be replayed. After the two sound clips of A and B are replayed, please select the scale to answer the question. Click the “REPLAY” and the stimuli will be replayed again.

- When you finish the question, please click “NEXT” to answer the next question.

- There will be a message indicating the end of the section if all the questions were finished.

- You can take a rest between sections.

- Please adjust gain on the interface to a suitable replay level. Please don’t adjust the level after the experiment begins.

バイノーラルシンセシスにおける音像幅について

実験機関名：国立大学法人東京藝術大学
実験責任者：蘇恒緯
東京藝術大学　大学院音楽研究科　音楽文化学専攻　音楽音響創造研究分野　博士 2 年

1　実験概要

この度は実験に参加していただき、誠にありがとうございます。本研究では、ヘッドホン聴取におけるバイノーラルシンセシスの音像幅について考察します。実験参加者の皆様にはマウスを使用して、パソコン上の指定のアプリケーションでの回答を行って頂きます。実験時間は 60 分前後を予定しております。実験参加者の心身の安全には十分注意をして実験を行いますが、万が一実験中に苦痛や不快感などを感じて、体調が悪くなる等の症状が出た場合はすぐに実験責任者にお知らせ下さい。途中で実験を中断しても構いません。尚、実験の中断により実験参加者に何ら不利益を被ることはございません。またこの実験で収集した情報は、この研究に関してのみ使用いたします。

2　実験の流れ

本実験では、提示された各刺激のペアに対して、音像幅の広さと音の自然さについて比較していただきます。各試問では、二つの音 A と B が順番に流され、そのうちにどちらの方がどの程度でより広いかについて判断してください。そして、刺激の音源は楽音の場合に、空間の自然さと音色の自然さについても比較して、どちらの方がどの程度でより自然なのかについて判断してください。表 1 に示した尺度を用いて 7 段階で評価してください。3 つのセッションがあり、1 セッションは 72 問があります。実験時間は 60 分程度です。

- GUI 画面上の「START」を押すと実験が始まり、問題 1 の刺激が流れます。音 A、B が流れ終わったら、尺度を選択して答えてください。「REPLAY」を押すと刺激がもう一度再生される。
- 回答が終われば、「NEXT」を押して、次の問題を回答してください。
- 回答する際に、選択のバーを動かないと「NEXT」ボタンが押せないので、必ずどちらのバーを動いて回答してください。それに、音が流れている際に、「NEXT」ボタンは無効になるので、流れ終わってから押してください。
- 全部の問題が終わると、画面上に本セッション完了を意味するメッセージが出ます。
- セッションの間に、ヘッドホンを外して休憩しても構いません。すぐ次のセッションに入っても構いません。
- 各セッションが始まる前に、お好きな音量をインターフェースによって調整してください。セッション中に音量を変わらないようにお願いします。
- 実験中にいつでも中断や中止をすることができます。

表 1　評価語と尺度

評価語	尺度						
Sound source width	A is strongly wider	A is moderately wider	A is slightly wider	A and B are equal	B is slightly wider	B is moderately wider	B is strongly wider
音像幅	A のほうがとても広い	A の方が広い	A のほうがやや広い	同じくらい	B のほうがやや広い	B の方が広い	B のほうがとても広い
Naturalness of spatial impression/ Naturalness of Timbre	A is strongly more natural	A is moderately more natural	A is slightly more natural	A and B are equal	B is slightly more natural	B is moderately more natural	B is strongly more natural
空間の自然さ/音色の自然さ	A のほうがとても自然	A の方が自然	A のほうがやや自然	同じくらい	B のほうがやや自然	B の方が自然	B のほうがとても自然

Experiment 3 (English translation)

Introduction

Thank you for participating in this experiment. This study aims to investigate the perceived source width of binaural synthesis for headphone reproduction. Please use the mouse to answer questions on the GUI on the computer. The estimated time for the experiment is about 1 hour.

Procedure

In this experiment, please evaluate the perceived source width, the degradation of timbre, and the degradation of the spatial quality respect to the 8 presented stimuli. There are 3 sections for the 3 evaluation items. In each question, there are 8 stimuli of A–H. Please click the corresponding button to switch the replay between stimuli.

For the evaluation of the degradation of timbre and the degradation of the spatial quality, the reference sound will be played automatically when each question start. Please compare the stimuli of A–H with the reference respectively. Please rate the stimulus which is almost the same as the reference 100 points, rate the stimulus which degrades the most 0 points, and rate other stimuli according to the degree of degradation based on the 0 and 100 points stimuli. The examples of the degradation in spatial quality can be the shift of localization, non-uniformity of sound image distribution, and separation of sound image. There are 4 comparisons in one session. For the evaluation of the perceived width, there is no reference, so please rate the stimulus which is widest 100 points, rate the stimulus which is narrowest 0 points, and rate other stimuli according to the perceived width based on the 0 and 100 points stimuli.

- Click the “START” on the GUI to start the experiment. When you finish the question, please click “NEXT” to answer the next question.
- There will be a message indicating the end of the section if all the questions were finished.
- You can take a rest between sections.
- Please adjust gain on the interface to a suitable replay level. Please don’t adjust the level after the experiment begins.

バイノーラルシンセシスにおける音像の広がり効果について

実験機関名：国立大学法人東京藝術大学
実験責任者：蘇恒緯
東京藝術大学　大学院音楽研究科　音楽文化学専攻　音楽音響創造研究分野　博士3年

1　実験概要

この度は実験に参加していただき、誠にありがとうございます。本研究では、バイノーラル聴取における音像の広がり効果について考察します。実験参加者の皆様にはマウスを使用して、パソコン上の指定のアプリケーションでの回答を行って頂きます。実験時間は60分前後を予定しております。実験参加者の心身の安全には十分注意をして実験を行いますが、万が一実験中に苦痛や不快感などを感じて、体調が悪くなる等の症状が出た場合はすぐに実験責任者にお知らせ下さい。途中で実験を中断しても構いません。尚、実験の中断により実験参加者に何ら不利益を被ることはございません。またこの実験で収集した情報は、この研究に関してのみ使用いたします。

2　実験の流れ

本実験では、提示された8つの刺激に対して、音像幅の広さ、音色の劣化、と空間品質の劣化について評価していただきます。それぞれの評価項目で3つのセッションが分かれています。各試問では、A–Hで8つの刺激が提示され、ボタンを押すと刺激再生の切り替えができます。

音色のと空間品質の劣化のセッションにおいて、各試問が始まるとレファレンスボタンが自動に押されてレファレンス音源が流れます。刺激A–Hそれぞれをレファレンス音源と比較し、レファレンスとほぼ変わらない刺激に100点、劣化が一番ひどい刺激に0点をつけ、そしてそれを基づいて他の刺激の劣化程度を評点してください。空間品質の劣化とは、「定位のずれ、幅の空間分布の不均一、音像分離」などで考えてください。各センションは4問があります。

音像幅のセッションにおいてレファレンスボタンはないので、A–Hの刺激の中に、音像が一番広い刺激に100点、一番狭い刺激に0点をつけ、そしてそれを基づいて他の刺激を評点してください。全部8問があります。

- GUI画面上の「START」を押すと実験が始まり、回答が終われば、「NEXT」を押して、次の問題を回答してください。
- 誤操作防止のため、回答する際に、評価用のスライダーを動かないと「NEXT」ボタンが押せません。
- 全部の問題が終わると、画面上に本セッション完了を意味するメッセージが出ます。
- セッションの間に、ヘッドホンを外して休憩しても構いません。すぐ次のセッションに入っても構いません。
- 各セッションが始まる前に、お好きな音量をインターフェースによって調整してください。セッション中に音量を変わらないようにお願いします。
- 実験中にいつでも中断や中止をすることができます。