Original Article

Basic study on improving sound localisation accuracy for musical tones by adding broadband noise

Hidenobu Takao*, Kanagawa Institute of Technology
Ryosuke Katayama, Kanagawa Institute of Technology

Abstract

Purpose: When presenting a virtual auditory display (VAD) to a user who has difficulty using visual information, such as a visually impaired individual, an acoustic head related transfer function (HRTF) is necessary for vertical localisation. However, measurement of the HRTF is time consuming and costly. Because HRTFs differ between individuals, the use of another person's HRTF reduces the accuracy of localisation. In this study, we propose and evaluate an extremely simple method for improving the accuracy of virtual sound localisation on the vertical plane using the same HRTF for various individuals, which may help with the use of a VAD.

Subjects and method: Six male university students (mean age of 22.5±2.5 years) with normal hearing and sight participated in the experiment, during which they wore blindfolds. Sound localisation accuracy was compared among three different conditions: presentation of a tone alone (tone condition), presentation of broadband noise alone (noise condition), and mixed presentation of a tone and broadband noise (mixed condition). Two types of playback system were used: a speaker array and virtual conditions (hearing binaurally through headphones). The stimuli were presented from a total of 13 different directions on the median plane at 10° intervals, from -60° (lower side) to 60° (upper side).

Results: As sound sources were presented farther from 0° in either direction, the target direction offset from 0° was more underestimated. It was found that in the virtual condition, the superimposition of musical tones and broadband noise significantly improved localisation accuracy compared with presenting each alone.

Conclusion: The present study demonstrated that when presenting virtual sound without personalising the HRTF, the extremely simple method of adding broadband noise markedly improves sound localisation accuracy on the vertical plane directly ahead of a subject.

Keywords: Virtual Auditory Display, Visual Impairment, Sound Localisation, 3D Sound, Binaural, Augmented Reality.

Introduction

When operating an information technology device such as a personal computer (PC) when the information on the screen is difficult to utilize, both normally sighted users and users with severe visual impairment may well want to choose and execute commands at will from a command list, such as a pull-down menu on a graphical user interface (GUI). This selection and execution of commands may also be conducted through hearing or touch rather than by sight. At present, the most common of such

Accepted on 8 May, 2016

*: Correspondence Author

methods employing auditory information is the conversion of on-screen content into speech using a software application called a screen reader, which utilizes text-to-speech (TTS) technology, examples of which include NVDA (NV Access, 2016) and PC Talker (Kochi System Development, 2016). However, merely reading aloud the menu content in order is a one-dimensional presentation of information, which lacks the intrinsic spatial perspicuity of a GUI and is inconvenient for users.

One possible means for resolving this issue is a virtual auditory display (VAD), which achieves perspicuity by adding three-dimensional spatial information to speech. One type of proposed VAD is a virtual sound screen, which displays virtual auditory information on a vertical plane in front of the user (Fujisawa, 2003). Another proposed type of VAD is a file manager in which the functional components of a GUI are arrayed on a virtual wall surrounding the user (Frauenberger, 2005). With these VADs, the sounds are displayed at 30° vertical intervals, with up to five items displayed at once; this number is lower than the number of items in a GUI menu on a new PC model.

There are two conceivable reasons for the low number of vertically arrayed items in a VAD. One is that the spatial resolution of the human auditory system is lower in the vertical direction than in the horizontal direction. According to a study conducted in a real acoustic space by Kurosawa et al., the difference limen for sound localisation in the median plane is on average approximately three-times that in the horizontal plane; the difference limen increases along with the angle of incidence, and is at its highest directly overhead (Kurosawa, 1981).

The second reason for the low number of vertically arrayed items in a VAD is the individual differences in the VAD transfer function. For the differentiation in a purely vertical direction, interaural difference information is useless. Therefore, listeners must rely on clues from the sound transfer characteristics of their head, outer ear, shoulders, and other parts of their body, which are called head-related transfer functions (HRTF) (Iida, 2010). Reflecting these HRTFs artificially in sound sources enables the production of virtual sound. However, there are individual differences in HRTFs; thus, when using another person's HRTFs, sound localisation is less accurate than when using one's own HRTFs (Sanada, 2007). Therefore, HRTFs tailored to an individual user would be ideal; this is impractical, however, owing to the special environment and length of time required for measuring the HRTFs. Although recently proposed software would allow the user to select HRTFs resembling their own from an HRTF corpus within a short period of time (Saitou, 2004), this method has not yet been generalized.

Therefore, assuming the use of HRTFs other than the user's own, we investigated the improvement of localisation accuracy by applying auditory contrivances to the presentations of stimuli. In a previous study, we created a three-dimensional auditory menu that opens vertically in front of the user in a manner such as that shown in Figure 1. In addition to speech presentations of menu items, the vertical spatial relationships between items are expressed using a scale of piano tones (Katayama, 2009). An evaluation experiment revealed that, compared to speech presentation alone, a tone-based expression of spatial relationships between items improves the spatial resolution and reduces the difference limen for the menu. In the present study, we focus on the use of spectral cues to further improve the sound localisation accuracy and spatial resolution. These spectral cues, which are a part of the amplitude spectrum of HRTF, are vital for perceiving the directions of the sound sources, particularly in perceiving sounds that are in front of, behind, above, or below the listener. A summation of multiple past studies revealed that spectral cues must include frequency components from 5 to 10 kHz (Iida, 2010).

Therefore, in this study, we propose an extremely simple method for improving the accuracy of sound localisation. When presenting speech language and musical tones from a spatial auditory menu, which is displayed vertically in front of the user, making it difficult to utilize the visual information, we simultaneously present broadband noise with the frequency band required for sound localisation on the median plane. In addition, we describe a basic investigation conducted to determine whether this technique improves the sound localisation accuracy.

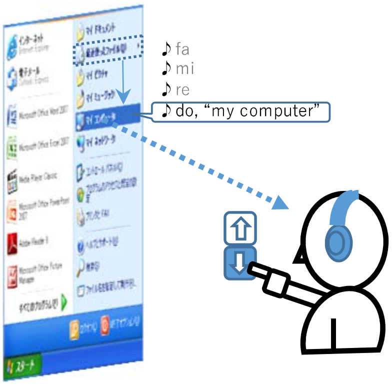

Figure 1: Conceptual rendering of virtual vertical auditory menu

Participants and Methods

1. Participants

The participants of this study were six male university students (mean age of 22.5±2.5 years) with normal hearing and sight. Each of the participants experienced all of the experimental conditions.

2. Experimental conditions

The experimental conditions are shown in Table 1. The sound localisation accuracy was compared among three different conditions: presentation of a tone alone (tone condition), presentation of broadband noise alone (noise condition), and a mixed presentation of a tone and broadband noise (mixed condition). These three types of presented content are referred to herein as the content conditions.

As described in detail later, these stimuli were presented from a total of 13 different directions on the median plane at 10° intervals, from -60° (lower side) to 60° (upper side). Two types of playback system were used: real conditions, in which real sounds were played from a speaker; and virtual conditions, in which virtual sounds were played binaurally from headphones. For the real sound conditions, the sound localisation was measured in a real space to generate reference values. By comparing these measurements to those for the virtual sound conditions, we examined the effects of the use of another person’s HRTFs on the level of accuracy.

Thus, we established a total of 3 × 2 = 6 conditions (RT, RN, RM, VT, VN, and VM).

Table 1: Experimental Design and Abbreviations

Content	Playback System	
	Real	Virtual
Tone	RT	VT
Noise	RN	VN
Mixed (Tone + Noise)	RM	VM

3. Experiment stimuli

The set of experiment stimuli used consisted of five repetitions per trial for the sounds described below, which were played for 1.2 s followed by 0.8 s of silence for each display angle. Therefore, the duration of each trial was

$$(1.2 + 0.8) \times 5 = 10 \text{ s.}$$

Tones

For the tones, the timbre of the sound sources presented was from a piano. The source of the sounds was a MIDI sound module. For the tones, we used a diatonic scale, with each of the 13 tones from D4 to B5 assigned to a different display angle. The sound sources created for the experiment were sampled on a PC at 16 bit/44.1 kHz and saved as monaural wave files.

Table 2: Display angles, pitch names, and centre frequencies for the tone condition

	Lower side							Upper side					
Display angle (degree)	-60	-50	-40	-30	-20	-10	0	10	20	30	40	50	60
Pitch names	D4 (re)	E4 (mi)	F4 (fa)	G4 (so)	A4 (la)	B4 (ti)	C5 (do)	D5 (re)	E5 (mi)	F5 (fa)	G5 (so)	A5 (la)	B5 (ti)
Centre frequency (kHz)	1.2	1.3	1.4	1.6	1.7	1.9	2.1	2.3	2.6	2.8	3.1	3.5	3.9

Broadband noise

The noise used for the broadband noise condition was created through the application of a bandwidth-limiting filter to white noise. The filter was created using audio filter creation software (WS-5510V Realtime Filter Designer). For the high-pass filter, the cutoff frequency was 4.8 kHz, and the damping ratio was −30 dB/oct. For the low-pass filter, the cutoff frequency was 9.6 kHz, and the damping ratio was −60 dB/oct. The broadband noise created for the experiment was sampled on a PC at 16 bit/44.1 kHz and saved as monaural wave files.

The frequency characteristics of the sound sources created for the content conditions are shown in Figures 2, 3, and 4. The frequencies were analysed using WaveSpectra (efu. 2012).

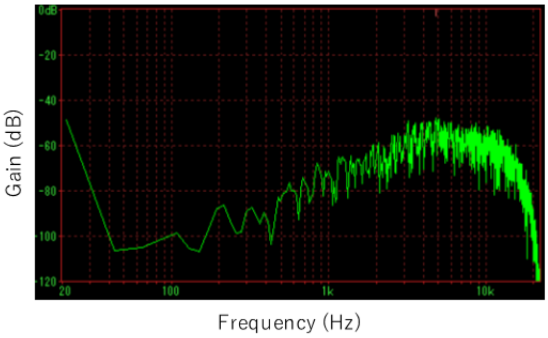

Figure 2: Broadband noise frequency response

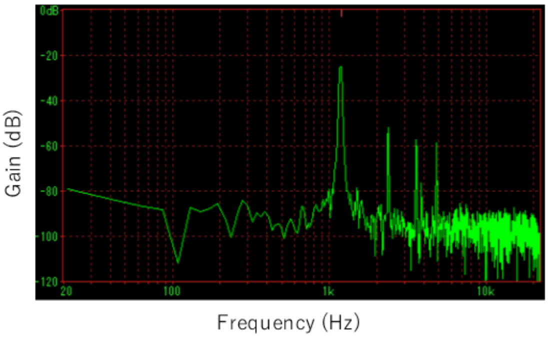

Figure 3: Piano tone D4 (re) frequency response

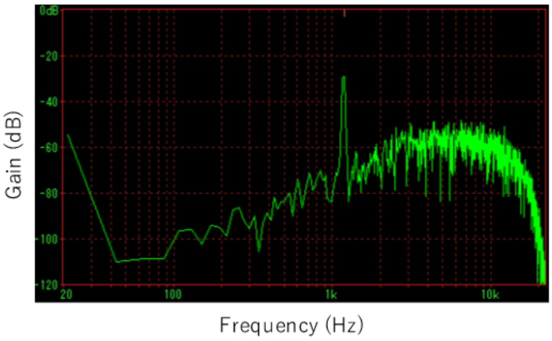

Figure 4: Mixed sound (noise + piano tone D4) frequency response

As an example of a piano tone, we show the frequency response for D4 (re). As an example of a mixed sound, we also show a mix of noise and D4 (re). A look at the piano tone frequency response reveals that the energy is high at the centre frequency and within its harmonic frequencies, and is low within all other bands. A look at the frequency response for the mixed condition reveals that the centre frequency of the piano tone remains intact without being concealed by noise.

Implementing 3D sound

The sound sources for the virtual condition were converted into three-dimensional sound using sound virtualisation technology. The previously described sound and broadband noise files were set to arrive from each direction, as described in Table 1, using CATT-Acoustic v.9.0a for room acoustic prediction and auralisation, and CATT-Walker v.1.1g for a real-time walkthrough auralisation, in virtual sound and space development environments, and virtualized for a binaural display. First, an accurate three-dimensional model of the laboratory was created using a three-dimensional computer-aided design function in the development environment. Next, we established coordinates for the virtual sound sources corresponding to speakers in a real space and for the listening points. We then conducted a sound simulation that incorporated the characteristics of sound reflected from the floor, ceiling, and walls. This simulation was then used to create environment-related transfer functions (ERTFs). At this point, the number of reflections up to the secondary reflection was calculated. The HRTFs used for binauralisation, which were prepared in the development environment in advance, differed from those of the participant. The reverberation duration was set at 0.08 s, identical to the duration in the real lab. These ERTFs were used to convolute the display sound files, which were sampled at 16 bit/44.1 kHz, saved as stereo wav files, and used as the experimental stimuli.

4. Measures

The accuracy of the perceived sound direction was measured and assessed as follows. First, the perceived direction was measured using the following method. Before beginning the experiment, with each participant in a seated position, we measured the height from the ground to the participant’s external auditory meatus. During the experiment, when a stimulus was displayed, the participant perceived the direction from which the sound originated without moving their head. Next, the participant extended their arm in the direction of the sound and pointed in that direction with a pointer held in their hand. Attached to the tip of this pointer was a laser pointer; the experimenter measured the height from the ground to the point where the laser shone (the indication point) using a tape measure and recorded the height. We then calculated the angle formed by the indication point and the height of the participant’s external auditory meatus. This angle was used as the perceived direction.

Next, based on the perceived direction, we assessed the perception accuracy as follows. To assess the perception accuracy for the sound source direction of arrival, we used the perception error in formula (1), which was applied in a study by Ikei et al. (Ikei, 2005). This perception error is the absolute value of the real sound target direction subtracted from the perceived direction. A smaller perception error ε_a signifies a greater level of accuracy in the sound source direction of arrival perception.

$$\varepsilon_a = |\text{ perceived direction} - \text{target direction }| \quad (1)$$

Statistical testing of the perception error was conducted using Excel Ver. 15 (Microsoft).

5. Experimental devices and environment

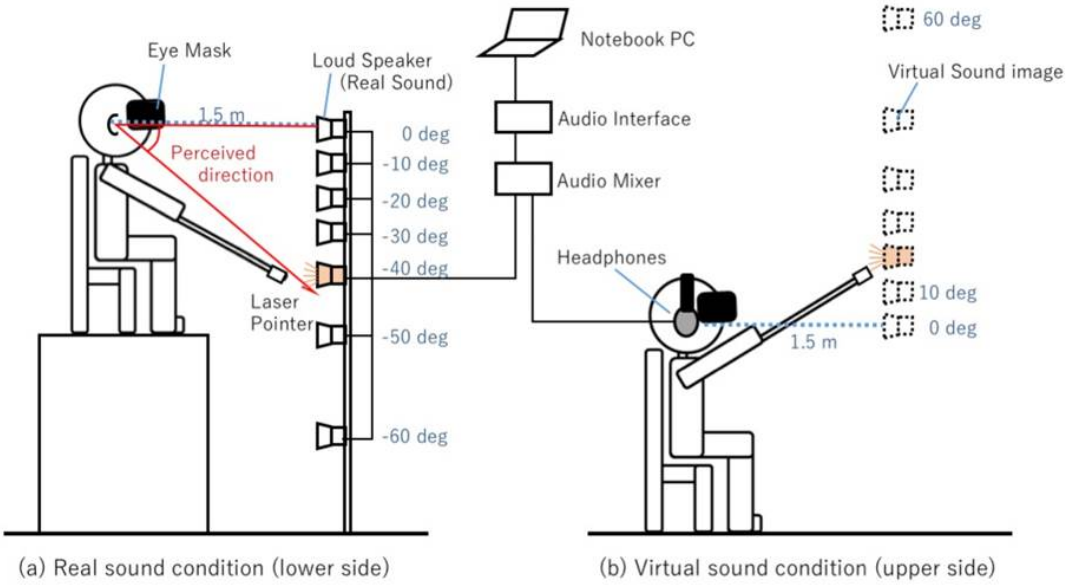

Figure 5: Experimental system design and stimuli presentation array

The experimental system design and stimuli presentation array are shown in Figure 5. Stimulus display software, which was created using Adobe Director MX, and the sound source files of the set of experimental stimuli were stored on a notebook PC (MacBook Air Apple). An audio interface (FireFace UC, RME) was connected to the notebook PC. For the real sound condition, the sound sources were distributed through this audio interface to the speakers (MSP3, Yamaha). For the virtual sound condition, the sound sources were presented binaurally through headphones (PFR-V1, Sony) connected to an audio interface. The distance from the participant's external auditory meatus to the sound source at 0° (at the height of the participant's external auditory meatus) was 1.5 m.

Using the height of the participant's external auditory meatus as a reference (0°), the angles at which the auditory stimuli were presented ranged from 60° to −60°. For the real sound condition, real sounds were presented from 13 speakers arranged within this range at 10° intervals. Owing to the limited height of the ceiling in the lab, the experimental system was divided into an upper side and a lower side. The sound pressure in the participant's ear was adjusted to 65 dB (SPL) when the sound source was presented at the reference height of 0°. The measurements were conducted in the Cognitive and Behavioural Science Research Lab at Kanagawa Institute of Technology. The lab is a soundproof room with a background noise level of <30 dB (A). The temperature and humidity of the room were controlled at 20°C and 50%, respectively. In addition, the reverberation duration was 0.08 s, which did not have a major effect on the sound localisation accuracy.

6. Procedure

The measurements were conducted separately for the upper and lower sides, first for the real sound condition, followed by the virtual sound condition.

6.1 Real sound condition

1) The participant sat down and put on an eye mask.

2) The height of the participant's external auditory meatus was measured; the height of the chair was adjusted so that the height of the participant's external auditory meatus was at 0°, as shown in Figure 5.

3) After the participant was given instructions for the experiment, a practice session was conducted through the presentation of a practice stimulus, and the participant was asked to use a pointer to indicate the sound image direction of arrival. The pointer was held in the participant’s dominant arm.

4) The measurements were started. The sound sources were presented in succession from seven directions, including 0°. Three trials were conducted for each direction; thus, a total of 7 × 3 = 21 trials were conducted in random order, which was treated as one set per condition.

5) After completing one set, the participant rested for 30 min.

6) Steps 1 through 5 were conducted for the remaining side (upper or lower).

7) Steps 1 through 6 were conducted for each of the three content conditions.

6.2 Virtual sound condition

The experiment was begun 24 h after completion of the measurements for the real sound condition.

1) The participant sat down, put on an eye mask, and opened the headphones.

2) Steps 2 through 7 were conducted for the real sound condition.

Results

1. Perceived angles

The mean values for the perceived angles measured for each condition are shown in Figures 6 through 11. Overall, as the sound sources were presented farther from 0° in either direction, the difference with the perceived direction increased, and the target direction was more greatly underestimated. For the VT condition (Figure 9), the differences between the target directions and the perceived directions for both the depression and elevation angles were particularly large, reaching greater than 40°. From about −20° onwards, changes in perceived direction became markedly small, meaning that the differences between angles were not being correctly perceived.

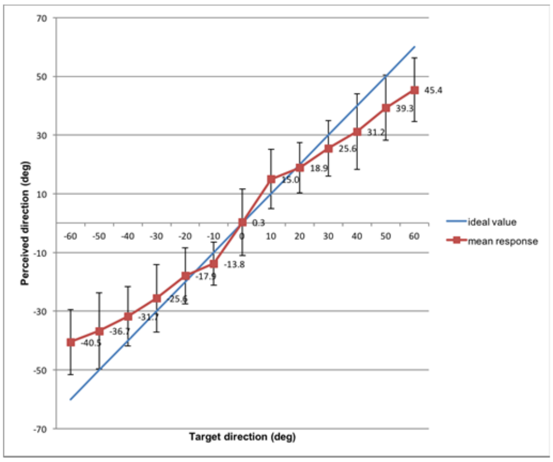

Figure 6: Perceived direction (RT condition)

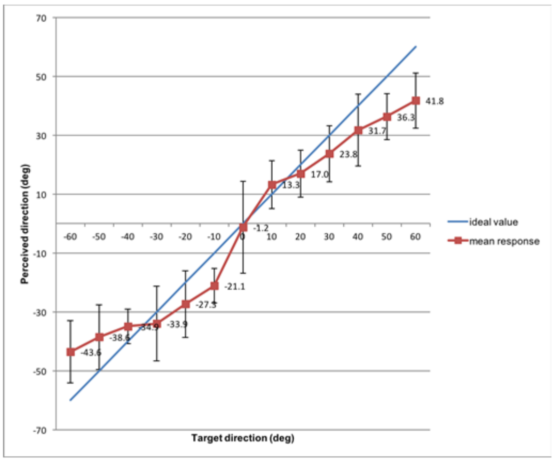

Figure 7: Perceived direction (RN condition)

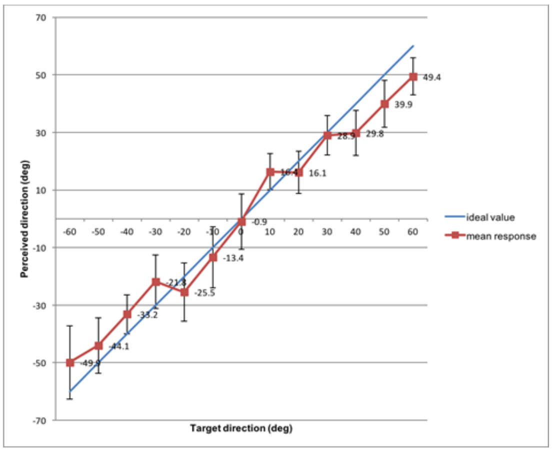

Figure 8: Perceived direction (RM condition)

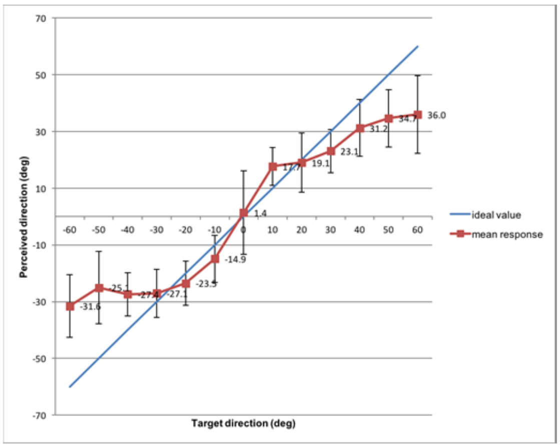

Figure 9: Perceived direction (VT condition)

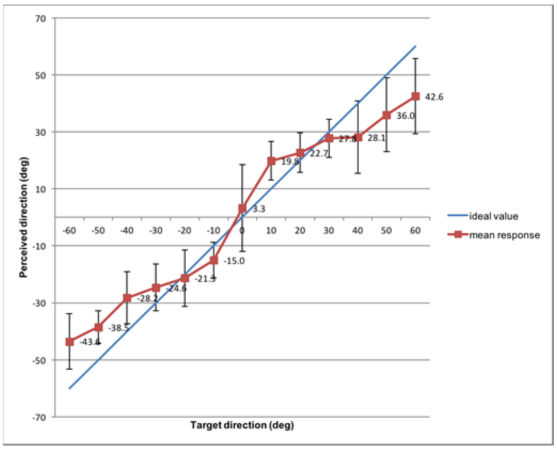

Figure 10: Perceived direction (VN condition)

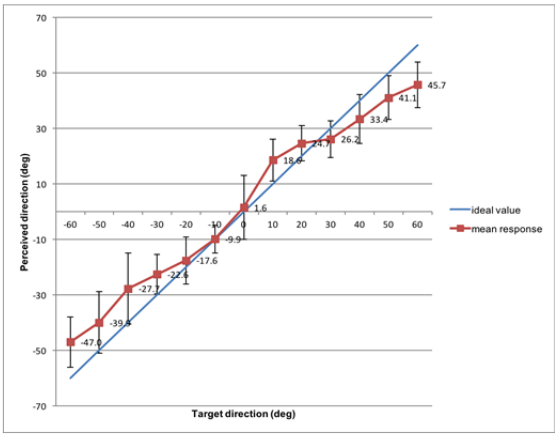

Figure 11: Perceived direction (VM condition)

2. Perception error

Next, based on the perceived directions obtained from the participants, we calculated the perception error and determined the mean values for each condition (Figure 12). First, paired t-tests between the playback conditions (real and virtual sounds) for each of the content conditions revealed a significant difference at the 99% level between the RT and VT conditions ($t(5) = -4.05$, $p = 0.002$) (Figure 12). A significant difference at the 99% level was also observed between the VT and VM conditions ($t(5) = 6.65$, $p = 0.001$). No significant differences were observed between any other pairs of conditions. In summary, for the tone condition, the error was significantly larger for the virtual sound condition than for the real sound condition, whereas for the noise and mixed conditions, the errors for the real and virtual sound conditions were approximately equal.

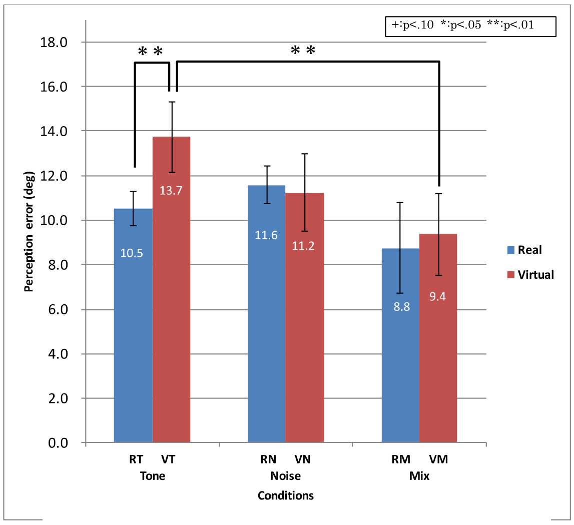

Figure 12: Perception error

Discussion

1. Noise condition results

The equivalent localisation accuracy for the VN and RN conditions can be explained based on the spectral cue theory. For a sound localisation experiment in which Iida et al. analysed multiple spectral peaks and notches in the HTRFs, and used parametric HRTFs recomposed of all or some of these peaks and notches, it was demonstrated that sound can be localized in the median plane by reproducing an HRTF using the lowest frequency notch within the ≥4 kHz frequency range (~6 kHz; N1), the next notch (~9 kHz; N2), and the peak around at 4 kHz (P1) (Iida, 2007). In the present study, although the low-bandwidth cutoff used for noise was 4.8 kHz, the noise was attenuated at a relatively gentle slope of 30 dB/oct. Therefore, at around P1, the noise was attenuated by only about 10 dB. Thus, it is likely that the participants were provided with the judgment information required to detect P1. In addition, within the band between 4.8 and 9.6 kHz, sound was presented without attenuation; therefore, it is likely that sufficient energy was obtained for detecting N1 and N2.

2. Tone condition results

Localisation accuracy for the piano tones was markedly low for the virtual sound condition. This may have arisen from the differences in the HRTFs used for sound virtualisation and the participant's own HRTFs. The correction of perception errors arising from differences in HRTFs requires as many spectral cues as possible to serve as information for making judgments. However, because the range of frequencies of the piano tones was narrow, there was little information for making such judgments, which may have resulted in the large error.

3. Mix condition results

Mixing tones with noise yielded the best accuracy outcomes for the virtual sound condition; these outcomes were equivalent to those for the real sound condition. Considering the large localisation error for the VT condition, this finding is difficult to explain through the spectral cue theory alone. However, the presence of prior knowledge, i.e., whether the sound is familiar, is known to play a role. One study reported that when participants were presented with speech from a familiar voice and speech from an unfamiliar voice, both from speakers in the median plane, the participants were able to perceive the direction of the familiar person's voice more accurately (Blauert, 1986). This result indicates that identifying a sound source may involve not only perception, but also higher-order cognitive function. Supposing that higher-order cognitive function was involved in the results of the present study, the involvement is inferred to be as follows. In the higher-order spatial perception process for deducing the direction of a familiar piano tone based on past experience, the use of spectral cue information, which has abundant broadband noise, as complementary information may have yielded major improvements in accuracy. In conclusion, the present study demonstrated that when presenting virtual sound using another person's HRTFs, the extremely simple method of adding broadband noise to the presented speech markedly improves the sound localisation accuracy.

4. Future issues

For the potential practical application of our results, broadband noise may reduce the degree of comfort. An investigation is necessary to determine how much the acoustic pressure level ratio of tone to broadband noise (S/N) can be increased without a loss in functionality. Shaping the temporal pattern of the sound in some way may also be useful. In addition, while the present study used piano tones, the use of tones from other instruments may produce different interactions. On another topic, in an experiment to assess an auditory pointing system for the visually impaired, which uses a pointer and spatial auditory icons presented through headphones, the perceived direction accuracy was higher when the participants were allowed to turn their heads than when they were not allowed to do so (Hirota, 2003). In our own research, we have already developed a headphone system that corrects for movements of the head (Katayama, 2008), and we intend to implement and assess this system in a future study.

Although the noise condition phenomenon is consistent with an interpretation based on the spectral cue theory, experiments must be conducted using different noise band-pass filter characteristics to determine whether the spectral cue theory is indeed applicable.

Similarly, with regard to the tone and mixed condition, an investigation must be conducted using tones from instruments, such as an organ, which can apply the spectral cue theory.

Acknowledgements: We would like to thank Ai Murakami, Satoru Kobayashi, Takahiro Tsuyuzaki, and Atsuhiro Kubota, along with Jiro Kubo of Acoustic Field, Inc., for their assistance with the performance of the experiment in the present study. The present study was funded with a Grant-in-Aid for Scientific Research (C) [23500661].

References

Blauert J, Morimoto M, Goto T, 1986. Spatial Hearing, Kajima Institute Publishing. 94.

Fraunberger C, Putz V, Holdrich R, Stockman T, 2005. Interaction patterns for auditory userinterfaces. Proc. of ICAD 05-Eleventh meeting, 154-161.

Fujisawa M, Itoh K, Senda T, Yonezawa Y, Hashimoto M, Kaneko H, 2003. Study on HRTFs by the short distance sound source in median plane, and application to auditory display for blind computer users. Technical report of IEICE, EA, 103, 397, 49-54.

Hirota K, Hirose M, 2003. Interaction with wearable systems based on auditory localization. IPSJ Journal, 44, 1, 156-165.

Ikei Y, Yamazaki H, Hirota K and Hirose M, 2005. vCocktail: A wearable-oriented multiplexed voice menu presentation method. Transaction of Human Interface Society, 7, 4, 571-581.

Iida K, Itoh M, Itagaki A and Morimoto M, 2007. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. Applied Acoustics, 68, 835-850.

Iida K, Morimoto M, 2010. Spatial Hearing. Corona Publishing, Co., Ltd., 6-7.

Katayama R, Takao H, Ishii H, 2008. Development of the navigation system for blind people using 3D sound user interface–Building a testbed. Proc. of the 16th Annual Meeting of System-Taikai, Technical Group of Japan Ergonomics Society, AE08-3.

Katayama R, Takao H, Ishii H, 2009. Effect of the tonal scale in the spatial orientation in the 3D sound field. Proc. of the 17th Annual Meeting of System-Taikai, Technical Group of Japan Ergonomics Society, 5-6.

Kochi System Development. PC-Talker, http://www.aok-net.com/screenreader/(accessed on 23 April, 2016)

Kurosawa A, Takagi T, Yamaguchi Z, 1982. On transfer function of human ear and auditory localization. Journal of Acoustical Society of Japan, 38, 3, 145-151.

Matsuo T, Yamaguchi Y, Arama H, 2013. Laterality of cerebral hemisphere damage on auditory space search in stroke patients. Journal of Japanese Occupational Therapy Association, 32, 5, 411-418.

Morimoto M, Saito A, 1977. On sound localization in the median plane–Effects of frequency range and intensity of stimuli. Transactions on Technical Committee of Psychological and Physiological Acoustics, H-40-1-3, 12-17.

NVDA, NV Access. http://www.nvaccess.org/

(accessed on 24 January, 2016)

Saitou K, Iwaya Y, Suzuki Y, 2004. The technique of choosing the individualized head-related transfer function based on localization. IEICE Technical Report, EA, 104, 247, 1-6.

Sanada D, Tamesue T, Itoh K, Hashimoto M, Kayama M, 2007. Arrangement of sound sources for construction of the virtual sound screen using HRTFs. IEICE Technical Report, EA, 107, 269, 49-54.

WaveSpectra. http://efu.jp.net/index.html (accessed on 1 April, 2016)