Proceedings of Meetings on Acoustics

Volume 4, 2008 http://asa.aip.org

155th Meeting
Acoustical Society of America
Paris, France
29 June - 4 July 2008
Session 3pPPa: Psychological and Physiological Acoustics

3pPPa1. Estimation of sound source elevation by extracting the vertical localization cues from binaural signals

Kazuhiro Iida

The author has proposed a parametric HRTF model for vertical sound localization. The parametric HRTF is recomposed only of the spectral peaks and notches extracted from the measured HRTF. The results of the median plane localization tests, which were carried out using the parametric HRTFs with various combinations of spectral peaks and notches, show that the pair of first and second notches (N1 and N2) above 5 kHz can be regarded as spectral cues. Then, utilizing these findings, estimation of the elevation of sound source in the upper median plane by extracting N1 and N2 frequencies from binaural input signal was carried out. The kinds of sound sources were female voice, male voice, music, white noise, and pink noise. The results show that the estimation is accurate for almost of all the elevations and of all the kind of sound sources.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.2980020]
Received 26 Jun 2008; published 20 Aug 2008

Estimation of Sound Source Elevation by Extracting the Vertical Localization Cues from Binaural Signals

Kazuhiro Iida[a]

[a] *Faculty of Engineering, Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino, Chiba 275-0016, Japan*

1. INTRODUCTION

The previous title of the proceedings is "A pair of spectral notches which plays a role as a spectral cue in the vertical localization, and its application to estimation of sound source elevation from binaural signals"

It is generally known that spectral information is a cue for median plane localization. Most previous studies showed that spectral distortions caused by pinnae in the high-frequency range approximately above 5 kHz act as cues for median plane localization [1-11]. Mehrgardt and Mellert [7] have shown that the spectrum changes systematically in the frequency range above 5 kHz as the elevation of a sound source changes. Shaw and Teranishi [2] reported that a spectral notch changes from 6 kHz to 10 kHz when the elevation of a sound source changes from -45 to 45°. Iida *et al.* [11] carried out localization tests and measurements of head-related transfer functions (HRTFs) with the occlusion of the three cavities of pinnae, scapha, fossa, and concha. Then they concluded that spectral cues in median plane localization exist in the high-frequency components above 5 kHz of the transfer function of concha.

Hebrank and Wright [5] carried out experiments with filtered noise and reported the following: spectral cues of median plane localization exist between 4 and 16 kHz; front cues are a 1-octave notch having a lower cut-off frequency between 4 and 8 kHz and increased energy above 13 kHz; an above cue is a 1/4-octave peak between 7 and 9 kHz; a behind cue is a small peak between 10 and 12 kHz with a decrease in energy above and below the peak. Moore *et al.* [12] measured the thresholds of various spectral peaks and notches. They showed that the spectral peaks and notches that Hebrank and Wright regarded as the cues of median plane localization are detectable for listeners, and thresholds for detecting changes in the position of sound sources in the frontal part of the median plane can be accounted for in terms of thresholds for the detection of differences in the center frequency of spectral notches.

Butler and Belendiuk [6] showed that the prominent notch in the frequency response curve moved toward the lower frequencies as the sound source moved from above to below the aural axis in the frontal half of the median plane. Raykar *et al.* [13] noted that one of the prominent features observed in the head-related impulse response (HRIR) and one that has been shown to be important for elevation perception are the deep spectral notches attributed to the pinna. They proposed a method of extracting the frequencies of pinna spectral notches from the measured HRIR, distinguishing them from other confounding features. The extracted notch frequencies are related to the physical dimensions and shape of the pinna.

The results of these previous studies imply that spectral peaks and notches due to the transfer function of concha in the frequency range above 5 kHz prominently contribute to the perception of sound source elevation. However, it has been unclear which component of HRTF plays an important role of as a spectral cue.

This study clarifies the spectral cues for vertical localization by systematic localization tests and careful observations of the characteristics of HRTFs. Then, the findings on the vertical localization cues are applied to the estimation of source elevation by extracting the vertical localization cues from the ear-input signals.

2. CUES FOR VERTICAL LOCALIZATION

The author has proposed a parametric HRTF model to clarify the contribution of each spectral peak and notch as a spectral cue for vertical localization. The parametric HRTF is recomposed only of the spectral peaks and notches extracted from the measured HRTF, and the spectral peaks and notches are expressed parametrically with frequency, level, and sharpness. Localization tests were carried out in the upper median plane using the subjects' own measured HRTFs and the parametric HRTFs with various combinations of spectral peaks and notches [14].

2.1 Parametric HRTFs

As mentioned above, the spectral peaks and notches in the frequency range above 5 kHz prominently contribute to the perception of sound source elevation. Therefore, the spectral peaks and notches are extracted from the measured HRTFs regarding the peaks around 4 kHz, which are independent of sound source elevation [2], as a lower frequency limit. Then, labels are put on the peaks and notches in order of frequency (e.g., P1, P2, N1, N2 and so on). The peaks and notches are expressed parametrically with frequency, level, and sharpness. The amplitude of the parametric HRTF is recomposed of all or some of these spectral peaks and notches.

In order to extract the essential spectral peaks and notches, the microscopic fluctuations of the amplitude spectrum of HRTF were eliminated by Eq. (1):

$$HRTF_w(k) = \sum_{n=-n_1}^{n_1} HRTF(k+n)W(n) \quad , \qquad (1)$$

where *W(n)* is a Gaussian filter defined by Eq. (2). k and n denote discrete frequency. The sampling frequency was 48 kHz, and the duration of HRTFs was 512 samples. In this study, n_1 and σ were set to be 4 and 1.3, respectively.

$$W(n) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-n^2}{2\sigma^2}} \qquad (2)$$

The spectral peak and notch are defined as the maximal and minimal levels of $HRTF_w$, respectively. Thus, the frequencies and the levels of the spectral peaks and notches are obtained. The sharpness of the peak and notch is set to be their envelopment fit with that of $HRTF_w$. Fig.1 shows examples of the parametric HRTFs recomposed of N1 and N2. As shown in the figure, the parametric HRTF reproduces all or some of the spectral peaks and notches accurately and has flat spectrum characteristics in other frequency ranges.

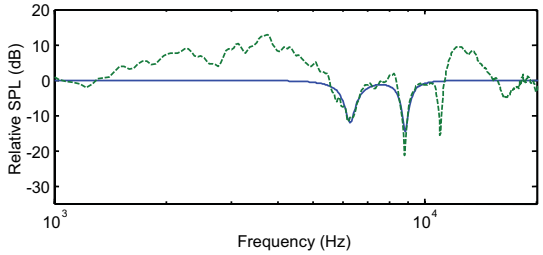


FIGURE 1. An example of parametric HRTF. Dashed line: measured HRTF, solid line: parametric HRTF recomposed of N1 and N2.

2.2 Method of Sound Localization Tests

Localization tests in the upper median plane were carried out using the subjects' own measured HRTFs and the parametric HRTFs. A notebook computer (Panasonic CF-R3), an audio interface (RME Hammerfall DSP), open-air headphones (AKG K1000), and the ear-microphones [14] were used for the localization tests.

The ear-microphones were fabricated using the subject's ear molds (Fig.2). Miniature electret condenser microphones of 5 mm diameter (Panasonic WM64AT102) and silicon resin were put into the ear canals of the ear molds and consolidated (Fig.3). The diaphragms of the microphones were located at the entrances of the ear canals. Therefore, this is so called the "meatus-blocked condition" [2], in other words, the "blocked entrances condition" [15].

The subjects sat at the center of the listening room. The ear-microphones were put into the ear canals of the subject. Then, the subjects wore the open-air headphones (Fig.4), and the stretched-pulse signals were emitted through them. The signals were received by the ear-microphones, and the transfer functions between the open-air headphones and the ear-microphones were obtained. Then, the ear-microphones were removed, and stimuli were delivered through the open-air headphones. Stimuli $P_{l,r}(\omega)$ were created by Eq. (3):

$$P_{l,r}(\omega) = S(\omega) \cdot H_{l,r}(\omega) / C_{l,r}(\omega), \qquad (3)$$

where $S(\omega)$ and $H_{l,r}(\omega)$ denote the source signal and HRTF, respectively. $C_{l,r}(\omega)$ is the transfer function between the open-air headphones and the ear-microphones.

The source signal was a wide-band white noise from 280 Hz to 17 kHz. The measured subjects' own HRTFs and the parametric HRTFs, which were recomposed of all or a part of the spectral peaks and notches, in the upper median plane in 30-degree steps were used. For comparison, stimuli without an HRTF convolution, that is, stimuli with $H_{l,r}(\omega)=1$, were included in the tests.

A stimulus was delivered at 60 dB SPL, triggered by hitting a key of the notebook computer. The duration of the stimulus was 1.2 s, including the rise and fall times of 0.1 s, respectively. A circle and an arrow, which indicated the median and horizontal planes, respectively, were shown on the display of the notebook computer. The subject's task was to plot the perceived elevation on the circle, by clicking a mouse, on the computer display. The subject could hear each stimulus over and over again. However, after he plotted the perceived elevation and moved on to the next stimulus, the subject could not return to the previous stimulus. The order of presentation of stimuli was randomized. The subjects responded ten times for each stimulus.

FIGURE 2. An ear mold of a subject

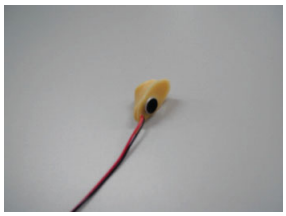


FIGURE 3. An ear-microphone.

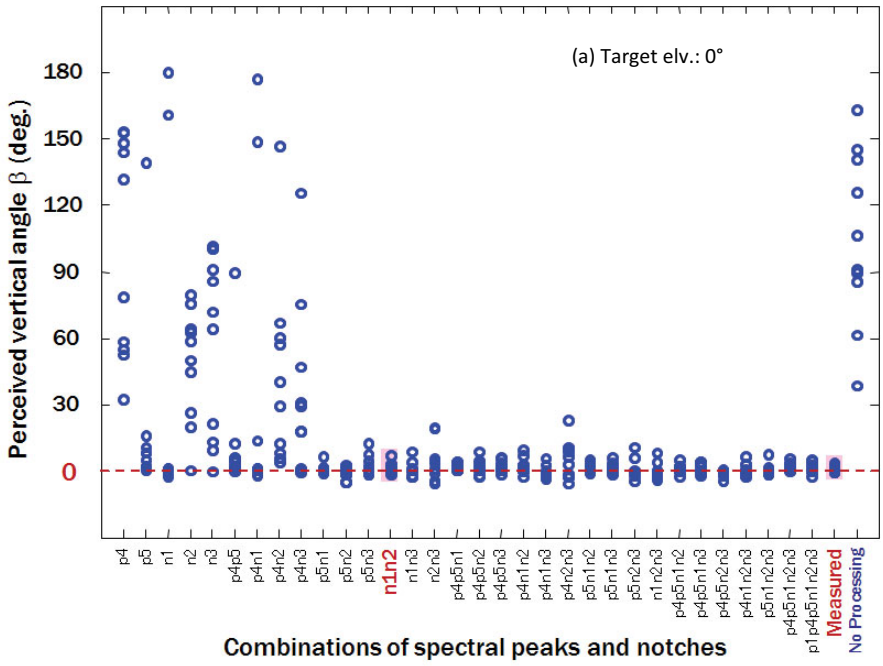
FIGURE 4. A subject wearing ear-microphones and ear-speakers.

2.3 Results of the Tests

Figure 5 shows the distributions of the responses of subject A (a male of 30 years of age) for target elevation of 0, 90, and 180°. The ordinate of each panel represents the perceived elevation, and the abscissa, the kind of stimulus. The 0° is ahead of the listener, and the 180° is behind. Hereafter, the measured HRTF and parametric HRTF are expressed as the mHRTF and pHRTF, respectively.

For the stimuli without an HRTF, the perceived elevation was not accurate, and the variance of responses was large. On the other hand, the subjects perceived the elevation of a sound source accurately at all the target elevations for the mHRTF. For the pHRTF(all), which is the parametric HRTF recomposed of all the spectral peaks and notches, the perceived elevation was as accurate as that for the mHRTF at all the target elevations. In other words, the elevation of a sound source can be perceived correctly when the amplitude spectrum of the HRTF is reproduced

by the spectrum peaks and notches. For the pHRTF recomposed of only one spectral peak or notch, the variances of the responses were large at all the target elevations. One peak or notch did not provide sufficient information for localizing the elevation of a sound source. The accuracy of localization improved as the numbers of peaks and notches increased. Careful observation of the results indicates that the pHRTF recomposed of N1 and N2 provides almost the same accuracy of elevation perception as the mHRTF at most of the target elevations.



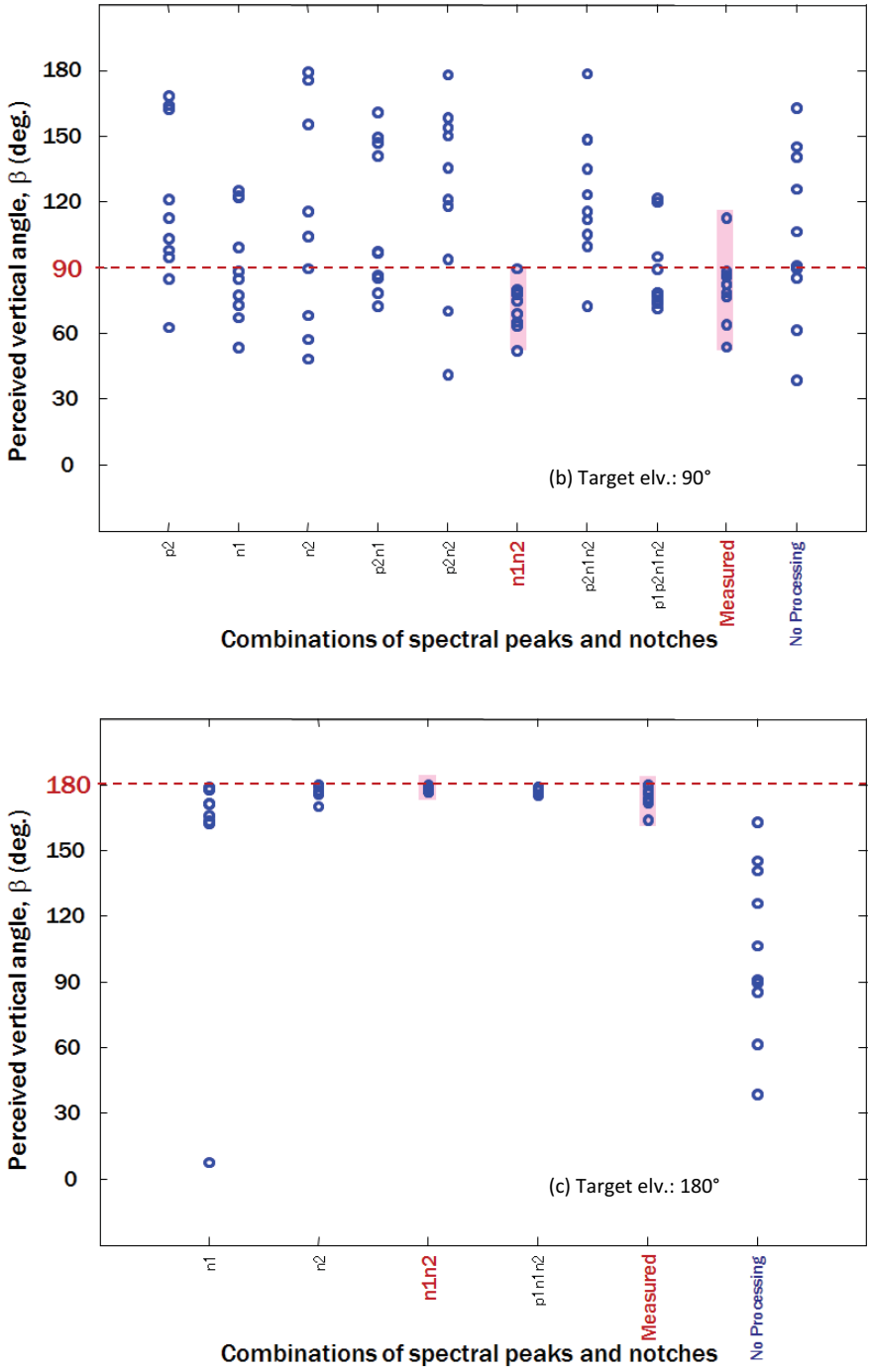


FIGURE 5. Responses to stimuli of measured HRTFs and parametric HRTFs (0, 90, and 180 deg.).

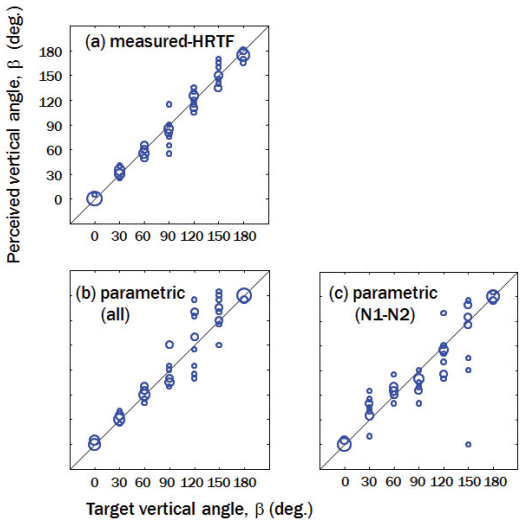


FIGURE 6. Responses to stimuli of measured HRTFs and parametric HRTFs in the median plane (subject A).

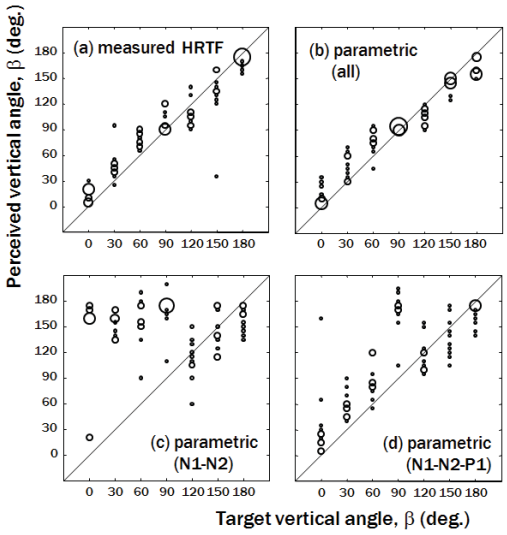


FIGURE 7. Responses to stimuli of measured HRTFs and parametric HRTFs in the median plane (subject B).

2.4 Discussions

The reason why some spectral peaks and notches markedly contribute to the perception of elevation is discussed. Fig.8 shows the distribution of the spectral peaks and notches of the measured HRTFs of subject A in the upper median plane. This figure shows that the frequencies of N1 and N2 change remarkably as the elevation of a sound source changes. Since these changes are non-monotonic, neither only N1 nor only N2 can identify the source elevation uniquely. It seems that the pair of N1 and N2 plays an important role as the vertical localization cues.

The frequency of P1 does not depend on the source elevation. According to Shaw and Teranishi [2], the meatus-blocked response shows a broad primary resonance, which contributes almost 10 dB of gain over the 4-6 kHz band, and the response in this region is controlled by a "depth" resonance of the concha. Therefore, the contribution of P1 to the perception of elevation cannot be explained in the same manner as those of N1 and N2. It could be considered that the hearing system of a human being utilizes P1 as the reference information to analyze N1 and N2 in the ear-input signals.

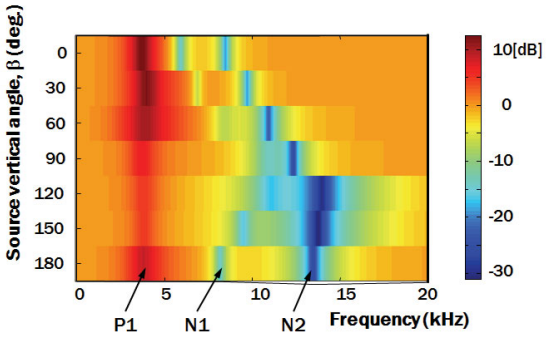


FIGURE 8. Distribution of frequencies of N1, N2, and P1 in the upper median plane.

2.4 Conclusions on the Cues for Vertical Localization

The author carried out sound localization tests using a parametric HRTF model. The results show that (1) perceived elevation for the parametric HRTF recomposed of all the spectral peaks and notches is as accurate as that for the measured HRTF; (2) some spectral peaks and notches play an important role in determining the perceived elevation, whereas some peaks and notches do not; (3) the parametric HRTF recomposed of the first and second notches (N1 and N2) and the first peak (P1) provides almost the same accuracy of elevation perception as the measured HRTFs.

Observations of the spectral peaks and notches of the HRTFs in the upper median plane indicate that (1) the frequencies of N1 and N2 change remarkably as the source elevation changes; (2) whereas, P1 does not depend on the source elevation.

From these results, it is concluded that (1) N1 and N2 can be regarded as spectral cues; (2) the hearing system of a human being could utilize P1 as the reference information to analyze N1 and N2 in ear-input signals.

3. ESTIMATION OF SOURCE ELEVATION BASED ON THE VERTICAL LOCALIZATION CUES

The findings mentioned above imply that the sound source elevation might be estimated with the spectrum of ear-input signals. There are a lot of previous studies on DOA (Direction Of Arrival) estimation, and some of them utilize interaural difference information [16]. In this chapter, the possibility of estimation of source elevation by extracting the vertical localization cues from the ear-input signals is examined.

3.1 Fundamental Strategy for Estimation of Source Elevation

In this study, the following fundamental strategy is adopted;

A) Input signals:

Only the ear-input signals are used. This intends to share the signals with binaural hearing system.

B) Signal processing:

Only the signal processing which is already known as the hearing mechanism is used. The validity and generality as the hearing mechanism are regarded more important than the improvement of estimation accuracy by the nonessential signal processing. The knowledge of the hearing mechanism used in this study is as follows:

- The hearing system utilizes N1 and N2 as cues for vertical localization [14].
- Vertical localization is based on the monaural spectral information. The spectral information is processed in the left and right ear, independently [17].
- The vertical localization mechanism does not require the *a priori* information on the kind of the source signal [18].

3.2 Algorithm

Signal processing is executed as following procedure:

- Transfer the ear-input signals (time domain) to spectrum information (frequency domain)
- Calculate monaural amplitude spectrum envelop
- Detect notch frequencies
- Estimate source elevation comparing the notch frequencies with the N1-N2 database.

The N1-N2 database expresses the frequencies of N1 and N2 as a function of the elevation angle of a sound source. Fig.9 shows the frequencies of N1 and N2 for a sound source in the upper median plane. The frequencies of N1 and N2 change remarkably as the elevation of a sound source changes. However, these changes are non-monotonic. This could be the reason why neither only N1 nor only N2 can identify the source elevation uniquely. These relations are expressed by 4^{th} order polynomial functions. The values of coefficient of determination for N1 and N2 are 0.98 and 0.99, respectively.

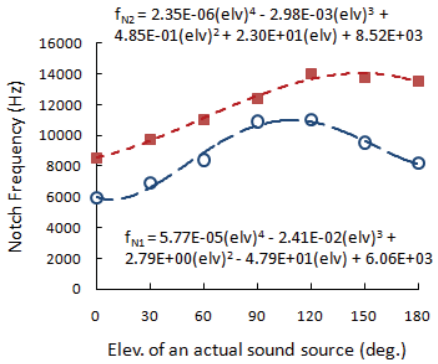


FIGURE 9. Relation between elevation of a sound source and frequencies of N1 and N2. Circles: N1, Squares: N2.

3.3 Simulation I - Estimation of Single Source Elevation under Free Field Condition-

In order to clarify the validity of the estimation method, simulations of estimating a single source elevation were carried out. Following five kinds of source signals were used; white noise, pink noise, male voice, female voice, and pop music. The duration of the signal was 1s. The ear-input signals were obtained by the convolution between the source signals and HRTFs in the upper median plane (0-180°, 30°step). No reflections or reverberations were included. Sampling frequency was 48 kHz.

Figure 10 shows an example of the process for extracting N1 and N2 from the ear-input signals. Fig.10 (a) shows an HRTF at elevation of 0°, (b) is the amplitude spectrum of the female voice with the HRTF convolution, i.e. the ear-input signal, and (c) shows the spectrum envelop and detected N1 and N2. Comparing the detected frequencies with the N1-N2 database, estimated elevation is obtained.

Figure 11 shows estimated elevation for the sound source located in the upper median plane. In general, estimated elevation was accurate regardless of the kind of the sound source. However, front-back estimation errors were observed in the cases of 0° for pop music and 30° for female voice. This error could be related to the fact that the behaviour of N1 and N2 frequencies in the front direction is similar to that in rear. This error is consistent with the human front-back confusion. Estimation accuracy improved remarkably when other 1s-duration parts of these two source signals were used. These interesting results indicate the instability in front-back estimation, which is common behaviour to human sound localization ability.

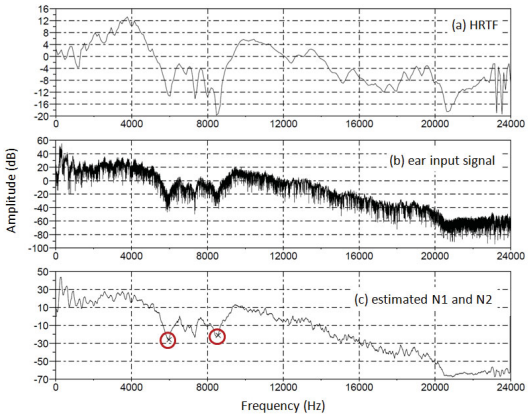


FIGURE 10. An example of detection process of N1 and N2 from ear-input signal. (a): HRTF, (b): ear input signal, (c): detected N1 and N2 (red circles)

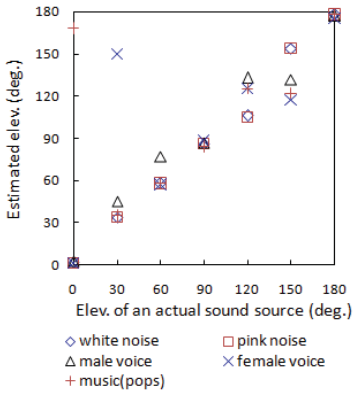


FIGURE 11. Estimated elevation for the sound source located in the upper median plane.

3.4 Simulation II - Estimation under Environmental Noise Condition -

In this section, the effect of non-target sound on the estimation accuracy of the target sound is examined.

Target sounds were white noise, male voice, and female voice, and the duration was 1s. The ear-input signals of the target sound were obtained by the convolution between the source signals and HRTFs in the upper median plane (0-180°, 30°step). No reflections or reverberations were included.

Non-target sound was an environmental noise recorded at a concourse of a station by 6-channel recording system [19]. Ear-input signals of the non-target sound were recorded with ear-microphones, which were located at the entrance of the ear canals of a listener, reproducing the 6ch-recorded signals through 6 loudspeakers in an anechoic chamber. The ear-input signals of the target sound and those of non-target signals were mixed in the time domain with S/N of 0, 10, 20, 30, and infinite dB.

Results of simulation are shown in Fig.12. For the target sound of white noise, the estimated elevation was as accurate as that for without non-target sound, in the case of 10dB < S/N. In the case S/N of 0dB, estimation accuracy is reduced. For the target sound of male and female voice, the estimated elevation was as accurate as that for without non-target sound, in the case of 15dB < S/N, except 0°. To obtain the same accuracy as without non-target sound, S/N of 30dB is required for 0°.

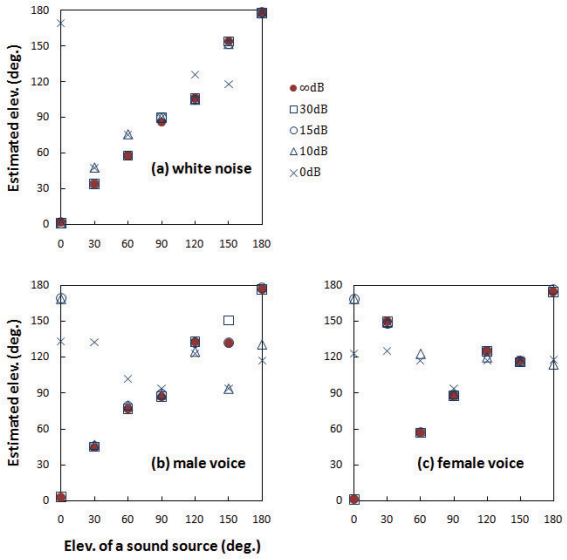


FIGURE 12. Estimated elevation for the sound source located in the median plane under environmental noise conditions. (a): white noise, (b): male voice, (c): female voice.

3.5 Discussions

Validity of the simulation results should be confirmed by comparing them with the human localization ability under noise condition. Good and Gilkey [20] reported on the effect of non-target sound, which was a wide-band noise provided from the front direction, on the localization accuracy of the target sound, which was a pulse train provided from the azimuth of 0 - 360°, and elevation of -45 - +90°. Their results show that the effect of non-target sound is less on the left-right perception, but much on the front-back perception. These results are consistent with those of this study qualitatively. It is hard, however, to make a quantitative comparison between them, because their experimental conditions are differ from this study. Sound localization tests under the same noise condition as the simulation should be done to clarify the validity of the estimation method.

3.6 Conclusions

The possibility of estimation of source elevation by extracting the vertical localization cues (N1 and N2) from the ear-input signals is examined. Simulations of estimating a source elevation in the median plane show that estimated elevation is accurate regardless of the kind of the sound source, in general. Then, the effect of non-target sound on the estimation accuracy of the target sound is examined. For the target sound of white noise, the estimated elevation was as accurate as that for without non-target sound, in the case of 10dB < S/N. For the target sound of male and female voice, the estimated elevation was as accurate as that for without non-target sound, in the case of 15dB < S/N, except 0°.

ACKNOWLEDGMENTS

The author wishes to thank Professor Masayuki Morimoto for meaningful discussion. The author thanks also to Dr. Motokuni Itoh and Atsue Itagaki for their cooperation in localization tests, and to Dr. Sakae Yokoyama for her help in making the environmental noise for simulations of source elevation estimation.

REFERENCES

1. K. Roffler, A. Butler, "Factors that influence the localization of sound in the vertical plane", *J. Acoust. Soc. Am.* 43, 1255-1259 (1968).
2. E. A. G. Shaw, R. Teranishi, "Sound pressure generated in an external-ear replica and real human ears by a nearby point source", *J. Acoust. Soc. Am.* 44, 240-249 (1968).
3. J. Blauert, "Sound localization in the median plane", *Acustica* 22, 205-213 (1969/70).
4. B. Gardner, S. Gardner, "Problem of localization in the median plane: effect of pinna cavity occlusion", *J. Acoust. Soc. Am.* 53, 400-408 (1973).
5. J. Hebrank, D. Wright, "Spectral cues used in the localization of sound sources on the median plane", *J. Acoust. Soc. Am.* 56, 1829-1834 (1974).
6. A. Butler, K. Belendiuk, "Spectral cues utilizes in the localization of sound in the median sagittal plane", *J. Acoust. Soc. Am.* 61, 1264-1269 (1977).
7. S. Mehrgardt, V. Mellert, "Transformation character-istics of the external human ear", *J. Acoust. Soc. Am.* 61, 1567-1576 (1977).
8. A. J. Watkins, "Psychoacoustic aspects of synthesized vertical locale cues", *J. Acoust. Soc. Am.* 63, 1152–1165 (1978).
9. M. Morimoto, H. Aokata, "Localization cues of sound sources in the upper hemisphere", *J. Acoust. Soc. Jpn (E)*. 5, 165-173 (1984).
10. J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics", *J. Acoust. Soc. Am.* 92, 2607-2624 (1992).
11. K. Iida, M. Yairi, M. Morimoto, "Role of pinna cavities in median plane localization", *Proc. 16th Int'l Cong. on Acoust*. 1998; 845-846 (1998).
12. B. C. J. Moore, R. Oldfield, G. J. Dooley, "Detection and discrimination of peaks and notches at 1 and 8 kHz", *J. Acoust. Soc. Am.* 85, 820-836 (1989).
13. V. C. Raykar, R. Duraiswami, B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses", *J. Acoust. Soc. Am.* 118, 364-374 (2005).
14. K. Iida, M. Itoh, A. Itagaki, M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues", *Applied Acoustics*, 68, 835-850 (2007).
15. D. Hammershøi, H. Møller, "Sound transmission to and within the human ear canal", *J. Acoust. Soc. Am.* 100, 408-427 (1996).
16. H. Nakashima, Y. Chisaki, T. Usagawa, M. Ebata, "Frequency domain binaural model based on interaural phase and level differences", *Acoust. Sci. & Tech.* 24, 172-178 (2003).
17. K. Iida, M. Itoh, E. Rin, M. Morimoto, " Extraction process of spectral cues from input signals to two ears in median plane localization", *Proc. 17th Int'l Cong. on Acoust*. (2001)
18. K. Iida, M. Morimoto, "*A priori* knowledge of the sound source spectrum in median plane localization", *J. Acoust. Soc. Am.* 105, 1391 (1999).
19. S. Yokoyama, K. Ueno, S. Sakamoto, H. Tachibana, "6-channel recording/reproduction system for 3-dimensional auralization of sound fields", *Acoust. Sci. & Tech.* 23, 97-103 (2002).
20. M. D. Good, R. H. Gilkey, "Sound localization in noise: The effect of signal-to-noise ratio", *J. Acoust. Soc. Am.* 99, 1108-1116 (1996).