

ACOUSTICAL LETTER

Model for estimating elevation of sound source in the median plane from ear-input signals

Kazuhiro Iida*

Faculty of Engineering, Chiba Institute of Technology,
2–17–1 Tsudanuma, Narashino, 275–0016 Japan

(*Received 9 July 2009, Accepted for publication 26 August 2009*)

Keywords: Elevation estimation, The median plane, Ear-input signal, Spectral cue, Head-related transfer function
PACS number: 43.66.Pn, 43.66.Qp, 43.60.Jn [doi:10.1250/ast.31.191]

1. Introduction

Several studies have shown that the azimuth of a sound source can be estimated by using interaural difference information extracted from signals obtained at both ears [1,2]. The elevation, however, could not be estimated from the interaural difference information because the interaural difference information includes only lateral information and lacks both front-back information and up-down information.

A number of studies have revealed that spectral distortions caused by pinnae in the high-frequency range above approximately 5 kHz act as cues for elevation perception [3–7]. Hebrank and Wright [3] carried out experiments with filtered noise and reported that spectral cues of median plane localization exist between 4 and 16 kHz, front cues are a one-octave notch having a lower cutoff frequency between 4 and 8 kHz and increased energy above 13 kHz, an overhead cue is a 1/4-octave peak between 7 and 9 kHz, and 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.* [4] measured the thresholds of various spectral peaks and notches. They showed that the spectral peaks and notches that Hebrank and Wright regarded as cues of median plane localization are detectable by listeners and that thresholds for detecting changes in the position of a sound source 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 [5] showed that the prominent notch in the frequency response curve moved toward 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.* [6] noted that deep spectral notches attributed to the pinna, which are prominent features observed in the head-related impulse response (HRIR), are important for elevation perception. 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.

Iida *et al.* [7] carried out localization tests in the median plane using a parametric head-related transfer function (HRTF), which is recomposed of all or some of the spectral peaks and notches extracted from the measured HRTF. Their results revealed that the parametric HRTF recomposed of the first and second notches (N1 and N2) and the first peak (P1), in order of frequency above 4 kHz, provides almost the same localization accuracy as the measured HRTFs. Observations of the spectral peaks and notches indicate that the frequencies of N1 and N2 change markedly with changes in the source elevation, whereas the frequency of P1 is independent of the source elevation. Thus, Iida *et al.* concluded that N1 and N2 could be regarded as spectral cues and that the human auditory system could use P1 as a reference for analyzing N1 and N2.

These findings imply that the elevation of a sound source might be estimated from the frequencies of N1 and N2 extracted from ear-input signals. In the present study, a model for estimating the elevation of a sound source in the median plane is proposed, and its validity is investigated.

2. Estimation model of sound source elevation

The proposed estimation model utilizes the following fundamental knowledge of the spatial hearing mechanism.

a) The human auditory system uses N1 and N2 as cues for the perception of elevation [7].

b) The perception of elevation is based on monaural spectral information. The spectral information is processed in the left and right ears independently [8].

c) The vertical localization mechanism does not use *a priori* information concerning the kind of sound source [9]. The perceived elevation depends only on the spectrum of the ear-input signal, regardless of the kind of sound source.

On the basis of the knowledge described above, the proposed model is used to estimate the elevation of a sound source in the median plane as follows.

1) Transform the input signal to the ear ipsilateral to the sound source in the time domain to spectral information by taking the Fourier transformation.

2) Obtain the envelope of the amplitude spectrum of each ear-input signal in order to eliminate the microscopic fluctuations by the moving average.

3) Extract all local minima of the amplitude spectrum envelope of the input signals to the ear above 4 kHz, and set the local minima as the candidates of N1 and N2.

4) Calculate the evaluation function, $E(\theta)$, for each candidate of N1 and N2 as follows.

*e-mail: kazuhiro.iida@it-chiba.ac.jp

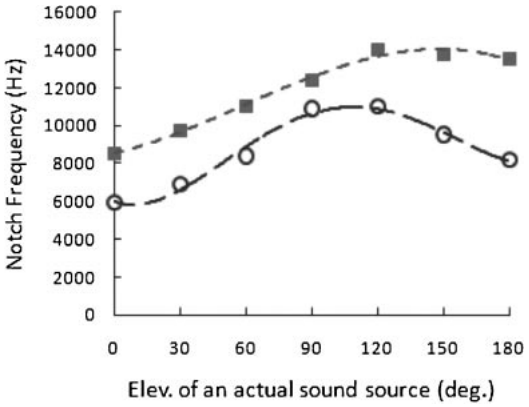

Fig. 1 Relationship between the elevation of a sound source and the frequencies of N1 and N2. The open circles and closed squares denote measured values of N1 and N2, respectively.

$$E(\theta) = \{|fN1c_i - fN1(\theta)| + |fN2c_j - fN2(\theta)|\}, \quad (1)$$
$$i = 1, N, \quad j = 1, N, \quad i < j,$$

where θ is the sound source elevation in degrees, $fN1c_i$ and $fN2c_j$ are the frequencies of the i-th N1 candidate and the j-th N2 candidate, respectively, $fN1(\theta)$ and $fN2(\theta)$ are the frequencies of N1 and N2 for a sound source at an elevation of θ obtained from the N1 and N2 databases, respectively, and N denotes the number of N1 and N2 candidates. Then, in order to obtain the estimated elevation, $\hat{\theta}$, use the value of θ when $E(\theta)$ is a minimum.

Figure 1 shows an example of an N1-N2 database, i.e., the frequencies of N1 and N2 at the left ear of the author's HRTFs for a sound source in the upper median plane. These frequencies change markedly as the elevation of the sound source changes. The relations between elevation and notch frequencies are expressed by 4th-order polynomial functions. Multiple regression equations are expressed by the following equations:

$$fN1(\theta) = 5.77 \times 10^{-5} \times \theta^4 - 2.41 \times 10^{-2} \times \theta^3 + 2.79 \times \theta^2 - 4.79 \times 10 \times \theta + 6.06 \times 10^3 \quad (2)$$

$$fN2(\theta) = 2.35 \times 10^{-6} \times \theta^4 - 2.98 \times 10^{-3} \times \theta^3 + 4.85 \times 10^{-1} \times \theta^2 + 2.30 \times 10 \times \theta + 8.52 \times 10^3, \quad (3)$$

where θ is the elevation of the sound source in degrees. The values of the coefficients of determination for N1 and N2 are 0.98 and 0.99, respectively.

3. Simulation of estimation of sound source elevation

3.1. Estimation of elevation of single sound source under free field condition

In order to clarify the validity of the estimation model described above, simulations of estimating a single sound source elevation were carried out. The following five kinds of source signals were used: white noise, pink noise, a male voice, a female voice, and popular music. The duration of each signal was 1 s. The input signals to the left ear were obtained by convolution between the source signals and the author's HRTFs in the upper median plane (0–180°, 30° steps). It is important to note that these HRTFs are the same as those used in the N1-N2 database. No reflections were included, and the sampling frequency was 48 kHz.

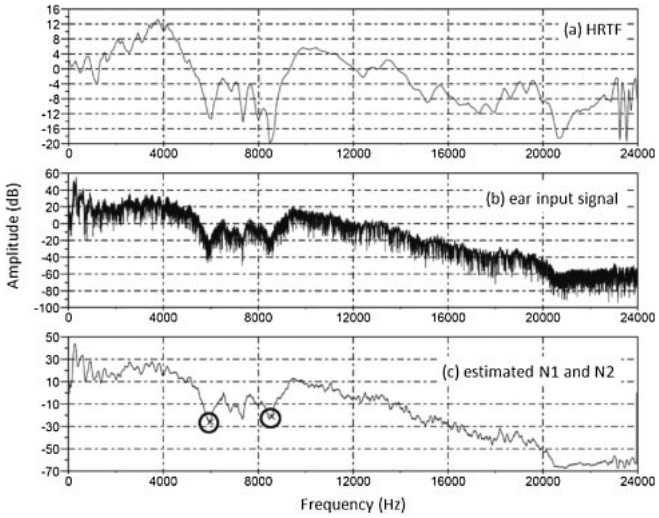

Fig. 2 Example of detection process of N1 and N2 from ear-input signal: (a) HRTF for elevation of 0°, (b) ear input signal, (c) amplitude spectrum envelope of input signal to left ear. Circles indicate detected N1 and N2.

Figure 2 shows an example of the process of N1 and N2 extraction from the ear-input signals. Figure 2(a) shows the author's HRTF for an elevation of 0°. Figure 2(b) is the amplitude spectrum of the female voice with the convolution of the HRTF, i.e., the ear-input signal, $|p|$. Figure 2(c) shows the amplitude spectrum envelope, $|p|_{\mathrm{env}}$, of the input signal to the left ear using the moving average, as shown in

$$|p(i+M-1)|_{\mathrm{env}} = \frac{1}{M}\sum_{j=1}^{M}|p(i+j)|, \quad i = 1, (L-M), \quad (4)$$

where M is the number of samples used for the moving average, and L is the number of samples used for estimation. In the present study, M and L were set to 50 and 48,000, respectively. The circles in Fig. 2(c) indicate the detected N1 and N2 at which $E(\theta)$ reaches minimum.

Figure 3 shows the estimated elevation for a sound source located in the upper median plane. In general, estimation was accurate regardless of the kind of sound source. However, front-back estimation errors were observed in the cases of 0° for popular music and 30° for the female voice. This error could be related to the fact that the behavior of N1 and N2 frequencies in the front direction is similar to that in the rear direction. Furthermore, this front-back error shows a similar tendency to human front-back confusion. The estimated direction was behind (around 170°) for some 1-s-long parts of popular music located at an elevation of 0° in the median plane, and front (around 0°) for other 1-s-long parts (Fig. 4). These results indicate instability in the front-back estimation, which is a well-known behavior in human sound localization [10].

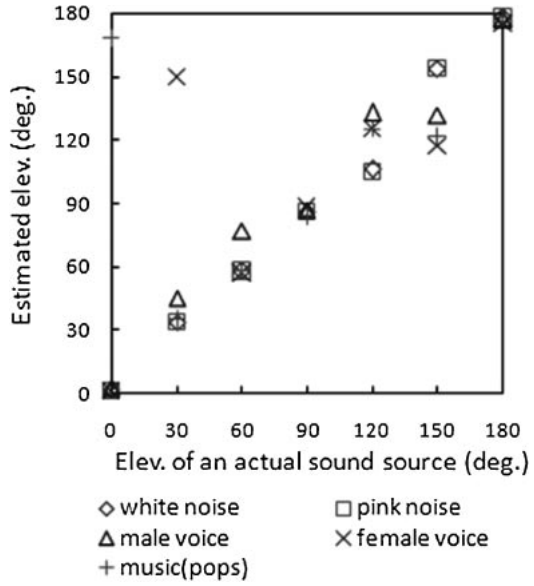

Fig. 3 Estimated elevation for sound source located in the upper median plane.

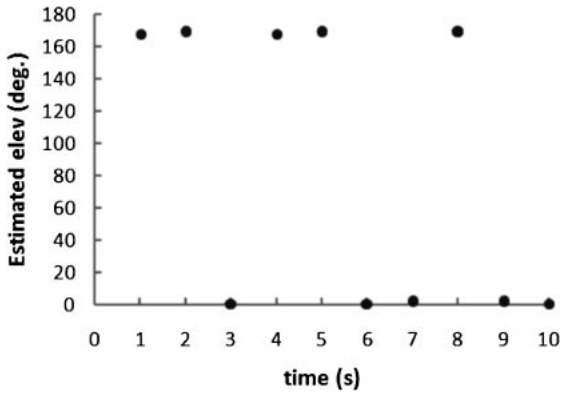

Fig. 4 Estimated elevation for various 1-s-long-parts of popular music located at elevation of 0° in the median plane.

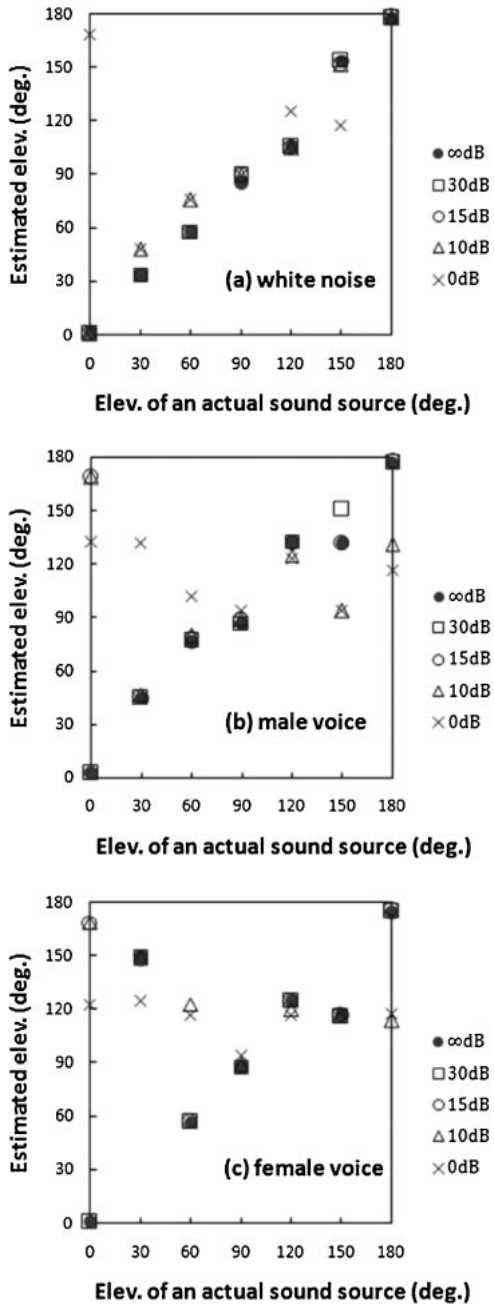

Fig. 5 Estimated elevation for sound source located in the median plane under environmental noise conditions: (a) white noise, (b) male voice, (c) female voice.

3.2. Estimation under environmental noise condition

The effect of a nontarget sound on the estimation accuracy of the target sound was examined.

The target sounds and the ear-input signals are the same as those described in Section 3.1. The nontarget sound was environmental noise recorded in the concourse of a railway station using a 6-channel recording system [11]. The ear-input signals of the nontarget sound were recorded using ear microphones, which were located at the entrance of the ear canals of the subject. Their HRTFs were used for the N1-N2 database, and the 6-channel recorded signals were reproduced through six loudspeakers in an anechoic chamber. The ear-input signals of the target sound and those of nontarget signals were mixed in the time domain with signal-to-noise ratios of 0, 10, 20, 30, and infinite dB.

The simulation results are shown in Fig. 5. For white noise as the target sound, the estimated elevation was as accurate as that estimated in the absence of the nontarget sound, for the case in which 10 dB < S/N. For the case in which S/N was 0 dB, the estimation accuracy was reduced. For male and female voices as the target sounds, the estimated elevation was as accurate as that estimated in the absence of the nontarget sound, for the case in which 15 dB < S/N, except at a target elevation of 0°. In order to obtain the same accuracy as that estimated in the absence of the nontarget sound, a S/N of 30 dB is required at a target elevation of 0°.

3.3. Discussion

The validity of the proposed model is examined by comparing the simulation results with the human localization ability under the noise condition. Good and Gilkey [12]

investigated the effect of nontarget sound, which was a wide-band noise presented from the front direction, on the localization accuracy of the target sound, which was a pulse train presented from the azimuth of 0 to 360° at an elevation of −45 to +90°. Their results revealed that the effect of the nontarget sound on left-right perception is slight, whereas the effect on front-back perception is significant. These tendencies are similar to those observed in the present study. However, it is difficult to compare the results of these two studies quantitatively, because the experimental conditions were different. Sound localization tests conducted under the same noise conditions as adopted in the simulation are required in order to clarify the validity of the estimation model.

4. Conclusions

The estimation of the elevation of a sound source in the median plane by extracting the elevation perception cues (N1 and N2) from the ear-input signals was investigated. The results of the simulation revealed that the estimated elevation is accurate regardless of the kind of sound source. The effect of nontarget sound on the estimation accuracy of the target sound was then investigated. For white noise as target sound, the estimated elevation was as accurate as that estimated in the absence of the nontarget sound when 10 dB < S/N. For the target sounds of male and female voices, the estimated elevation was as accurate as that estimated in the absence of the nontarget sound when 15 dB < S/N, except at a target elevation of 0°.

Acknowledgments

The authors would like to thank Dr. Sakae Yokoyama for her help in producing the environmental noise used in the simulation of source elevation estimation. This study was supported by "Academic Frontier" Project for Private Universities: matching fund subsidy from MEXT (Ministry of Education, Culture, Sports, Science and Technology).

References

[1] J. Blauert and W. Cobben, "Some consideration of binaural cross correlation analysis," *Acustica*, **39**, 96–104 (1978).
[2] H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, "Frequency domain binaural model based on interaural phase and level differences," *Acoust. Sci. & Tech.*, **24**, 172–178 (2003).
[3] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, **56**, 1829–1834 (1974).
[4] B. C. J. Moore, R. Oldfield and G. J. Dooley, "Detection and discrimination of peaks and notches at 1 and 8 kHz," *J. Acoust. Soc. Am.*, **85**, 820–836 (1989).
[5] A. Butler and K. Belendiuk, "Spectral cues utilizes in the localization of sound in the median sagittal plane," *J. Acoust. Soc. Am.*, **61**, 1264–1269 (1977).
[6] V. C. Raykar, R. Duraiswami and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am.*, **118**, 364–374 (2005).
[7] K. Iida, M. Itoh, A. Itagaki and M. Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Appl. Acoust.*, **68**, 835–850 (2007).
[8] K. Iida, M. Itoh, E. Rin and M. Morimoto, "Extraction process of spectral cues from input signals to two ears in median plane localization," *Proc. 17th Int. Congr. Acoust.* (2001).
[9] K. Iida and M. Morimoto, "*A priori* knowledge of the sound source spectrum in median plane localization," *J. Acoust. Soc. Am.*, **105**, 1391 (1999).
[10] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, **105**, 2841–2853 (1999).
[11] S. Yokoyama, K. Ueno, S. Sakamoto and H. Tachibana, "6-channel recording/reproduction system for 3-dimensional auralization of sound fields," *Acoust. Sci. & Tech.*, **23**, 97–103 (2002).
[12] M. D. Good and R. H. Gilkey, "Sound localization in noise: The effect of signal-to-noise ratio," *J. Acoust. Soc. Am.*, **99**, 1108–1116 (1996).