BINAURAL LOCALIZATION OF SPEECH SOURCES IN THE MEDIAN PLANE USING CEPSTRAL HRTF EXTRACTION

Dumidu S. Talagala, Xiang Wu, Wen Zhang and Thushara D. Abhayapala

Research School of Engineering, CECS, The Australian National University. Canberra. Australia.
Email: {dumidu.talagala, xiang.wu, wen.zhang, thushara.abhayapala}@anu.edu.au

ABSTRACT

In binaural systems, source localization in the median plane is challenging due to the difficulty of exploring the spectral cues of the head-related transfer function (HRTF) independently of the source spectra. This paper presents a method of extracting the HRTF spectral cues using cepstral analysis for speech source localization in the median plane. Binaural signals are preprocessed in the cepstral domain so that the fine spectral structure of speech and the HRTF spectral envelope can be easily separated. We introduce (i) a truncated cepstral transformation to extract the relevant localization cues, and (ii) a mechanism to normalize the effects of the time varying speech spectra. The proposed method is evaluated and compared with a convolution based localization method using a speech corpus of multiple speakers. The results suggest that the proposed method fully exploits the available spectral cues for robust speaker independent binaural source localization in the median plane.

Index Terms— Binaural localization, cepstral transformation, head related transfer function (HRTF), median plane.

1. INTRODUCTION

The human auditory system localizes a sound source by exploring the interaural time/level differences (ITD/ILD) and the monaural spectral cues of binaural signals received at the ears [1]. The acoustic transfer function, i.e., the head-related transfer function (HRTF), encompasses these cues and has been widely adopted to design binaural source localization systems. In general, HRTF based localization algorithms identify a source location by maximizing the correlation between the binaural signals [2], or the ITD/ILD estimates [3], in the range of possible source locations. Although this approach is known to be robust in the horizontal (azimuth) plane dominated by interaural cues [4], work on median (elevation) plane localization has been limited and challenging, due to the diminishing interaural differences and the dominance of the spectral localization cues [5, 6].

Human perceptual experiments have shown that elevation is perceived as resonances (peaks) and cancellations (notches) of certain frequencies, which are mainly caused by scattering and diffraction of sound waves in the pinna at high frequencies [7]. Numerous methods from parametric models [7, 8] to finite-difference time-domain models of the head [9] have been proposed to map the frequencies of these notches to the source location. However, the accuracy of this mapping is critically affected by perturbations in elevation and the reverberant nature of the acoustic environment. In addition, the more robust algorithms [10] are also more computationally complex, due to the nature of the HRTF frequency response and the influence of the source spectra on the localization cues.

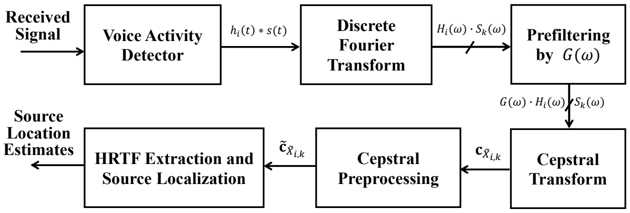

Received Signal
Voice Activity Detector
$h_i(t) * s(t)$
Discrete Fourier Transform
$H_i(\omega) \cdot S_k(\omega)$
Prefiltering by $G(\omega)$
$G(\omega) \cdot H_i(\omega) \cdot S_k(\omega)$
Cepstral Transform
$\mathbf{c}_{\tilde{X}_{i,k}}$
Cepstral Preprocessing
$\tilde{\mathbf{c}}_{\tilde{X}_{i,k}}$
HRTF Extraction and Source Localization
Source Location Estimates

Fig. 1. The system block diagram. $\widetilde{\mathbf{c}}_{\widetilde{X}_{i,k}}$, $\mathbf{c}_{\widetilde{X}_{i,k}}$ are the cepstral domain representations of a received signal $h_i(t) * s(t)$ in the k^{th} time frame. $G(\omega)$ is the prefilter described in (4) and (9).

In this paper, we present a method of localizing speech sources in the median plane, by extracting the HRTF spectral cues through a cepstral transformation of the binaural signals that can minimize the influence of the source spectra. Two speech characteristics needed to be considered: i) the majority of the energy is concentrated at low frequencies, where spectral cues are negligible and ii) the non-stationary nature of speech. The first characteristic suggests that speech can be fairly well localized in the median plane, due to less variability at frequencies above 3–4 kHz, and the reduced likelihood of the speech spectrum contaminating the spectral cues. The second characteristic suggests a short-time approach, i.e., a short-time Fourier transform, is best suited to model the variability of the speech spectrum. In this context, the proposed cepstral processing of binaural signals retains both the fine low frequency structure of the speech source and the HRTF magnitude information (spectral envelope) at two distinctive parts of the cepstrum. This, together with the logarithmic operation of the cepstral transform that acts as a weighting function at higher frequencies, enables a simpler separation of the two components and the use of the enlarged fluctuations in the HRTF magnitude response for localization. Finally, the cepstral prefiltering concept has been shown to be robust to the effects of reverberation in previous studies [11, 12], suggesting that the proposed method may also be suitable for localization in reverberant conditions.

2. SYSTEM MODEL

The received signal at each of the binaural receivers is a convolution of the source signal $s(t)$ and the corresponding head related impulse response $h_i(t)$ for $i \in \{l, r\}$ representing the left and right ears. Thus, for a single active source, the received signal can be expressed in the frequency domain as

$$X_{i,k}(\omega) = H_i(\omega) \cdot S_k(\omega) + N_{i,k}(\omega), \tag{1}$$

where $X_i(\omega)$, $H_i(\omega)$ and $S(\omega)$ represent the received signal, HRTF and source spectra at a frequency ω. $N_{i,k}(\omega)$ represents the additive noise term and $k = 1 \ldots K$ represents the frame number. The speech signals are separated into K frames, such that the frame length is less than the stationary time duration of the signal, typically 10–50 ms for voiced speech [13].

In this study, we consider the source signals to be pure human speech; thus, the majority of the energy in the received signals at the ears are concentrated below 4 kHz (generally the formant frequencies) during voiced speech [13]. However, the localization cues relevant to the median plane are prevalent at frequencies above 3 kHz [5, 6], and the HRTF features must therefore be extracted from the high frequency region of the received signal spectrum. The cepstral preprocessing approach we propose to extract these features and localize a source is illustrated in Fig. 1 and described below.

2.1. Cepstral Transformation

The transformation of a signal into the cepstral domain is an inverse Fourier transform of the absolute magnitude spectrum of that signal. Thus,

$$\mathbf{c}_{X_{i,k}} \triangleq \mathcal{F}^{-1} \left\{ \log_{10} |\mathbf{x}_{i,k}| \right\}, \tag{2}$$

where $\mathbf{x}_{i,k} = [\ X_{i,k}(0),\ \cdots,\ X_{i,k}(\omega_{\max})\ ]$ is the signal spectrum and $\mathcal{F}$ represents the Discrete Fourier Transform (DFT). The cepstral transformation of (1) can be expressed approximately as a sum of the cepstral coefficients of the HRTF, speech and noise components respectively, due to the logarithmic operation of the cepstral transform, and is defined as [11, 12]

$$\mathbf{c}_{X_{i,k}} \triangleq \mathbf{c}_{H_i} + \mathbf{c}_{S_k} + \mathbf{c}_{\widetilde{N}_{i,k}}, \tag{3}$$

where $\mathbf{c}_{\widetilde{N}_{i,k}} \triangleq \mathcal{F}^{-1} \left\{ \log_{10} |1 + N_{i,k} / \left[H_i(\omega) S_k(\omega) \right]| \right\}$.

The magnitude response of the HRTF could therefore be reconstructed by extracting the cepstral coefficients corresponding to $\mathbf{c}_{H_i}$. In practise however, the influence of the speech spectrum will distort the reconstructed HRTF, and requires a statistical normalization of the effects of speech. Hence, to reduce this distortion, we introduce a prefilter $G(\omega)$, as shown in Fig. 1 (the derivation of G is described in Section 2.3). In order to simplify the derivation we will ignore the effect of noise ($\mathbf{c}_{\widetilde{N}_{i,k}} \rightarrow \mathbf{0}$), but will include and evaluate its effect on the localization performance in Section 4. Thus, the prefiltered received signals can be simplified as

$$\widetilde{X}_{i,k}(\omega) = G(\omega) \cdot H_i(\omega) \cdot S_k(\omega). \tag{4}$$

The corresponding cepstral domain representation becomes

$$\mathbf{c}_{\widetilde{X}_{i,k}} = \mathbf{c}_G + \mathbf{c}_{H_i} + \mathbf{c}_{S_k}, \tag{5}$$

where $\mathbf{g} = [\ G(0),\ \cdots,\ G(\omega_{\text{max}})\ ]$ is the spectrum of the prefilter and $\mathbf{c}_G \triangleq \mathcal{F}^{-1} \left\{ \log_{10} |\mathbf{g}| \right\}$.

2.2. Truncation of Cepstral Coefficients

The lower order cepstral coefficients in (3) typically model the envelope of the received signal spectrum, while the higher order coefficients model its rapid spectral fluctuations. In the case of speech sources, the higher order coefficients are predominantly speech information [13] corresponding to the pitch and formant structure of a particular speech frame. The spectral envelope of the HRTF could therefore be extracted by appropriately truncating the cepstral coefficients in (5).

Thus, we define a truncation operation $\mathcal{T}\{\cdot\}$, that retains sufficient HRTF magnitude information for source localization. The truncated cepstral coefficients now become

$$\widetilde{\mathbf{c}}_{\widetilde{X}_{i,k}} \triangleq \mathcal{T} \left\{ \mathbf{c}_{\widetilde{X}_{i,k}} \right\} = \mathcal{T} \left\{ \mathbf{c}_G \right\} + \mathcal{T} \left\{ \mathbf{c}_{H_i} \right\} + \mathcal{T} \left\{ \mathbf{c}_{S_k} \right\}, \tag{6}$$

where the truncation order is determined by the number of cepstral coefficients required to model the magnitude response of $h_i(t)$.

2.3. Cepstral Preprocessing and Speech Normalization

Including the effects of the cepstral truncation operation described in the previous section, (6) can be expressed as

$$\widetilde{\mathbf{c}}_{\widetilde{X}_{i,k}} = \mathbf{c}_G + \widehat{\mathbf{c}}_{H_i} + \mathcal{T} \left\{ \mathbf{c}_{S_k} \right\}, \tag{7}$$

where $\widehat{\mathbf{c}}_{H_i} \triangleq \mathcal{T} \left\{ \mathbf{c}_{H_i} \right\}$ and $\mathbf{c}_G \triangleq \mathcal{T} \left\{ \mathbf{c}_G \right\}$. $\widehat{\mathbf{c}}_{H_i}$ characterizes the cepstral approximation of the HRTF magnitude response, i.e., the spectral envelope, but source localization requires the estimation of $\widehat{\mathbf{c}}_{H_i}$ in the presence of time varying speech.

As stated previously, we introduce the prefilter $G(\omega)$ to normalize the effects of the speech component $\mathcal{T} \left\{ \mathbf{c}_{S_k} \right\}$, such that

$$\mathbf{c}_G + E \left[\mathcal{T} \left\{ \mathbf{c}_{S_k} \right\} \right] = \mathbf{c}_0, \tag{8}$$

where $\mathbf{c}_0 = [\ c_0,\ 0,\ \cdots,\ 0\ ]$ is a constant vector and $E\left[\cdot\right]$ is the expectation operator. c_0 is an arbitrary constant, selected such that $G(\omega)$ normalizes the distribution of speech energy across frequency, resulting in only the zero-th order cepstral coefficient. Thus, the prefilter can be expressed as

$$\mathbf{g} = 10^{\mathcal{F}\{\mathbf{c}_0 - \mathbf{c}_S\}}, \tag{9}$$

where $\mathbf{c}_S \triangleq E\left[\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\right]$ is obtained empirically from the analysis of speech data obtained from multiple speakers.

3. SOURCE LOCATION ESTIMATION

3.1. Extracting the HRTF Spectral Envelope

The design of the prefilter and the behaviour of the speech spectrum can now be exploited to extract the truncated cepstral HRTF coefficients $\widehat{\mathbf{c}}_{H_i}$ as shown below. For a particular speaker, the expectation of (7) can be expressed as

$$\begin{aligned}\widehat{\mathbf{c}}_{\widetilde{X}_i} &\triangleq \frac{1}{K}\sum_{k=1}^{K}\widetilde{\mathbf{c}}_{\widetilde{X}_{i,k}} = \mathbf{c}_G + \widehat{\mathbf{c}}_{H_i} + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\\ &= \mathbf{c}_0 + \widehat{\mathbf{c}}_{H_i} - \mathbf{c}_S + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}, \end{aligned}\tag{10}$$

where $\mathbf{c}_G = \mathbf{c}_0 - \mathbf{c}_S$ from (8). We exploit the property of speech, where for sufficiently large K the spectrum approaches a general distribution [14, 15], such that

$$\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left\{\mathbf{c}_{S_k}\right\} \rightarrow E\left[\mathcal{T}\left\{\mathbf{c}_{S_k}\right\}\right] = \mathbf{c}_S. \tag{11}$$

Thus, (10) becomes

$$\widehat{\mathbf{c}}_{\widetilde{X}_i} \approx \mathbf{c}_0 + \widehat{\mathbf{c}}_{H_i}, \tag{12}$$

where $\mathbf{c}_0$ is zero for all but the first element, i.e., a cepstral representation of a uniform spectrum. The HRTF spectral envelope can therefore be extracted by applying an inverse cepstral transformation to (12), and is given by

$$\widehat{\mathbf{h}}_i = 10^{\mathcal{F}\left\{\widehat{\mathbf{c}}_{\widetilde{X}_i} - \mathbf{c}_0\right\}}. \tag{13}$$

3.2. Sound Source Localization

In this paper, the correlation between the extracted HRTF spectral envelope and the HRTFs in a pre-measured database, is adopted as a metric to determine the actual source location. However, since the localization cues that differentiate the source locations in the median plane are both subtle and predominantly located at higher frequencies, we first combine the extracted HRTF spectral envelopes obtained from the received binaural signals in (13) for the relevant frequency range $\omega \in \{\omega_a, \ldots, \omega_b\}$. Thus, the binaural "estimated HRTF spectral envelope" becomes

$$\widehat{\mathbf{h}} = \left[\ \widehat{H}_l(\omega_a),\ \cdots,\ \widehat{H}_l(\omega_b),\ \widehat{H}_r(\omega_a),\ \cdots,\ \widehat{H}_r(\omega_b)\ \right], \tag{14}$$

where $\widehat{\mathbf{h}}_i = \left[\ \cdots,\ \widehat{H}_i(\omega_a),\ \cdots,\ \widehat{H}_i(\omega_b),\ \cdots\ \right]$.

The spectral envelopes of the median plane HRTFs in the database can be expressed in a similar fashion, by applying the cepstral truncation operation $\mathcal{T}$. Thus, the binaural "evaluated HRTF spectral envelope" in the direction Θ becomes

$$\mathbf{h}(\Theta) = \left[\ \widetilde{H}_l(\Theta,\omega_a),\ \cdots,\ \widetilde{H}_l(\Theta,\omega_b),\ \widetilde{H}_r(\Theta,\omega_a),\ \cdots\ \right], \tag{15}$$

where $\mathbf{h}_i(\Theta) \triangleq 10^{\mathcal{F}\left\{\widetilde{\mathbf{c}}_{H_i}(\Theta)\right\}} = \left[\ \cdots,\ \widetilde{H}_i(\Theta,\omega),\ \cdots\ \right]$.

The sample cross-correlation can therefore be used to calculate the source localization spectrum, i.e., the correlation between (14) and (15) for all possible source locations, as

$$P(\Theta) = \widehat{\mathbf{h}} \oplus \mathbf{h}(\Theta), \tag{16}$$

where $\oplus$ denotes the cross-correlation operation. Hence, for a single active sound source, the estimated source location in the median plane can be determined by evaluating (16) for all Θ, and is given by

$$\widehat{\Theta} = \arg\max_{\Theta}\ \{P(\Theta) \geq \gamma\}. \tag{17}$$

Considering the possible influence of noise and the case of no active sound sources, an appropriate threshold is chosen, i.e., $\gamma = \max\{0.95 \cdot \max\{P(\Theta)\}, 0.5\}$, to identify the peak of the spectrum and the relevant source location.

4. EVALUATION

4.1. Simulation Setup

We evaluate the performance of the proposed localization technique in the median plane through simulations, using MIT's HRTF measurement database of the KEMAR dummy-head [16]. The simulated clean speech signals are obtained from a corpus of 34 speakers, each with 600 utterances sampled at 16 kHz, consisting of a sequence of words in a sentence. The speech data is obtained from the sample recordings used in the "PASCAL 'CHiME' Speech Separation and Recognition Challenge" [17], and the binaural signals are produced by convoluting the KEMAR head related impulse responses with these speech signals. We apply a voice activity detector to identify the voiced speech frames, and perform the Fourier and cepstral transform operations using a 20 ms Hamming window at 10 ms intervals, i.e., a window length of 320 samples corresponding to the sampling rate of 16 kHz. The prefilter $\mathbf{g}$ is computed from the average speech cepstrum as per (11), and is used as a common prefilter for all 34 speakers.

The localization performance is compared with the state-of-the-art convolution based binaural localization scheme described in [18]; a superior, more noise-robust variant of the Source Cancellation Algorithm [2] and the classical matched filtering approach [19]. We evaluate the performance in the 3.5–7.5 kHz audio bandwidth, where spectral cues are known to dominate [5], at the indicated Signal to Noise Ratios (SNRs). The effect of reverberation is not explicitly considered, but the cepstral methods are known to be robust to its effects [11, 12], due to the truncation operation's removal of the rapid spectral fluctuations. The empirical threshold used

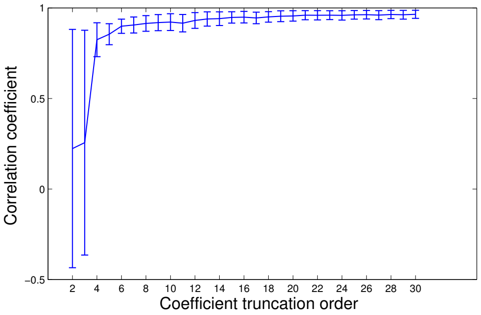


Fig. 2. Correlation between the truncated approximation and the actual HRTF, for source locations in the median plane. It should be noted that the correlation peaks and falls with increasing truncation order (not shown in the figure).

to identify a source detection is common to both techniques, and is described in (17).

4.2. Effects of Cepstral Preprocessing

The effects of truncating the HRTF magnitude response in the cepstral domain, as described in Section 2.2, is presented in Fig. 2. The mean correlation and standard deviation between the actual HRTF and the truncated approximation of the HRTF in the 3.5–7.5 kHz audio bandwidth is summarized for the KEMAR's median plane source locations. As expected, we observe that increasing the cepstral truncation order improves the correlation, due to the inclusion of the rapid fluctuations in the HRTFs. However, our objective is to extract a smooth form of the spectral cues in the magnitude response of the HRTF, i.e., model the general shape of the HRTF with a sufficiently small number of coefficients. We observe that the mean correlation stabilizes beyond approximately the first 25 cepstral coefficients, and therefore use a cepstral truncation order of 25 to extract the HRTF spectral envelopes of the binaural signals.

4.3. Single Source Localization Performance

Fig. 3 illustrates the comparison of the source localization spectra for the convolution based and proposed methods for a single trial at 40 dB SNR. The vertical dashed line at 10° indicates the actual source location in the median plane. Although the peak correlation for both techniques correspond to the actual source location, it can be observed that the convolution based method provides a flatter source localization spectrum, in contrast to the distinctive peak of the proposed method. Naturally, the flat spectrum increases the uncertainty of the estimated source location, i.e., the confidence interval of the estimate. For simplicity, we use a single standard deviation of the distribution of the detected source locations as a metric to quantify this uncertainty, and is denoted by the error bars in the subsequent figures.

Fig. 4 illustrates the average source localization performance for a source located in the median plane at 10° intervals between -40° and 220°, in the 3.5–7.5 kHz audio bandwidth. 0°, 90° and 180° indicate the source locations directly in front, above and behind the KEMAR in the median plane. The results are the averages of multiple trials using different speech segments of the speakers in the 'CHiME' speech corpus for each source location. The dashed line indicates the actual median plane source locations in the different experiment scenarios, and is the performance benchmark the localization algorithms should ideally follow.

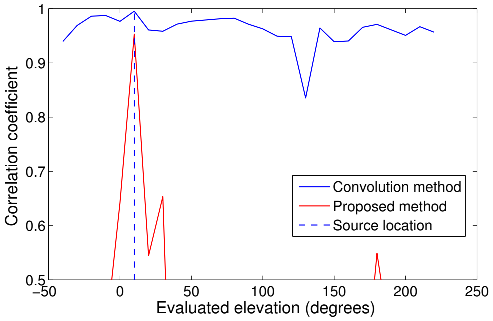
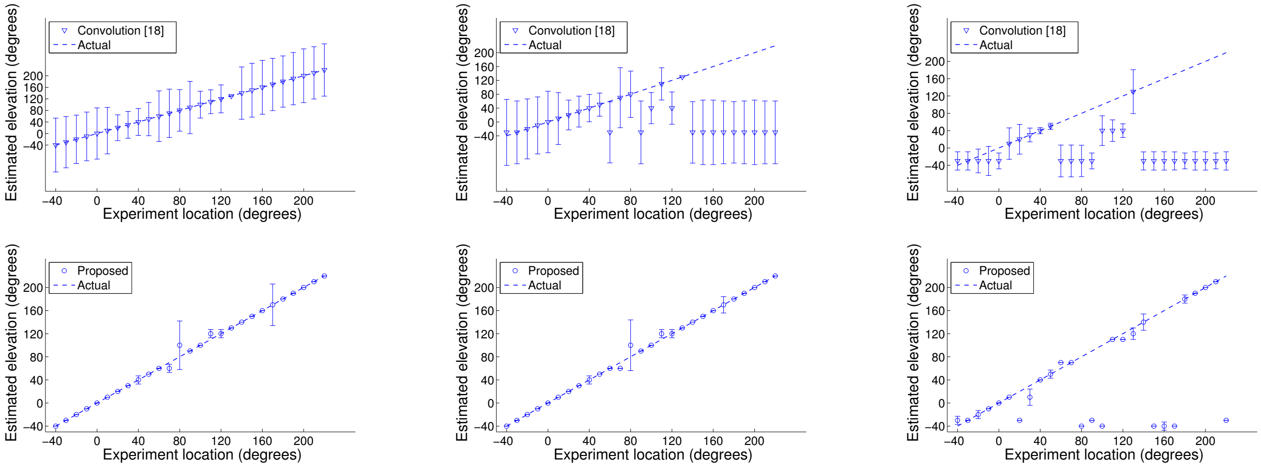


Fig. 3. Source localization spectra of the proposed and convolution based methods for a source at 10° in the median plane.

The localization performance results in Fig. 4(a) show similar localization accuracy for both techniques, but a much greater uncertainty of the estimated location for the convolution based method at 40 dB SNR. In general, decreasing SNR in Fig. 4(b) and (c) results in a degradation in the performance, but the proposed method is shown to be superior and more robust to the effects of noise. Crucially, the uncertainty of the source location estimate of the proposed method is not visibly affected. However, at 20 dB SNR in Fig. 4(c), the performance of the proposed method diverges from the ideal benchmark in the region directly above and behind the KEMAR head. This is consistent with the known localization capability of humans [1], and is mainly due to the lack of rich spectral cues in these regions. The greater uncertainty of the convolution based method is primarily due to it favouring the higher energy region of the signal spectrum, and is a deficiency that is exacerbated with decreasing SNR. Overall, the greater accuracy of the proposed method at low SNR indicates a more efficient exploitation of the spectral localization cues for binaural localization in the median plane.

5. CONCLUSION

In this paper, we present an effective technique to extract the HRTF spectral cues using cepstral processing for binaural source localization in the median plane. We introduce the concept of truncating the binaural signals and HRTF data in the cepstral domain, and show that a small number of cepstral components retain the key localization information. Further, we show that the variable spectral characteristics of speech can be normalized in a straightforward manner to further reduce its influence on spectral cue extraction.



(a) SNR = 40 dB

(b) SNR = 30 dB

(c) SNR = 20 dB

Fig. 4. Average single source localization performance in the 3.5–7.5 kHz audio bandwidth at (a) 40 dB, (b) 30 dB and (c) 20 dB SNR. The figures indicate the source location estimate and the level of uncertainty of the estimate for different trials, where the source is located at the actual elevations between -40° and 220° in the median plane.

The median plane localization performance of the proposed method is compared with the state-of-the-art convolution based approach, and is shown to produce a more confident, noise-robust location estimate. This implies that the proposed method effectively extracts the spectral cues in the HRTF to achieve speaker independent binaural source localization in the median plane. Future work will investigate extending the concept of cepstral HRTF extraction to multiple source localization, as well as localization in both azimuth and elevation.

REFERENCES

[1] J. C. Middlebrooks and D. M. Green, “Sound localization by human listeners,” *Annu. Rev. Psychol.*, vol. 42, no. 1, pp. 135–159, 1991.

[2] F. Keyrouz and K. Diepold, “An enhanced binaural 3D sound localization algorithm,” in *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Darmstadt, Germany, 2006, pp. 662–665.

[3] M. Raspaud, H. Viste, and G. Evangelista, “Binaural source localization by joint estimation of ILD and ITD,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[4] J. Woodruff and D. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[5] J. Blauert, “Sound localization in the median plane,” *Acustica*, vol. 22, pp. 205–213, 1969/70.

[6] B. Rakerd, W. M. Hartmann, and T. L. McCaskey, “Identification and localization of sound sources in the median sagittal plane,” *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2812–2820, 1999.

[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,” *Applied Acoustics*, vol. 68, no. 8, pp. 835–850, 2007.

[8] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, “Sound localization for humanoid robots - building audio-motor maps based on the HRTF,” in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, 2006, pp. 1170–1176.

[9] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida, “Mechanism for generating peaks and notches of head-related transfer functions in the median plane,” *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3832–3841, 2012.

[10] D. S. Talagala, W. Zhang, and T. D. Abhayapala, “Broadband DOA estimation using sensor arrays on complex-shaped rigid bodies,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1573–1585, 2013.

[11] A. Stéphenne and B. Champagne, “A new cepstral prefiltering technique for estimating time delay under reverberant conditions,” *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[12] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, “Cepstrum prefiltering for binaural source localization in reverberant environments,” *IEEE Signal Processing Lett.*, vol. 19, no. 2, pp. 99–102, 2012.

[13] P. Taylor, *Text-to-Speech Synthesis*, Cambridge University Press, Cambridge, 2009.

[14] G. R. Doddington, “Speaker recognition - identifying people by their voices,” *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[15] Cox R. M. and Moore J. N., “Composite speech spectrum for hearing aid gain prescriptions,” *J. Speech Hear Res.*, vol. 31, no. 1, pp. 102–107, 1988.

[16] B. Gardner and K. Martin, “HRTF measurements of a KEMAR dummy-head microphone,” Tech. Rep., MIT Media Lab Perceptual Computing, 1994.

[17] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech Language*, vol. 27, no. 3, pp. 621–633, 2013.

[18] M. Usman, F. Keyrouz, and K. Diepold, “Real time humanoid sound source localization and tracking in a highly reverberant environment,” in *Proc. 9th International Conference on Signal Processing (ICSP)*, Beijing, China, 2008, pp. 2661–2664.

[19] F. Keyrouz, Y. Naous, and K. Diepold, “A new method for binaural 3D localization based on HRTFs,” in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Darmstadt, Germany, 2006.