ROBUST FULL-SPHERE BINAURAL SOUND SOURCE LOCALIZATION

Benjamin R Hammond, Philip J.B. Jackson

CVSSP
University of Surrey
Guildford, GU2 7XH, UK

ABSTRACT

We propose a novel method for full-sphere binaural sound source localization that is designed to be robust to real world recording conditions. A mask is proposed that is designed to remove diffuse noise and early room reflections. The method makes use of the interaural phase difference (IPD) for lateral angle localization and spectral cues for polar angle localization. The method is tested using different HRTF datasets to generate the test data and training data. The method is also tested with the presence of additive noise and reverberation. The method outperforms the state of the art binaural localization methods for most testing conditions.

Index Terms— Binaural, Localization, HRTF, Cepstrum.

1. INTRODUCTION

There are many applications that benefit from the ability for a machine to localize a sound source using a binaural recording of a sound arriving from any direction on the full-sphere around the listener, such as assisting the hearing impaired or being implemented in an augmented reality system [1, 2, 3]. The head related transfer function (HRTF) describes the frequency based filtering effect of the listener's head, pinna, and torso at the ear canals from a point in space. The time domain equivalent is called the head related impulse response (HRIR).

A HRTF dataset consists of a collection of measured HRTFs at different directions of arrival around the listener. The HRTF recorded at one ear and the corresponding HRTF recorded at the other ear comprise a HRTF pair. From this HRTF pair, interaural parameters can be derived, including the frequency dependent interaural phase difference (IPD) and interaural level difference (ILD). These interaural parameters are frequently used in binaural sound source localization methods. It was shown in [4] that while the state of the art methods which only use interaural parameters are able to accurately estimate the lateral angle of a sound, they are not able to localize its elevation. Spectral cues provide an additional cue for elevation localization [5]. In particular, the peaks and notches found in the HRTFs and their relative levels are used to discern the direction of arrival (DOA) of the sound source in elevation [6].

When recording impulse responses, unwanted artefacts are introduced into the recording. These artefacts are associated with the acoustic environment; the measurement procedure and equipment; and the post-processing of the data [7, Chapter 8]. Some localization methods are tested by generating their test data and training data synthetically from the exact same HRTF dataset [8, 9]. This is referred to as the matched HRTF condition throughout this paper. The mismatched HRTF condition refers to the case of testing a method using different HRTF datasets to generate the training and test data. For this mismatched condition, the HRTF datasets are captured using the same model of dummy head, but in different rooms, using different measurement equipment. For full-sphere binaural sound source localization to be possible in a real world case with the use of a pre-measured HRTF dataset, the method must be robust to additive and convolutive noise provided by the recording equipment. Therefore it is necessary to test using the mismatched HRTF condition. The state of the art full-sphere binaural sound source localization methods have not been tested using this mismatched condition. Either they have been tested using the matched HRTF condition [8, 9], or the test sounds and template HRTFs were recorded in the same location, using the same recording equipment [10].

In this paper, we propose a full-sphere binaural sound source localization method that is designed to be robust to this mismatched HRTF condition. It uses the IPD for lateral angle localization and spectral cues for elevation localization using a training dataset of HRIR pairs. The proposed method is tested for both the matched and mismatched HRTF conditions, and is additionally tested with reverberation and additive noise.

2. PROPOSED METHOD OVERVIEW

In anechoic conditions, the signal received at the ear of a listener, $y_\zeta(t)$ from a single sound source in space, $s(t)$ is filtered by the head related impulse response (HRIR), $h_\zeta(t)$, where the channel index, $\zeta \in \{l, r\}$ denotes the left and right ear respectively. This describes the direct path of the sound. In reverberant environments, the total impulse response is comprised of the HRIR and an additional component provided by the acoustic reflections, $\epsilon_\zeta(t)$. The acoustic reflections consist of early reflections, which have directionality and later reflections which are diffuse [7]. If this sound is recorded at the ears, the recording equipment and procedure may also introduce convolutive noise, $\nu_\zeta(t)$ and additive noise, $\chi_\zeta(t)$. Thus, the signal received at the ear of the listener is given by: $y_\zeta(t) = s(t) * (h_\zeta(t) + \epsilon_\zeta(t)) * \nu_\zeta(t) + \chi_\zeta(t)$. Let the discrete time domain equivalent of $y_\zeta(t)$ be $y_\zeta(n)$. The short-time Fourier transform of $y_\zeta(n)$ is $Y_\zeta(p, m)$, where p and m are the frequency index and time index respectively. For each time-frequency unit (p, m), the IPD is defined as: $\phi(p, m) = \angle(Y_l(p, m)/Y_r(p, m)) \in (-\pi, \pi]$.

2.1. Mask

The aim of the mask is to identify time-frequency units that are dominated by the sound from the direct path and those which are dominated by noise or acoustic reflections [11, Chapter 1]. If a sound source is active in a given frequency band, the level of the sound

EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1).

from the direct path should be higher than that of the early reflections and diffuse reverberation. In principle, the probability density of the IPD for diffuse reverberation and stereo uncorrelated noise should be relatively flat [12]. Conversely, for a given frequency band, a high ratio of time-frequency units dominated by the direct path of the sound to time-frequency units dominated by reflections or noise yields a peak in the probability density of the IPD. The authors in [12] created a mask that removes diffuse noise based on the probability density of the interaural parameters. We have designed our mask to remove early reflections and diffuse noise. The diffuse noise can be removed even if it shares the same IPD as the direct sound. We achieve this by additionally considering the level at the left and right ears, $S(p,m) = \max\{Y_l(p,m), Y_r(p,m)\}$. In each frequency band, p, the time indices, $\hat{m} \in \{1,2,3...\}$ are ordered by their corresponding level, $S(p,\hat{m})$, such that the first index, $\hat{m}=1$ corresponds to the highest level in frequency band, p and the second index, $\hat{m}=2$ corresponds to the second highest level, etc. A kernel density estimator is used to estimate the probability density function, $R(\phi;p,b)$ of the IPD, ϕ, using a Gaussian kernel smoothing function. To account for phase circularity, the PDF is estimated using all aliases of the IPD in the range $(-3\pi, 3\pi]$ for each time-frequency unit:

$$R(\phi;p,b) = \frac{1}{\rho\sqrt{2\pi\sigma^2}} \sum_{d=-1}^{1} \sum_{\hat{m}=b\rho+1}^{b\rho+\rho} e^{-\frac{(\phi-\phi(p,\hat{m})+2\pi d)^2}{2\sigma^2}}, \quad (1)$$

where b refers to the sample, $b \in \{0,1,2,3,...,Q\}$ of sample size ρ of time-frequency units for a given frequency band, and σ is a fixed standard deviation for all of the Gaussian components that make the kernel distribution. The value $\upsilon_p = \arg\max_{\phi\in(-\pi,\pi]}(R(\phi;p,b=0))$ is stored for each frequency band. To create the mask, $M(p,\hat{m})$, a threshold, η is chosen, such that if $R(\phi=\upsilon_p; p=A, b=0) < \eta$, for frequency band, A, then $M(p=A,\hat{m}) = 0, \forall\hat{m}$, and if $R(\phi=\upsilon_p; p=A, b) > \eta, \forall b$, for frequency band, A, then $M(p=A,\hat{m}) = 1, \forall\hat{m}$. Otherwise, the sample, B_p for each frequency band is found, such that: $B_p = \arg\min_b(R(\phi=\upsilon_p;p,b) < \eta)$,and the mask is created from this:

$$M(p,\hat{m}) = \begin{cases} 1 & (\upsilon_p - \xi < \phi(p,\hat{m}) < \upsilon_p + \xi) \wedge (\hat{m} < \rho B_p) \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where ξ is a fixed constant.

2.2. Cone of confusion

The interaural-polar coordinate system is used to describe the direction of arrival of the sound source. It does so with a lateral angle, $\lambda \in [-90^\circ, 90^\circ]$, and polar angle $\theta \in [-180^\circ, 180^\circ)$. For the polar angles: 0° is at the front of the listener, 90° is above the listener, and 180° is at the back of the listener [13]. For differing lateral angles, there is a diversity in IPD, for a given frequency band. However, there is a similarity in IPD in different regions on the polar dimension, which gives rise to the cone of confusion [14, Chapter 1]. For time-frequency units containing only sound from the direct path within a given frequency band, the IPD should be the same for each of these units, irrespective of the sound source, for a given DOA.

For non-moving sound sources, the polar angle must be estimated using spectral cues. For the proposed method, a HRTF dataset is used as training data, and the spectral cues of the test sound are compared with the spectral cues of HRTF pairs in the training dataset. In contrast to the IPD, the spectral cues vary depending on the sound source, which may provide confounding information for a comparison with the spectral cues of the HRTF pairs in the training dataset. Because the IPD is sound source agnostic, the lateral angle of the sound can be estimated with greater reliability than the polar angle. As such, the cone of confusion is firstly estimated and the polar angle is then estimated as a point on the cone of confusion. For the proposed method then, the HRTF dataset is used as training data to estimate the cone of confusion by determining the most likely HRTF pair for each polar angle in the dataset. A grid is formed on the lateral-polar plane, consisting of points in 2° increments in the lateral and polar dimensions. $\alpha \in \{1,2,3,...,\gamma\}$ and $\beta \in \{1,2,3,...,\mu\}$ are the indices of the lateral and polar angles of the points on the grid, respectively. The HRIR pair in the HRTF dataset with the closest direction of arrival to each grid point is referred to as $h_\zeta^{\alpha\beta}(n)$, which has a corresponding HRTF pair $H_\zeta^{\alpha\beta}(p)$. The FFT size used to transform the HRIRs to HRTFs is the same as the FFT size used to generate each time frame for $\phi(p,m)$, as such the frequency index, p can be used for both $H_\zeta^{\alpha\beta}(p)$ and $\phi(p,m)$. From this, the IPD templates are generated: $\Upsilon^{\alpha\beta}(p) = \angle(H_l^{\alpha\beta}(p)/H_r^{\alpha\beta}(p)) \in (-\pi,\pi]$. The IPD difference, $\Gamma^{\alpha\beta}(p,m)$ is given by: $\Gamma^{\alpha\beta}(p,m) = \min_{d\in\{-1,0,1\}}(|\Upsilon^{\alpha\beta}(p) - (\phi(p,m) + 2\pi d)|)$. For each polar angle β_κ, the corresponding lateral angle α_κ that lies on the cone of confusion is given by the maximum likelihood of the masked IPD difference:

$$(\alpha_\kappa, \beta_\kappa) = \underset{\alpha\in\{1,...,\gamma\},\beta=\kappa}{\arg\max} \sum_{p,m} M(p,m).\ln(\mathcal{N}(\Gamma^{\alpha\beta}(p,m)|0,1)), \quad (3)$$

where $\kappa \in \{1,2,3,...,\mu\}$ are the indices of the cone of confusion, which have corresponding lateral and polar angles $(\alpha_\kappa, \beta_\kappa)$. It was found experimentally that the optimum result for cone of confusion estimation was yielded by using an upper limit on the frequency indices, p corresponding to a frequency of 11kHz.

2.3. Spectral pre-processing

The test sound is created by reconstructing the binauralized signal in the time domain, $\hat{y}_\zeta(n)$. In order to do this, the mask is applied in the time-frequency domain: $\hat{Y}_\zeta(p,m) = M(p,m).Y_\zeta(p,m)$, after which the inverse short-time Fourier transform is applied. For the spectral pre-processing stage, the same operations are performed on both the test sound and the HRIR pairs on the estimated cone of confusion, for both the left and right channels. Let $\Psi(k)$ denote the log-magnitude Fourier spectrum of a signal, $\psi(n)$, such that $\Psi(k) = \log_{10}|\mathcal{F}\{\psi(n)\}|$, where $\mathcal{F}$ is the Discrete Fourier Transform (DFT), and k is the frequency index in the discrete frequency domain. The log-magnitude Fourier spectrum of the reconstructed signal is then denoted as $\hat{\Psi}(k)$.

The authors in [5] show that the main cues for polar angle localization on the median plane are the locations and relative levels of the spectral peaks and notches in the HRTFs. To identify the peaks and notches, the rapid fluctuations present in the HRTFs and the spectrum of the test sound need to be removed. Additionally, in the spectrum of the test sound and the HRTF templates exists a component that slowly varies in magnitude throughout the frequency range. This kind of slow varying component is present in the spectrum of most natural sounds and is also introduced by convolutive noise from recording equipment. This component has such a slow variation that it does not obscure the location of the peaks and notches or the relative level of neighbouring peaks and notches. In order to compare the relative positions of the peaks and notches in the test sound and templates then, this slow varying component needs to be removed. Two variants of this method are proposed which both remove the rapid fluctuations and the slow varying component present in the spectrum,

which is denoted by $\grave{\Psi}(k)$ in the general case. For both variants of the spectral pre-processing stage, it was found experimentally that the optimum results were achieved using an upper limit, J for the frequency index, k corresponding to a frequency of 11kHz. The frequency domain consists of indices corresponding to frequency bands in the mask that contain only zeros, k_D and indices corresponding to frequency bands that contain non-zeros, k_C.

The first variant is referred to as the Gaussian Filter variant. For this variant, consider that applying a high-pass lifter to the real cepstrum allows for the removal of the slow varying component in the frequency domain [15, Chapter 31] [16, Chapter 13]. As the transformation from the frequency domain to the cepstral domain involves a logarithm, the magnitude in the frequency domain cannot contain any zeros. To avoid zeros in the frequency domain and to preserve the slow varying component in the log-magnitude Fourier spectrum, linear interpolation is applied to the log-magnitude Fourier spectrum over the frequency region k_D, which yields $\check{\Psi}(k)$. To remove the rapid fluctuations in the log-magnitude Fourier spectrum, a Gaussian filter is used [5]. Defining the Gaussian filter function as:

$$G_W\{G(q); x_1, \sigma_g\} = \frac{1}{\sqrt{2\pi\sigma_g^2}} \sum_{x=-x_1}^{x1} G(q+x).e^{\frac{-x^2}{2\sigma_g^2}}. \quad (4)$$

The smoothed log-magnitude Fourier spectrum, $\tilde{\Psi}(k)$ is given by: $\tilde{\Psi}(k) = G_W\{\check{\Psi}(k); x_1, \sigma_g\}$. It was experimentally found that a value of σ_g corresponding to a frequency of 450Hz yielded the optimum results. The slow varying component in the log-magnitude Fourier spectrum is removed by the applying a high-pass lifter in the cepstral domain and transforming the resulting signal back to the frequency domain:

$$\grave{\Psi}(k) = \mathcal{F}\{w(v).\mathcal{F}^{-1}\{\tilde{\Psi}(k)\}\}, w(v) = \begin{cases} 1 & 2 \leq v \leq N-2 \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where $v \in \{0, ..., N-1\}$ are the indices of the cepstral coefficients, c_v in the cepstral domain.

The second variant is referred to as the Cepstrum Regression variant. For this variant, consider that applying a bandpass lifter in the cepstral domain removes both the slow varying component in the log-magnitude Fourier spectrum and the rapid fluctuations. However, in order to transform a signal to the cepstral domain, the magnitude in the Fourier domain must be non-zero throughout the range. Instead of transforming a signal to the cepstral domain and applying a bandpass lifter, multiple linear regression can be used to estimate the cepstral coefficients. Exploiting the fact that for the real cepstrum, $c_v = c_{N-v}$, the Fourier transform of the real cepstrum is given by: $\Psi(k) = c_0 + 2c_1\cos(\omega(k)) + 2c_2\cos(2\omega(k)) + ...$, where $\omega(k)$ is normalized discrete frequency, defined as $\omega(k) = \pi k/J$. To find the cepstral coefficients, $a = [c_0, ..., c_u]$, we generate a model matrix as: $X = [1 \quad 2\cos(\omega(k_C)) \quad 2\cos(2\omega(k_C)), ..., 2\cos(u\omega(k_C))]$. The coefficient estimates, $\hat{a}$ can be obtained by the usual least squares method: $\hat{a} = (X^T X)^{-1} X^T.\hat{\Psi}(k_C)$, where X^T is the transpose of X [17]. The rapid fluctuations and slow varying component of the log-magnitude Fourier spectrum are removed by reconstructing the log-magnitude Fourier spectrum without using the lower order or higher order coefficients:

$$\grave{\Psi}(k) = 2\sum_{P=2}^{u} c_P \cos(P\omega(k)). \quad (6)$$

It was experimentally found that $u = 8$ yielded the optimum results.

For the specific cases, let $\grave{\Psi}(k)$ be denoted by $\grave{Y}_\zeta(k)$ and $\grave{H}_\zeta^\kappa(k)$ for the log-magnitude Fourier spectrum of the test sound and HRIR pairs on the cone of confusion respectively with spectral pre-processing applied.

2.4. Polar angle estimation

The spectral difference, $\Theta_\zeta^\kappa(k)$ is given by: $\Theta_\zeta^\kappa(k) = \grave{Y}_\zeta(k) - \grave{H}_\zeta^\kappa(k)$. For the polar angle estimation, it was found experimentally that the optimum results were achieved using a lower limit, I and upper limit, J for the frequency index, k corresponding to frequencies of 4kHz and 11kHz respectively. The log-likelihood distribution, $Z(\beta_\kappa)$ of each polar angle on the cone of confusion is given by: $Z(\beta_\kappa) = \sum_\zeta \sum_{k=I}^{J} \ln(\mathcal{N}(\Theta_\zeta^\kappa(k)|0,1))$. The distribution is noisy and it was found that the results improved by smoothing the distribution. The smoothed log-likelihood distribution is given by: $\tilde{Z}(\beta_\kappa) = G_W\{Z(\beta_\kappa); x_1, \sigma_g\}$. It was experimentally found that a value of σ_g corresponding to 10° yielded the optimum results. The HRTF template index, $\hat{\kappa}$ corresponding to the estimated lateral and polar angle of the test sound is given by:

$$\hat{\kappa} = \arg\max_\kappa(\tilde{Z}(\beta_\kappa)). \quad (7)$$

3. TESTING PROCEDURE

Two reference methods are tested: The first reference method is the Cross-Convolution method [18, 9], which is a full-sphere localization method, and the second reference method is the Speech Prefilter method which is a median plane localization method [8]. To develop our implementation of both of these methods, we tested using the same HRTF dataset that the authors in both papers used to obtain results [19]. Our implementation of the cross-convolution method produced similar results to those reported in [18, 9] using speech sounds. For the Speech Prefilter method, we found that the performance of our implementation of this method was strongly dependent on the specific speech utterance used. As the specific speech utterances used and the sample size are not specified in [8], we could not directly compare our implementation to that of the authors.

10 environmental sounds have been selected for testing from the sounds used in [20]. Namely, they are: Waves crashing, electric saw cutting, water pouring, train moving, chopping wood, typing on keyboard, ice dropping into glass, bells chiming, cars honking and sheep baaing. Speech is one of the most important everyday sounds and is tested under its own category. 10 speech samples have been chosen from the CSTR VCTK corpus [21]. The speech samples have been selected to give an equal balance of male and female voices and a diverse set of accents.

The lateral angular error is the lateral angle between the ground truth test position and the estimated location of the sound source. The central angular error is the angle between the ground truth test position and the estimated position of the sound source, from the point of view of the listener [4]. As lateral angular errors have been shown to be relatively small when compared with errors along the cone of confusion, the central angular error is a good indication of the error in the polar dimension [22]. The HRTF dataset used to generate training data for all conditions is the Gauss-Legendre 2° dataset, measured in [23]. For the matched HRTF condition, the test sound is produced synthetically by convolving the HRIR pair from the same dataset as the training data with each of the mono sound sources above at the DOAs of interest. For the mismatched HRTF condition, the test sound is produced using the dataset measured at

the University of York as part of the SADIE project [24, 25]. Additionally, to test the presence of reverberation, a Binaural Room Impulse Response (BRIR) condition is tested. For this condition the test sound is produced using the BRIRs from the dataset measured in [26], with the dummy head facing the front of the room. A full sphere localization test is conducted using the DOAs of the loudspeakers in [26] for the matched HRTF, mismatched HRTF and BRIR conditions. Stereo uncorrelated white noise is added to the test sounds at 4 different signal-to-noise ratios (SNRs). Additionally, a median plane localization test is conducted with test sounds at 30dB SNR.

4. RESULTS AND DISCUSSION

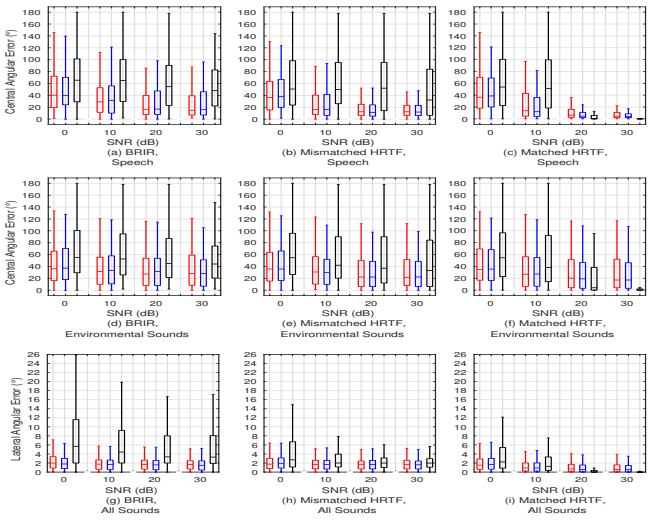

Fig. 1. Central angular error (°) (a-f) and lateral angular error (°) (g-i), as a function of signal-to-noise ratio (dB) for localization on the full sphere. Results are shown for test sounds generated with BRIRs, mismatched HRTFs, and matched HRTFs. The proposed method is shown with both variants: Gaussian Filter (red), and Cepstrum Regression (blue), as well as the reference method: Cross-Convolution (black). Horizontal line within box: median; box: inter-quartile range (IQR); whisker: within quartile ± 1.5.IQR; outliers are not shown.

Figure 1 shows the central angular error and lateral angular error for the full sphere localization test. Figure 2 shows the central angular error for the median plane localization test. For all cases, the two variants of the proposed method perform at a similar level to each other in all test conditions. In most test conditions, the proposed method outperforms the reference methods. For the majority of cases, for the proposed method, the lateral angular error is smaller than the lateral angle that humans are able to discern [27]. This is also true for the central angular error for speech, for all test conditions at lower noise levels. This demonstrates that the method is robust to convolutive and additive noise at lower noise levels. The localization accuracy for environmental sounds was slightly worse than for the speech sounds. This can be attributed to the spectrum of certain environmental sounds and certain HRTFs having peaks and notches in similar positions. This can also be attributed to the low sound level of the spectrum in the high frequency range, resulting in the sound in the high frequency range falling below the noise floor.

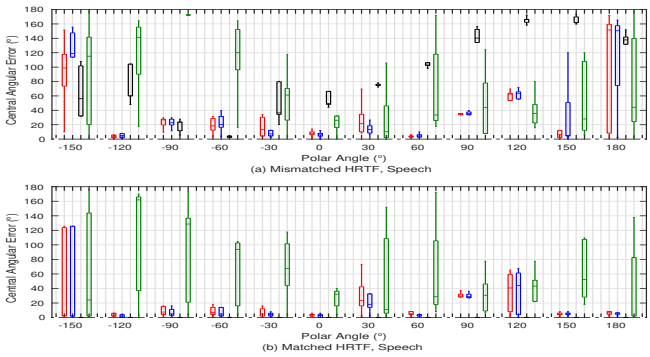

Fig. 2. Central angular error (°) , as a function of polar angle (°) for localization on the median plane. Results are shown for test sounds generated with mismatched HRTFs, and matched HRTFs, shown for speech sounds at 30dB SNR. The proposed method is shown with both variants: Gaussian Filter (red), and Cepstrum Regression (blue), as well as the reference methods: Cross-Convolution (black) and Speech Prefilter (green). Horizontal line within box: median; box: inter-quartile range (IQR); whisker: within quartile ± 1.5.IQR; outliers are not shown.

The Speech Prefilter method yielded reasonable results for sound sources above the listener and yielded the worst results for sound sources below the listener. For the Neumann KU-100 dummy head, there is a pinna notch present in the 3.5kHz -7.5kHz region of the HRTFs with DOAs below the dummy head, which is the region used by this method. As the method does not try to remedy the additive noise present, it is likely that the noise present in the same frequency region as the pinna notch provides confounding information, resulting in higher errors. It can be difficult to see on the plot, that the Cross-Convolution method, for the matched HRTF condition in the median plane localization test, estimates the DOA with a central angular error close to 0° for all angles. However the method performs poorly for the mismatched HRTF condition. The cross-convolution method relies on a comparison between the left channel and right channel, but as the head is theoretically almost symmetrical, the HRTF for the left channel should be similar to its right channel pair at all angles on the median plane. This indicates that for the matched HRTF condition, the method is using the unique measurement noise present in each HRTF pair in the dataset as a localization cue. This would not occur in a real life setting and demonstrates why developing a binaural localization method for use with the matched HRTF condition should be avoided.

5. CONCLUSION

In this paper we present a full-sphere localization method with two variants that is designed to localize a sound source in real world conditions. The method uses a mask designed to remove early reflections and diffuse noise. The method is effective in estimating the lateral angle of the sound source in all test conditions. The polar angle is estimated well for speech sounds in the mismatched HRTF condition and with a low level of noise and reverberation. Additionally, we show the disparity between results generated by the reference methods for the matched and mismatched HRTF conditions and make the case that future binaural localization methods should be developed for use with the mismatched HRTF condition.

6. REFERENCES

[1] A. Harma, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc*, vol. 52, no. 6, pp. 618–639, June 2004.

[2] P.F. Hoffmann, F. Christensen, and D. Hammershoi, "Insert earphone calibration for hear-through options," in *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*, August 2013.

[3] D. Jain, L. Findlater, J. Gilkeson, B. Holland, R. Duraiswami, D. Zotkin, C. Vogler, and J.E. Froehlich, "Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, New York, NY, USA, 2015, CHI '15, pp. 241–250, ACM.

[4] B.R. Hammond and P.J.B. Jackson, "Full-sphere binaural sound source localization by maximum-likelihood estimation of interaural parameters," in *Audio Engineering Society Convention 142*, May 2017.

[5] 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," *Applied Acoustics*, vol. 68, no. 8, pp. 835 – 850, August 2007.

[6] E. Blanco-Martin, F.J. Casajus-Quiros, J.J. Gomez-Alfageme, and L.I. Ortiz-Berenguer, "Estimation of the direction of auditory events in the median plane," *Applied Acoustics*, vol. 71, no. 12, pp. 1211–1216, December 2010.

[7] H. Kuttruff, *Room acoustics*, Spon Press, Abingdon, UK, fifth edition, 2009.

[8] D.S. Talagala, X. Wu, W. Zhang, and T.D. Abhayapala, "Binaural localization of speech sources in the median plane using cepstral hrtf extraction," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, September 2014, pp. 2055–2059.

[9] M. Usman, F. Keyrouz, and K. Diepold, "Real time humanoid sound source localization and tracking in a highly reverberant environment," in *2008 9th International Conference on Signal Processing*, October 2008, pp. 2661–2664.

[10] D.S. Talagala, W. Zhang, T.D. Abhayapala, and A. 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, March 2014.

[11] D. Wang and G.J. Brown, *Computational auditory scene analysis: principles, algorithms, and applications*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006.

[12] M.I. Mandel, R.J. Weiss, and D.P.W Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, February 2010.

[13] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, August 2014.

[14] B. Xie, *Head-related transfer function and virtual auditory display*, J. Ross Publishing, Inc., Florida, USA, second edition, 2013.

[15] D. Havelock, *Handbook of signal processing in acoustics*, Springer, New York, NY, USA, 2008.

[16] A. Oppenheim, *Discrete-time signal processing*, Pearson, Upper Saddle River, NJ, USA, third edition, 2010.

[17] E.B. Brooks, V.A. Thomas, R.H. Wynne, and J.W. Coulston, "Fitting the multitemporal curve: A fourier series approach to the missing data problem in remote sensing analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 9, pp. 3340–3353, September 2012.

[18] M. Rothbucher, D. Kronmuller, K. Diepold, M. Durkovic, and T. Habigt, "HRTF sound localization," in *Advances in Sound Localization*, P. Strumillo, Ed., chapter 5. INTECH Open Access Publisher, 2011.

[19] W.G Gardner and K.D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[20] B. Gygi, G.R. Kidd, and C.S. Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, pp. 839–855, August 2007.

[21] C. Veaux, J. Yamagishi, K. MacDonald, et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," http://dx.doi.org/10.7488/ds/1994, 2017, Accessed: 2018-02-10.

[22] T.R Letowski and S.T Letowski, "Auditory spatial perception: Auditory localization," Tech. Rep., DTIC Document, May 2012.

[23] B. Bernschutz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA) Conference on Acoustics*, March 2013, p. 29.

[24] G. Kearney and T. Doyle, "An HRTF database for virtual loudspeaker rendering," in *Audio Engineering Society Convention 139*. Audio Engineering Society, October 2015.

[25] G. Kearney, "The perception of auditory height in individualised and non-individualized dynamic cross-talk cancellation," in *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control*. Audio Engineering Society, July 2016.

[26] C. Pike and M. Romanov, "An impulse response dataset for dynamic data-based auralization of advanced sound systems," in *Audio Engineering Society Convention 142*, May 2017.

[27] E. Hendrickx, M. Paquier, V. Koehl, and J. Palacino, "Ventriloquism effect with sound stimuli varying in both azimuth and elevation," *The Journal of the Acoustical Society of America*, vol. 138, no. 6, pp. 3686–3697, December 2015.