ROBUST FULL-SPHERE BINAURAL SOUND SOURCE LOCALIZATION USING INTERAURAL AND SPECTRAL CUES

Benjamin R. Hammond, Philip J.B. Jackson

CVSSP, University of Surrey, Guildford, GU2 7XH, UK

ABSTRACT

A binaural sound source localization method is proposed that uses interaural and spectral cues for localization of sound sources with any direction of arrival on the full-sphere. The method is designed to be robust to the presence of reverberation, additive noise and different types of sounds. The method uses the interaural phase difference (IPD) for lateral angle localization, then interaural and spectral cues for polar angle localization. The method applies different weighting to the interaural and spectral cues depending on the estimated lateral angle. In particular, only the spectral cues are used for sound sources near or on the median plane.

Index Terms— Binaural, localization, HRTF

1. INTRODUCTION

To localize a sound source on the full-sphere i.e. from any direction of arrival (DOA) around the listener, for non-moving sources, the main localization cues are the interaural and spectral cues. The head related transfer function (HRTF) describes the frequency based filtering effect of the listener's morphology at the listener's ear canal from a point in space. The time domain equivalent is the head related impulse response (HRIR). A HRTF dataset consists of a collection of measured HRTFs at different DOAs around the listener for both ears [1, Chapter 1]. The interaural and spectral cues can be derived from these HRTFs. The interaural and spectral cues change between listeners, as the morphology of each listener is different. Therefore, in order to localize a non-moving sound source on the full-sphere, the listener's unique HRTF dataset is needed. The method in this paper is based on the method in [2]. The proposed method uses the interaural phase difference (IPD) to estimate the lateral angle of the sound source. The elevation is then estimated using a weighted combination of the interaural and spectral cues. The method aims to be robust to additive noise and diffuse reverberation by consideration of the probability density function of the IPD in each frequency band. Additionally, the method aims to be robust to reverberation by using data smoothing techniques on the localization cues. Finally, for robustness to different sound types and convolutive noise provided by the recording equipment, linear regression is used to remove the slow varying component in the spectra of the binaural sound received at both ears of the listener. The proposed method is compared with three state of the art reference methods [3, 4, 5, 6], which have been selected for their different approaches to binaural sound source localization.

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

2. PROPOSED METHOD

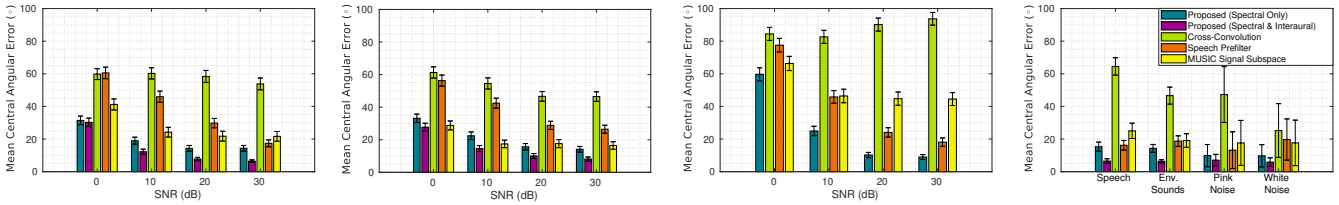
In anechoic conditions, the signal received at the ear of a listener, $y_\zeta(t)$ from a single sound source in space, $s(t)$ is filtered by the head related impulse response (HRIR), $h_\zeta(t)$, where the channel index, $\zeta \in \{l, r\}$ denotes the left and right ear respectively, and t denotes continuous time. 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 [2, 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: $\phi(p, m) = \angle(Y_l(p, m)/Y_r(p, m)) \in (-\pi, \pi]$. The magnitude ratio (MR) is a bounded form of the ILD [8]. As the ILD can result in extreme values, the MR is used in its place. The three localization cues used in this paper then are the IPD, MR and spectral cues. Rapid fluctuations in the frequency domain of the localization cues of the test sound are dissimilar to those in the frequency domain of the localization cues extracted from the HRTF template with the same DOA as the test sound. However, the slow varying component in the frequency domain of each localization cue extracted from the test sound is similar to the slow varying component in the frequency domain of each localization cue extracted from the HRTF template with the same DOA as the test sound. Cepstral liftering can be used to extract this slow varying component in the frequency domain from both the spectral cues and the MR [9, Chapter 31] [10, Chapter 13]. However, cepstral liftering cannot be used to extract this slow varying component from the IPD, as it is a circular variable. Instead, kernel density estimation is used with a Gaussian kernel [11, Chapter 7]. Consider the level at the left and right ears, $S(p, m) = \max(\{Y_l(p, m), Y_r(p, m)\})$. In each frequency band, $p \in \{1, 2, 3, ..., Q\}$, 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, $R(\phi, f)$ of the IPD, ϕ, as a function of continuous frequency, f, 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. In [2], univariate kernel density estimation was used to estimate the PDF of the IPD in each frequency band. For the proposed method, it was found that the IPD of the direct component of the sound was better

estimated by using bivariate kernel density estimation to estimate the PDF of the IPD as a function of frequency:

$$R(\phi,f)=\frac{1}{6\pi\rho Q\sigma_\phi\sigma_f}\sum_{d=-1}^{1}\sum_{\hat{m}=1}^{\rho}\sum_{p=1}^{Q}$$
$$\exp\left(-\left(\frac{(\phi-\phi(p,\hat{m})+2\pi d)^2}{2\sigma_\phi^2}+\frac{(f-f(p,\hat{m}))^2}{2\sigma_f^2}\right)\right), \quad (1)$$

where all Gaussian components that make the kernel distribution are assigned the same values for σ_ϕ and σ_f. It was experimentally found that a value of $\rho = 30$ yielded the best results. The PDF is then assessed along the IPD dimension only at the frequencies, $f_p \in \{f_1, f_2, f_3, ..., f_Q\}$ corresponding to each frequency index, p, which yields a one-dimensional PDF, $\hat{R}(\phi;p)$ as a function of IPD, ϕ for each frequency index, p: $\hat{R}(\phi;p)=\frac{R(\phi,f=f_p)}{\int_{\phi=-\pi}^{\pi}R(\phi,f=f_p)\,d\phi}$. The estimated IPD of the direct component of the sound for each frequency band, p is given by: $\upsilon(p)=\arg\max_{\phi\in(-\pi,\pi]}(\hat{R}(\phi;p))$, and the maximum value of the PDF in each frequency band is given by: $A(p)=\hat{R}(\phi=\upsilon(p);p)$. 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 component of the sound to time-frequency units dominated by reflections or stereo uncorrelated noise yields a peak in the probability density of the IPD. Additionally, in typical reverberant environments, for a given frequency band, time-frequency units that are dominated by the direct component of the sound should have a level that is higher than time-frequency units dominated by reflections or stereo uncorrelated noise. Using this information, the frequency bands within which the sound source is active can be identified. To create the mask, $M_x(p)$, a threshold, η_x is chosen. The value of η_x used to generate the frequency mask for the IPD, MR and the left and right spectral cues used to estimate the polar angle are denoted by η_ϕ, η_Ξ, η_l and η_r respectively. Additionally, $\eta_{\tilde{\phi}}$ denotes the value used to generate the frequency mask for the IPD used to estimate the cone of confusion. The values η_ϕ, η_Ξ, η_l, η_r and $\eta_{\tilde{\phi}}$ produce the frequency masks: $M_\phi(p)$, $M_\Xi(p)$, $\hat{M}_l(p)$, $\hat{M}_r(p)$ and $M_{\tilde{\phi}}(p)$ respectively. For the general case, the mask is denoted by: $M_x(p)$ and the corresponding threshold is denoted by η_x. The mask, $M_x(p)$ is given by: $M_x(p)=1$ for $A(p)>=\eta_x$ and $M_x(p)=0$ for $A(p)<\eta_x$. A certain percentage of samples are needed in order for the localization cues to be useful. If needed, the values of η_ϕ, η_Ξ, η_l, η_r and $\eta_{\tilde{\phi}}$ are lowered to ensure that the percentage of $M_x(p)=1$ is above a certain value. For frequency indices, p corresponding to frequencies outside of the frequency range of $4kHz < f < 18kHz$, $\hat{M}_l(p)=0$, and $\hat{M}_r(p)=0$; for frequency indices, p corresponding to frequencies outside of the frequency range of $0Hz < f < 18kHz$, $M_\phi(p)=0$, $M_\Xi(p)=0$; and for frequency indices, p corresponding to frequencies outside of the frequency range of $0Hz < f < 11kHz$, $M_{\tilde{\phi}}(p)=0$. These values were found experimentally to yield the optimum results. The interaural-polar coordinate system is used to describe the direction of arrival of the sound source. It does so with a lateral angle, $\lambda \in [-90°, 90°]$, and polar angle $\theta \in [-180°, 180°)$. 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 great diversity in the interaural parameters, for a given frequency band. However, for a fixed lateral angle, there is a similarity in the interaural parameters in different regions on the polar dimension, which gives rise to the cone of confusion [1, Chapter 1]. For time-frequency units containing only sound from the direct path within a given frequency band, the interaural parameters should yield the same values for each of these units, irrespective of the sound source, for a given DOA. Because the interaural parameters are sound source agnostic and diverse for differing lateral angles, 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, a HRTF dataset is used to generate training data to estimate the cone of confusion by estimating 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 $\beta \in \{1, 2, 3, ..., \iota\}$ 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: $\hat{\Upsilon}^{\alpha\beta}(p)=\angle(H_l^{\alpha\beta}(p)/H_r^{\alpha\beta}(p)) \in (-\pi,\pi]$. The process applied to the IPD of the test sound, $\phi(p,\hat{m})$ to give $\upsilon(p)$ is also applied to the IPD templates $\hat{\Upsilon}^{\alpha\beta}(p)$ using a value of $\rho=1$ to give $\Upsilon^{\alpha\beta}(p)$. The IPD difference, $\Gamma_\phi^{\alpha\beta}(p)$ is given by: $\Gamma_\phi^{\alpha\beta}(p)=\angle(e^{j\upsilon(p)}e^{-j\Upsilon^{\alpha\beta}(p)})$. 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: $(\alpha_\kappa,\beta_\kappa)=\arg\max_{\alpha\in\{1,...,\gamma\},\beta=\kappa}\sum_p M_{\tilde{\phi}}(p).\ln(\mathcal{N}(\Gamma_\phi^{\alpha\beta}(p)|0,1))$, where $\kappa \in \{1, 2, 3, ..., \iota\}$ are the indices of the cone of confusion, which have corresponding lateral and polar angles $(\alpha_\kappa,\beta_\kappa)$.

The log-magnitude Fourier spectrum of the HRIRs in the training dataset, $\acute{H}_\zeta^{\alpha\beta}(p)$ are found directly by using the same FFT size used to generate each time frame for $\phi(p,m)$: $\acute{H}_\zeta^{\alpha\beta}(p)=10\log_{10}(|H_\zeta^{\alpha\beta}(p)|^2)$, where $H_\zeta^{\alpha\beta}(p)=\mathcal{F}\{h_\zeta^{\alpha\beta}(n)\}$, where $\mathcal{F}$ is the Discrete Fourier Transform (DFT). The MR, $E^{\alpha\beta}(p)$ of each HRIR pair in the training dataset is given by: $E^{\alpha\beta}(p)=2\times(|H_l^{\alpha\beta}(p)|/(|H_l^{\alpha\beta}(p)|+|H_r^{\alpha\beta}(p)|)-1/2)$ In [2], the method used all time-frequency units estimated to be above the noise floor to estimate the spectrum of the direct component of the test sound. One problem with this approach is that some frequency bands contained more time-frequency units estimated to contain a signal above the noise floor than other frequency bands. This had the unintended consequence of yielding higher levels for the spectrum in frequency bands with more estimated time-frequency units containing a signal above the noise floor, due to the number of time-frequency units alone. In the proposed method, for active frequency bands, the same number of time-frequency units are used in each frequency band to estimate the spectrum and the MR of the direct component of the test sound. The MR of the direct component of the test sound is estimated by: $\hat{\Xi}(p)=2\cdot\left(\frac{\sqrt{\sum_{\hat{m}=1}^{g}|Y_l(p,\hat{m})|^2}}{(\sqrt{\sum_{\hat{m}=1}^{g}|Y_l(p,\hat{m})|^2}+\sqrt{\sum_{\hat{m}=1}^{g}|Y_r(p,\hat{m})|^2})}-\frac{1}{2}\right)$. It was found experimentally that $g=10$ yielded the optimum results. If the IPD of a time-frequency sample is not close to the value of υ_p, then it is considered to be highly affected by reflections. Using this knowledge, an additional constraint is used which was found to better estimate the spectrum of the direct component of the sound. Let $\tilde{m}$ denote time indices with $\phi(p,m)$ in the range: $\upsilon(p)-\tau < \phi(p,m) < \upsilon(p)+\tau$, ordered by their correspond-



(a) (b) (c) (d)

Fig. 1: Mean central angular error (°) for (a,b,d) localization on the full-sphere, (c) localization on the median plane, as a function of: (a-c) SNR, (d) sound category with binaural test sound signals generated at 30dB SNR. The results are shown for binaural test sound signals synthetically generated using: (a-c) all of the monaural sound sources, (d) the monaural sound sources specified by the sound category in the abscissa. The proposed method is shown using both variants: Spectral cues only, and spectral/interaural cues, as well as the reference methods: Cross-Convolution, Speech Prefilter, and MUSIC Signal Subspace. The error bars correspond to 95% confidence intervals. The results are shown for: (a,c,d) anechoic condition, (b) reverberant condition.

ing level, $S(p,m)$ in descending order. The direct component of the log-magnitude Fourier spectrum of the test sound is estimated by: $\hat{\Psi}_\zeta(p) = 10\log_{10}(\sum_{\tilde{m}=1}^{g}|Y_\zeta(p,\tilde{m})|^2)$. It was found experimentally that $g = 10$ and $\tau = 0.5^c$ yielded the optimum results. Let $\Psi(p)$ denote the log-magnitude Fourier spectrum of a signal, $\psi(n)$. The same operations are now performed on both the test sound and the HRIR pairs on the estimated cone of confusion, for both the left and right channels. To identify the peaks and notches in the spectrum, the rapid fluctuations present in the spectrum of the HRTFs and 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 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 to those in the HRTF templates then, this slow varying component needs to be removed [2, 14]. The real cepstrum of the signal, $\psi(n)$ is given by: $c_v = \mathcal{F}^{-1}\{\Psi(p)\}$, where $v \in \{0,...,N-1\}$ are the indices of the cepstral coefficients, c_v in the cepstral domain. Consider that applying a low-pass lifter in the cepstral domain removes the rapid fluctuations in the log-magnitude Fourier spectrum. 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 low-pass lifter, multiple linear regression can be used to estimate the cepstral coefficients. The indices $\hat{p}_{B,\zeta}$ correspond to frequency indices where $\hat{M}_\zeta(p) = 1$. To find the cepstral coefficients, $a = [c_0 \ . \ . \ . \ c_u]^T$, we generate a model matrix as: $X = [1 \quad 2\cos(\omega(\hat{p}_{B,\zeta})) \quad 2\cos(2\omega(\hat{p}_{B,\zeta})) \ . \ . \ . \ 2\cos(u\omega(\hat{p}_{B,\zeta}))]$, where $\omega(p)$ is normalized discrete frequency, defined as $\omega(p) = \pi p/J$, and J is the frequency index corresponding to the Nyquist frequency. The coefficient estimates, $\hat{a}$ can be obtained by the usual least squares method: $\hat{a} = (X^TX)^{-1}X^T.\Psi(\hat{p}_{B,\zeta})$, where X^T is the transpose of X [15]. The rapid fluctuations of the log-magnitude Fourier spectrum are removed by reconstructing the log-magnitude Fourier spectrum without using the higher order coefficients: $\check{\Psi}(p) = 2\sum_{v=1}^{u} c_v\cos(v\omega(p))$. It should be noted that c_0 is not used in the reconstruction of the log-magnitude Fourier spectrum, resulting in $\check{\Psi}(p)$ having a mean of zero. It was experimentally found that $u = 25$ yielded the optimum results. Let $\Xi(p)$ denote the MR of a signal, $\psi(n)$. The same procedure performed on $\Psi(p)$ to yield $\check{\Psi}(p)$ is performed on $\Xi(p)$ to yield $\check{\Xi}(p)$. For the specific cases, let $\check{\Xi}(p)$ be denoted by $\grave{\Xi}(p)$ and $\grave{E}^\kappa(p)$ for the test sound and HRIR pairs on the cone of confusion respectively. The authors in [14] note that the first peak in the spectrum of the HRTF acts as a reference position for the other spectral peaks and notches. The proposed method uses the same principle. The lower limit for the frequency range for the spectrum, $\hat{f}_\zeta$ is chosen as the frequency that has a corresponding frequency index which yields the maximum level in $\check{\Psi}(\hat{p}_{B,\zeta})$ for the test sound in the 4kHz - 6kHz frequency range. 4kHz - 6kHz is the approximate frequency range of the first peak in the spectrum of the HRTFs. This lower limit, which is found for the spectrum of the test sound is also used for the HRTF templates. The mask $M_\zeta(p)$ has the same values as the mask $\hat{M}_\zeta(p)$, with the exception that for frequency indices, p corresponding to frequencies outside of the frequency range of $\hat{f}_\zeta < f < 18kHz$, $M_\zeta(p) = 0$. The indices $p_{B,\zeta}$ correspond to frequency indices where $M_\zeta(p) = 1$. The slow varying component of $\check{\Psi}(p)$ is denoted by: $\check{\Psi}^s(p)$. It is found by simple linear regression of $\check{\Psi}(p_{B,\zeta})$ using the ordinary least squares method. This slow varying component is removed to give $\grave{\Psi}(p)$, so that the peaks and notches of the spectra of the test sound and HRTF templates can be compared, i.e. $\grave{\Psi}(p) = \check{\Psi}(p) - \check{\Psi}^s(p)$. For the specific cases, let $\grave{\Psi}(p)$ be denoted by $\grave{Y}_\zeta(p)$ and $\grave{H}^\kappa_\zeta(p)$ for the test sound and HRIR pairs on the cone of confusion respectively. The spectral difference, $\hat{\Gamma}^\kappa_\zeta(p)$ is given by: $\hat{\Gamma}^\kappa_\zeta(p) = \grave{Y}_\zeta(p) - \grave{H}^\kappa_\zeta(p)$. Let $\sigma\{.\}$ denote the standard deviation operation. The normalized spectral difference is given by: $\Gamma^\kappa_\zeta(p) = \grave{Y}_\zeta(p)/\sigma\{\grave{Y}_\zeta(p_{B,\zeta})\} - \grave{H}^\kappa_\zeta(p)/\sigma\{\grave{H}^\kappa_\zeta(p_{B,\zeta})\}$, and the normalized MR difference is given by:$\Gamma^\kappa_\Xi(p) = \grave{\Xi}(p)/\sigma\{\grave{\Xi}(p_D)\} - \grave{E}^\kappa(p)/\sigma\{\grave{E}^\kappa(p_D)\}$, where the indices p_D correspond to frequency indices where $M_\Xi(p) = 1$. Let $\hat{\lambda}$ denote the median of the lateral angles of the estimated entries on the cone of confusion, κ. The lateral angles, $|\lambda| < 10^\circ$ are considered to be close to the median plane, such that the interaural cues are unreliable. For this region, only the spectral cues are used, and the weighting is equal for both ears. The non-normalized spectra are used for localization in this region, as the spectra of HRTFs with DOAs above the listener are relatively flat. Normalization of the spectra of the test sound at these DOAs places an emphasis on the spectral shape of the noise rather than the HRTF, which yields a confounding result for the estimated DOA of the test sound. For DOAs with a lateral angle away from the median plane, the DOA of the test sound is better estimated using the normalized spectra, as in these regions there isn't a HRTF with a flat spectrum and normalization allows for a better comparison of the shape of the spectra of the test sound

and HRTF templates. For $|\hat{\lambda}| < 10^{\circ}$ the log-likelihood distribution, $Z(\beta_\kappa)$ of each polar angle on the cone of confusion, β_κ is given by: $Z(\beta_\kappa) = 0.5 \times \sum_{\zeta \in \{l,r\}} \sum_p M_\zeta(p).\ln(\mathcal{N}(\hat{\Gamma}^\kappa_\zeta(p)|0,1))$. The other regions are: $(V \times 10)^{\circ} <= |\hat{\lambda}| < ((V+1) \times 10)^{\circ}$, for $V < 8$, and $(V \times 10)^{\circ} <= |\hat{\lambda}| <= ((V+1) \times 10)^{\circ}$, for $V = 8$, where $V \in \{1, ..., 8\}$. The channel index is now denoted in terms of the ipsilateral ear, I and contralateral ear, K, i.e. $\zeta \in \{I, K\}$. For $\hat{\lambda} >= 0^{\circ}$, the left ear is the ipsilateral ear and the right ear is the contralateral ear, and for $\hat{\lambda} < 0^{\circ}$, the right ear is the ipsilateral ear and the left ear is the contralateral ear. For $|\lambda| >= 10^{\circ}$ it was found that the spectral cue of the contralateral ear provided confounding information due to the high amount of reverberation and noise present in its spectrum. Using this information, for $|\hat{\lambda}| >= 10^{\circ}$ the weights: w_I^V, w_Ξ^V, w_ϕ^V are used to weight the normalized spectral difference of the ipsilateral ear, $\Gamma_I^\kappa(p)$, the normalized MR difference, $\Gamma_\Xi^\kappa(p)$ and the IPD difference, $\Gamma_\phi^\kappa(p)$ respectively to create the log-likelihood distribution, $Z(\beta_\kappa)$ of each polar angle on the cone of confusion, β_κ. For $|\hat{\lambda}| >= 10^{\circ}$, the log-likelihood distribution is given by: $Z(\beta_\kappa) = \sum_{x \in \{I,\Xi,\phi\}} w_x^V \cdot \sum_p M_x(p) \,.\, \ln(\mathcal{N}(\Gamma_x^\kappa(p)|0,1))$. The HRTF template index, $\hat{\kappa}$ corresponding to the estimated lateral and polar angle of the test sound is given by: $\hat{\kappa} = \arg\max_\kappa(Z(\beta_\kappa))$. A brute force approach is taken to learn these weights in each region, V, using validation data consisting of binaural sounds generated with the second tokens of the environmental sounds in [16], and 10 speech samples not included in the test data from the CSTR VCTK corpus [17]. The HRTF dataset used to generate the validation data is the dataset measured at IRCAM in 2014 as part of the "Club Fritz" project [18]. The testing condition in the method shown above uses the interaural and spectral cues for localization on the full-sphere. A second testing condition uses only the spectral cues to resolve the cone of confusion, i.e. $w_\Xi^V = 0$, and $w_\phi^V = 0$.

3. TESTING PROCEDURE

In this paper, the proposed method is used to estimate the cone of confusion, $(\alpha_\kappa, \beta_\kappa)$. The reference methods then estimate which of the DOAs, $(\alpha_\kappa, \beta_\kappa)$ is the true DOA of the sound source. In this paper, slight modifications are made to the reference methods to improve their robustness to the testing conditions, while the spirit of the methods are retained. The reference methods are the Cross-Convolution method [4, 3], the Speech Prefilter method [5] and the MUSIC (MUltiple SIgnal Classification) Signal Subspace method [6]. Our implementation of these methods is described in [19]. In order to test the robustness of the localization methods, a diverse range of monaural sound sources are used to generate the binaural test sound signals. These include 10 environmental sounds taken from [16], 10 speech samples chosen from the CSTR VCTK corpus [17], white noise and pink noise. The HRTF dataset used to generate training data for all conditions is the Gauss-Legendre 2° dataset, measured in [20]. The HRTF dataset used to generate the binaural test sound signals for the anechoic condition is the dataset measured at RIEC, Tohoku University as part of the "Club Fritz" project [18]. For the reverberant condition the binaural test sound signals are produced using the BRIRs from the dataset measured in [21], with the dummy head facing the front of the room. A full sphere localization test is conducted using the DOAs of the loudspeakers in [21] for the reverberant condition. For the anechoic condition, the HRIR pairs in the RIEC dataset with DOAs nearest to the DOAs used in the reverberant condition are used. Additionally, a median plane localization test is conducted using the RIEC dataset and the 22 monaural sound sources to generate the binaural test sound signals. For this condition, 35 HRIR pairs are used, all of which lie on the median-plane and are spaced in 10° increments, with the exception of a HRIR pair at $\theta = -90^{\circ}$, which is absent. For the median plane localization test, as shown in Figure 1c, only HRTF templates that lie on the median plane are used for training data. These HRTF templates are directly used as the entries for the cone of confusion. The binaural test sound signals are created synthetically by convolving the HRIR pairs at each of the test positions with each of the 22 monaural sound sources. Stereo uncorrelated pink noise is added to the binaural test sound signals to give signal-to-noise ratios (SNRs) from 0dB to 30dB in 10dB steps.

4. RESULTS AND DISCUSSION

For all methods, 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 [22]. The proposed method using interaural and spectral cues outperforms the reference methods in all testing conditions. The cross-convolution method performs poorly in most testing conditions. This could be due to the method only using an interaural comparison, and not using any technique to be robust to noise or reverberation. Figure 1a and Figure 1b show the mean central angular error as a function of SNR for localization on the full-sphere, for the anechoic and reverberant testing conditions respectively. It can be seen that the proposed method using the interaural and spectral cues outperforms the proposed method using only spectral cues, showing that the interaural cues can improve localization in the polar dimension. For the proposed method, the results for the anechoic conditions are similar to those in the reverberant condition, showing that the method is robust to the presence of reverberation. Figure 1c shows the mean central angular error as a function of SNR for localization of sound sources on the median plane. Around the median plane, a slight positioning error when measuring HRTFs can result in a large difference in the interaural cues between different HRTF datasets. The Cross-Convolution, and MUSIC Signal Subspace methods both implicitly use interaural cues for localization, which results in large errors for this testing condition. The proposed method and the Speech Prefilter method both use spectral cues only for median plane localization, resulting in more accurate localization estimates in this region. Figure 1d shows the mean central angular error as a function of sound category. The Speech Prefilter method is fairly robust against speech, though the method trains a prefilter as the average speech spectrum in an attempt to remove the slow varying component of the spectrum, which can result in higher errors for speech sounds that have spectra that do not resemble the average speech spectrum. It can be seen that the proposed method is robust to the different sounds. One reason is because the proposed method uses linear regression to estimate the slow varying component in the spectrum, and as such is more generalizable to different sound types.

5. CONCLUSION

The proposed method outperforms the state of the art binaural sound source localization methods in all testing conditions. The proposed method is robust to reverberation, additive noise and sound category. Using spectral cues only for localization of sound sources on the median plane allows for better localization estimates in this region. Away from the median plane, it was found that incorporating interaural parameters increases the performance of the method.

6. REFERENCES

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

[2] B. R. Hammond and P. J. B. Jackson, "Robust full-sphere binaural sound source localization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 86–90.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] Y. Luo, D. N. Zotkin, and R. Duraiswami, "Gaussian process models for HRTF based 3d sound localization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2858–2862.

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

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

[11] H. Ivanka, K. Jan, and Z. Jiri, *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing*, World Scientific Publishing Co Pte Ltd, 2012.

[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] 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.

[15] 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.

[16] 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.

[17] C. Veaux, J. Yamagishi, K. MacDonald, et al., "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017.

[18] A. Andreopoulou, D. R. Begault, and B. F. G Katz, "Inter-laboratory round robin HRTF measurement comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 895–906, 2015.

[19] B. R. Hammond and P. J. B. Jackson, "Robust median-plane binaural sound source localization," in *DCASE 2018 Workshop: Workshop on Detection and Classification of Acoustic Scenes and Events*, Nov 2018.

[20] 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.

[21] 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.

[22] 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.