A SUBBAND ENERGY MODIFICATION METHOD FOR ELEVATION CONTROL IN MEDIAN PLANE

Dingding Yao[1,2]*, Junfeng Li*[1,2]*, Huaxing Xu*[3]*, Risheng Xia*[1] *and Yonghong Yan*[1,2]

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]School of Electrical Engineering, Zhengzhou University, Zhengzhou, China

ABSTRACT

Elevation perception is crucial for binaural reproduction. A recent study proposed an elevation control method by modifying the energy of HRTFs in each auditory scale subband, such as the ERB and Mel subband. However, this subband division is designed based on auditory excitation patterns and may not be consistent with the elevation localization cues. To this end, this study proposes a novel subband division strategy which emphasizes the physiological information involved in elevation localization based on a statistical analysis of the HRTF. Then, the elevation controlled HRTFs are constructed by modifying the energy of the HRTF magnitudes in each subband. Results of the listening test demonstrate that our method with the proposed subband division strategy outperforms the method with ERB scale subdivision in terms of the accuracy for controlling the perceived elevation of sound image.

Index Terms— Elevation perception, Head-related transfer function, Spectral cues, Elevation control, Binaural synthesis

1. INTRODUCTION

It is well established that the perceptual cues for sound localization in the vertical direction are primarily related to spectral cues caused by sound scattering around the listener's head, pinna, torso and shoulders [1, 2]. All these perceptual cues can be typically represented by a complex frequency response function, known as head-related transfer function (HRTF).

Over the last few decades, many studies have investigated the role of spectral cues for vertical localization in the median plane. By smoothing the HRTF magnitude spectra with different degrees of simplification and different boundary frequencies, Asano *et al.* found that the information in macroscopic pattern of HRTF is important for vertical localization and the major elevation perception cues exist in the frequency region above 5 kHz [3]. Talagala *et al.* presented an effective method to extract the HRTF spectral cues using cepstral processing for binaural source localization in the median plane [4]. In parallel, the importance of spectral peaks and notches has also been studied [5, 6, 7, 8, 9]. Iida *et al.* showed that a parametric HRTF recomposed of the first two notches (N_1 and N_2) and the first two peaks (P_1 and P_2) yields almost the same localization accuracy as the measured HRTF [6, 8]. Moreover, Langendijk *et al.* examined the importance of various frequency bands to different elevation perception and found that the most important cue discriminating up-down elevations is in the middle 1-octave band (i.e., $5.7 - 11.3$ kHz) [10]. Talagala *et al.* introduced a method for extracting the directional information in the subbands of a broadband signal and observed that the diversity in the frequency domain played an important role in the localization of a source in the vertical plane [11].

Considering the important contribution of spectral cues of HRTFs to elevation perception, several approaches have been reported to control the perceived elevation of sound images [12, 13, 14, 15]. For instance, Tan *et al.* divided frequency spectrum into five regions and manipulated the spectrum of HRTFs by boosting or attenuating levels of corresponding band to control the elevation of reproduced sound [13]. By parametrically modeling and modifying the spectral features (spectral peaks and notches) of the direct component in BRIRs, Yao *et al.* proposed an elevation control approach for binaural reproduction [15].

Recently, Karapetyan *et al.* adopted a 24 channels Mel filter bank for analyzing the early reflections of elevated BRIRs and adjusted the energy in each subband for elevation control [16]. Their study on elevation control have demonsrated the feasibility of adjusting auditory subband energies as a simple and easy method for controlling the perceived elevation of sound. However, Mel filter banks are designed based on auditory excitation patterns while HRTF represents the physical properties of anthropometric parameters and encodes the cues needed in spatial localization. Therefore, a better fre-

This work is partially supported by the National Key Research and Development Program (Nos. 2017YFB1002803, 2016YFB0801203, 2016YFC0800503), the National Natural Science Foundation of China (Nos. 11804309, 11590770-4, 11722437, 61650202, U1536117, 61671442, 11674352, 11504406, 61601453) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 2016A03007-1).

quency division which emphasizes physiological information of HRTFs along with elevation is demanded.

In an attempt to hightlight subbands of HRTFs strongly correlated to the elevation perception, a novel frequency division strategy based on the statistical analysis of HRTFs is presented in this study. Specically, with F-ratio statistical analysis of HRTFs, the proposed frequency division strategy considers the dependencies between frequency components and elevation, enabling different resolution in different frequency regions. Then based on the frequency division, the energy of each subband of HRTFs is modelled as a function of elevation. Finally, by modifying the subband energies of HRTFs, the elevation of rendered sound image can be controlled. Subjective listening tests show that the proposed method works well in terms of elevation perception.

2. ELEVATION CONTROL METHOD BASED ON SUBBAND ENERGY MODIFICATION

2.1. Principle of elevation control using subband energy modification

Previous studies have demonstrated that the localization in the median plane is strongly correlated to the macroscopic patterns of HRTFs [1, 3, 4, 10]. For this reason, by modifying the subband energies of HRTFs, the perceived elevation angle of the transfer function could be changed. Mathematically, it can be expressed as,

$$H_m(\omega, b_n, \theta_m) = \quad T(b_n, \theta_m) H_o(\omega, b_n) \quad n = 1, 2, \cdots, P \quad (1)$$

where $H_o(\omega, b_n)$ denotes the original HRTF at angular frequency ω in subband b_n. $H_m(\omega, b_n, \theta_m)$ denotes the modified HRTF with target elevation θ_m. P is the total number of subbands. $T(b_n, \theta_m)$ is the band energy modification function which is defined as,

$$T(b_n, \theta_m) = \quad \frac{E_n(\theta_m)}{E_n(\theta_o)} \quad (2)$$

where $E_.(\cdot)$ denotes a subband energy model of HRTFs which would be further described in Section 2.3. $E_n(\theta_m)$ and $E_n(\theta_o)$ denote the energy of subband b_n of HRTF at target elevation θ_m and that at original elevation θ_o, respectively.

In order to control the elevation of the perceived elevation of sound image, the energies of original HRTF in subbands $b_1, b_2, \cdots, b_P$ are boosted or attenuated, as shown in Eqs. (1). Consequently by this simple and easy way, each subband energy of original HRTFs can be modified to perceptually approximate to those of HRTFs at target elevation.

2.2. Subband division of HRTFs based on F-ratio analysis

According to the method by Karapetyan *et al.* [16], a 24 channels auditory scale filter bank was used to obtain the energies of the early reflections of elevated BRIRs. The subband division strategy that used is based on the excitation patterns in human auditory system, which has high resolution in low frequency regions and low resolution in high frequency regions. However, HRTF represents the physical properties of anthropometric parameters that involve physiological acoustics. Practically, it is unnecessary to maintain high resolution in frequency regions which are physiologically irrelevant to the elevation.

In order to emphasize subbands of HRTFs which are strongly correlated to elevation perception, a novel frequency division strategy based on the statistical analysis of HRTFs is presented. As widely used for feature extraction in pattern recognition, the Fisher's F-ratio is adopted in this study, which is defined by

$$F\_ratio = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{M \cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_i^j - u_i)^2} \quad (3)$$

where x_i^j is the magnitude spectra of the jth subject in the database of elevation i with $j = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, M$. u_i and u are the magnitude spectra averages for elevation i and for all selected elevations, respectively, which are defined by

$$u_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j; u = \frac{1}{M \cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N} x_i^j \quad (4)$$

Conceptually, the F-ratio is a function of frequency which represents the inter-elevation variance to intra-elevation variance. The larger score of F-ratio means more vertical information is encoded in corresponding frequency ranges.

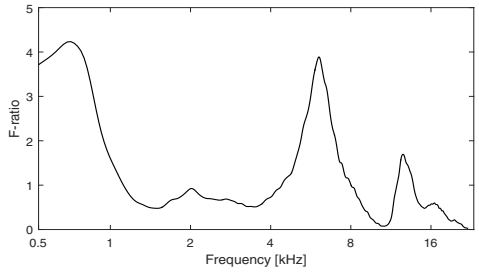

Fig. 1. F-ratio calculated by HRTFs in CIPIC [17] database.

We performed F-ratio calculation on the CIPIC HRTF database [17], and the F-ratio curve with a log frequency scale is plotted in Fig. 1. From Fig. 1, it can be seen that F-ratio is not monotonic, indicating vertical information encoded in HRTFs are dynamic. Therefore, a non-uniform subband division is proposed in which different regions require different frequency resolution based on the dependencies between frequency components and elevation.

According to F-ratio, the frequency range (500 Hz to 18 kHz) is divided into n regions. More specifically, the integration of the log scale F-ratio curve is first calculated and divided into n parts. Then the end point of each part is picked out as center frequency of each subband which would be used to generate gammatone-shaped band-pass filters.

Based on the frequency division strategy described above, the designed non-uniform subband filters for band energy calculation are shown in Fig. $2(a)$ which are denoted as NUSF (non-uniform subband filters). For comparison, the filter banks based on the excitation patterns in human auditory system are also plotted in Fig. $2(b)$ which are denoted as EPSF (excitation patterns based subband filters). From Fig. 2, it is easy to find that the resolution of EPSF decreases as frequency increases, whereas the NUSF has high resolution in the frequency regions with large F-ratio. That is, the proposed method shows high resolution in frequency regions which encode more elevation information.

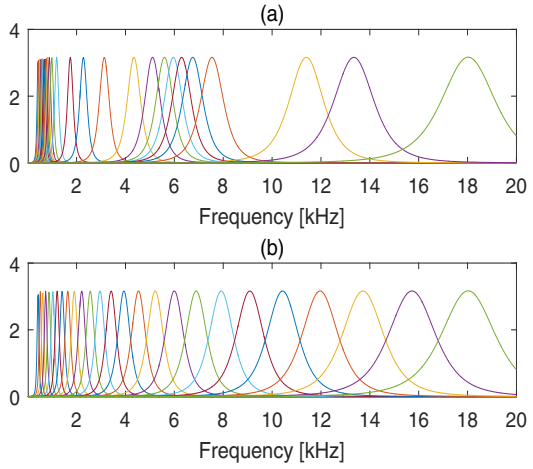

Fig. 2. Illustrations for NUSF [(a)] and EPSF [(b)]. A higher resolution in frequency regions with large F-ratio can be find in NUSF.

2.3. Energy model of HRTF subband

In this section, the subband energies of different elevations are first obtained by filtering vertical HRTFs with the filter banks, then formulated as a function of elevation. For a given elevation θ, the band energy E_n of subband b_n can be estimated using the polynomial regression method given by

$$\hat{E_n}(\theta, A_n) = \sum_{i=0}^{K} a_{i,n}\theta^i, \tag{5}$$

where $A_n = \{a_{0,n}, a_{1,n}, \cdots, a_{K,n}\}$ denotes the polynomial regression coefficients and K is the order of polynomial regression.

Further considering the independence among different band, the regression coefficients A_n of subband b_n can be obtained in the sense of least square error, given by

$$\hat{A_n} = \arg\min_{A_n \in \Re} \sum_{r\in\Omega} \sum_{\theta\in\Theta} \left(E_n(\theta) - \hat{E_n}(\theta, A_n) \right)^2, \tag{6}$$

where $E_n(\theta)$ denotes the subband energy of the measured HRTFs with elevation θ of subband b_n, Ω and Θ are the sets of HRTFs and elevations. These polynomial regression coefficients characterise the changes of the subband energies along with elevation which can be used in Section 2.1 for extracting the band energy modification function $T(b_n, \theta_m)$.

3. EXPERIMENTAL EVALUATION

Listening tests were conducted to evaluate the localization performance of the proposed elevation control method. For comparison, the subband energies were obtained by using the filter banks in Fig. $2(a)$ and (b).

Fifteen young adults (aged 22-31, 12 males and 3 females) with normal hearing sensitivity served as paid volunteers in the experiment. In order to reduce the effect of non-individual HRTFs on sound source localization, the method developed by Akagi *et al.* [18] was adopted to find the best HRTFs from CIPIC database for listening test. The source signal is a 44.1 kHz 250-ms Gaussian noise bandpass filtered between 200 Hz and 18 kHz with 20-ms cosine square on-set and off-set ramps.

The HRTFs at elevation of 0° were used as the original HRTFs and by modifying the subband energies of these HRTFs, the HRTFs of different elevation can be synthesized. The virtual acoustic stimuli were generated by convolving the sound source signals with the modified HRTFs and then delivered to subjects at 70 dB SPL via Audio-technica ATH-A2000Z headphone whose acoustic transfer property was compensated. Target virtual elevations of the stimuli consisted of thirteen directions from -45° to 90° in increments of 22.5° in the frontal median plane.

Subjects were instructed to respond to the perceived source position of the synthesized stimuli via graphical user interface. Prior to the test, each subject listened to a set of reconstructed samples to get familiar with the testing procedure. During the formal test, each subject had to listen to a total of 104 stimuli (2 methods * 13 directions * 4 times). The stimuli were presented randomly and could be listened to several times until the subject made a decision. Moreover, each subject was also instructed to check the box on the display when he/she perceived a sound image inside his/her head.

Fig. 3 shows the sound localization results averaged across all subjects directions, in which the area of each circle is proportional to the number of responses and the ordinate and the abscissa reprent the subject's responded elevation, the

Table 1. Averaged localization errors for 13 test directions

Method	Error (°)												
Target elevation	-45	-33.75	-22.5	-11.25	0	11.25	22.5	33.75	45	56.25	67.5	78.75	90
EPSF-based method	36.17	29.60	23.60	20.54	14.20	16.47	23.37	18.94	20.30	20.76	23.97	19.77	27.89
NUSF-based method	31.87	29.01	15.09	17.98	12.68	16.76	17.90	17.60	18.66	13.28	12.95	15.28	13.21

target elevation of stimuli, respectively. The results shows that most of the responses of both methods are distributed near the diagonal line, demonstrating the feasibility of the subband energy modification method for elevation control. Further observation of Fig. 3 reveals that the proposed NUSF-based method performs better than the EPSF-based method, especially at elevation of $-45°$, $-22.5°$, $67.5°$ and $90°$. The averaged localization errors for each test direction are listed in Table. 1. As can be seen, averaged localization errors of proposed NUSF-based method is much smaller than those of the EPSF-based method at $-22.5°$, $22.5°$, $56.25°$, $67.5°$, $78.75°$ and $90°$.

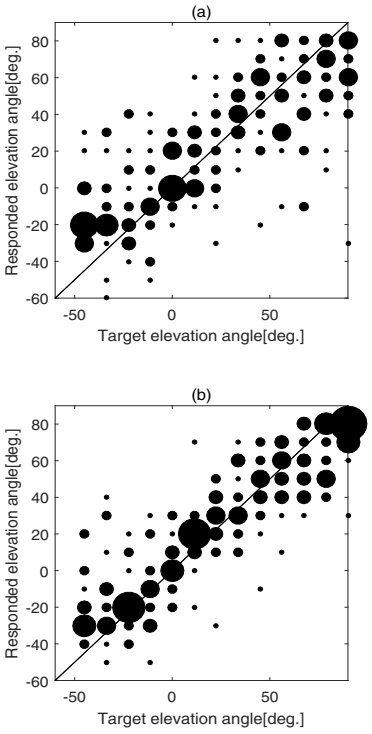

Fig. 3. The sound localization results averaged across all subjects for test directions. (a) represents the localization performance of the EPSF-based method; (b) represents the localization performance of the proposed NUSF-based method.

To furher examine the effects of these two methods and directions (13 directions), the absolute localization errors are subjected to statistical analysis using the localization error as the dependent variable, and method and direction as the two within-subjects factor. Two-way analysis of variance (ANOVA) with repeated measures indicates the significant effects of method $[F(1,14) = 5.085, p = 0.0407]$, and direction $[F(12,168) = 2.954, p = 0.0009]$. There are no significant interactions between method and direction $[F(12,168) = 0.734, p = 0.7171]$.

4. DISCUSSION

The previous sections describe an effective method for elevation control in the median plane. It should be noted that the HRTFs of sagittal planes with different lateral angles from the right- to the left-hand side in equal intervals of $15°$ are also analyzed. For lateral angle $\phi \in [-45°, 45°]$, the F-ratio analysis on HRTFs shows quite similar tendency as that in the median plane. For lateral angle $\phi \in (-45°, -90°) \cup (45°, 90°)$, the F-ratio analysis reveals that the contribution of low frequency components of ipsilateral HRTFs increases which is consistent with the results from Algazi *et al.* that there exists elevation-dependent features at low frequencies when the source is located away from the median plane [19].

Although F-ratio can evaluate the role of spectral cues for elevation perception, the calculation of F-ratio requires several subjects' HRTFs. Therefore, F-ratio defined in this study could be used as a tool for extracting common elevation-dependent features among subjects for elevation perception but has limits for explaining listener-specific spectral cues.

5. CONCLUSION

This paper presents an effective and simple elevation control method for binaural reproduction by modelling and modifying the subband energy of HRTFs. First, a novel frequency division strategy for subband is adopted which gives different frequency resolution to different frequency regions. Then, based on this frequency division strategy, the band energy of corresponding subbands are modified to control the elevation perception. Finally, subject listening tests show that the proposed elevation control approach yields a better elevated sound source perception and it is a viable option for efficient elevation control for spatial audio and applications in VR audio.

6. REFERENCES

[1] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[2] Bosun Xie, *Head-related transfer function and virtual auditory display*, J. Ross Publishing, 2013.

[3] Futoshi Asano, Yoiti Suzuki, and Toshio Sone, "Role of spectral cues in median plane localization," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 159–168, 1990.

[4] Dumidu S Talagala, Xiang Wu, Wen Zhang, and Thushara D Abhayapala, "Binaural localization of speech sources in the median plane using cepstral hrtf extraction," in *22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 2055–2059.

[5] EAG Shaw, TR Anderson, and RH Gilkey, "Binaural and spatial hearing in real and virtual environments, chapter acoustical features of human ear," *RH Gilkey and TR Anderson, Lawrence Erlbaum Associates, Mahwah, NJ, USA*, pp. 25–47, 1997.

[6] Kazuhiro Iida, Motokuni Itoh, Atsue Itagaki, and Masayuki Morimoto, "Median plane localization using a parametric model of the head-related transfer function based on spectral cues," *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, 2007.

[7] Vikas C Raykar, Ramani Duraiswami, and B Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 364–374, 2005.

[8] Kazuhiro Iida and Yohji Ishii, "Effects of adding a spectral peak generated by the second pinna resonance to a parametric model of head-related transfer functions on upper median plane sound localization," *Applied Acoustics*, vol. 129, pp. 239–247, 2018.

[9] Hironori Takemoto, Parham Mokhtari, Hiroaki Kato, Ryouichi Nishimura, and Kazuhiro Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3832–3841, 2012.

[10] Erno HA Langendijk and Adelbert W Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002.

[11] Dumidu S Talagala, Wen Zhang, Thushara D Abhayapala, and Abhilash Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1207–1217, 2014.

[12] Peter H Myers, "Three-dimensional auditory display apparatus and method utilizing enhanced bionic emulation of human binaural sound localization," Mar. 28 1989, US Patent 4,817,149.

[13] Chong-Jin Tan and Woon-Seng Gan, "User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems," *Electronics letters*, vol. 34, no. 25, pp. 2387–2389, 1998.

[14] Thomas McKenzie and Gavin Kearney, "Preliminary investigations into binaural cue enhancement for height perception in transaural systems," in *Audio Engineering Society Conference: 61st International Conference: Audio for Games*, Feb 2016.

[15] Dingding Yao, Junfeng Li, and Risheng Xia, "A parametric elevation control approach for binaural reproduction," *Applied Acoustics*, vol. 148, pp. 360 – 365, 2019.

[16] Aleksandr Karapetyan, Felix Fleischmann, and Jan Plogsties, "Elevation control in binaural rendering," in *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[17] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano, "The cipic hrtf database," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 99–102.

[18] Masato Akagi and Hideki Hisatsune, "Admissible range for individualization of head-related transfer function in median plane," in *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 2013, pp. 326–329.

[19] V Ralph Algazi, Carlos Avendano, and Richard O Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, 2001.