BINAURAL LOCALIZATION OF SPEECH SOURCES IN 3-D USING A COMPOSITE FEATURE VECTOR OF THE HRTF

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

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

‡CVSSP, University of Surrey, Guildford, United Kingdom. Email: d.talagala@surrey.ac.uk

ABSTRACT

Binaural localization of speech sources in 3-D, using head-related transfer functions (HRTFs), always suffers elevation ambiguity due to the limited high frequency spectral information available at the receivers. This paper presents a method that overcomes this limitation by exploiting the interaural phase and magnitude features present in the HRTF. We (i) introduce a new feature vector that combines these two sets of features in a non-linear fashion, and (ii) propose a mechanism to extract this feature vector free from distortion by the speech spectra. The performance of the proposed method is evaluated and compared with a correlation-based HRTF database matching approach and a two-step localization technique for multiple source positions, HRTFs (individuals) and speech inputs. The results suggest that up to 20% improvement in localization performance can be achieved for moderate signal-to-noise ratios.

Index Terms— Binaural localization, cepstral transformation, generalized cross-correlation (GCC), head related transfer function (HRTF), phase transform (PHAT).

1. INTRODUCTION

Through the course of time, the human auditory system has evolved to efficiently localize sound sources in 3-D using just two signals received at the ears. However, understanding the underlying mechanisms and replicating this ability has been an ongoing process, mainly due to the complex fashion in which the location information is imprinted on the received binaural signals. Numerous techniques have been proposed to mimic the human localization process [1–3], motivated by developments ranging from humanoid robotic systems and target tracking systems to artificial hearing aids.

Of the localization cues contained in the HRTF, the interaural time and level difference (ITD and ILD) are believed to be the most important pieces of information used to determine the azimuth location of a sound source [4, 5]. However, these cues are known to dominate at lower frequencies and are insufficient for localizing the elevation in 3-D, due to the existence of a 'cone-of-confusion' exhibiting similar interaural cues [6]. On the other hand, the high frequency spectral cues are known to facilitate elevation estimation [7], and therefore plays a crucial role in 3-D binaural localization. Yet, in a practical scenario (e.g., localizing speech sources), fully utilizing the richness of the spectral cues may not be possible due to the limited bandwidth of speech sources.

The computation of the correlation between the two received signals and a HRTF dataset is one of the most straightforward localization mechanisms [8]. This however does not consider the localization cue distribution within the HRTF, and is typically inaccurate in noisy environments. Methods that do consider the existence and dispersion of the localization cues generally adopt a joint [9] or two-step process [2, 10] to first estimate the azimuth direction, and the elevation afterwards. The problems with this method are, 1) multiple potential source locations that exhibit similar ITD and ILD characteristics, and 2) the lack of a complete set of spectral cues for elevation estimation. Thus, identifying the most relevant localization information and simultaneous estimation of both azimuth and elevation becomes essential for robust binaural source localization.

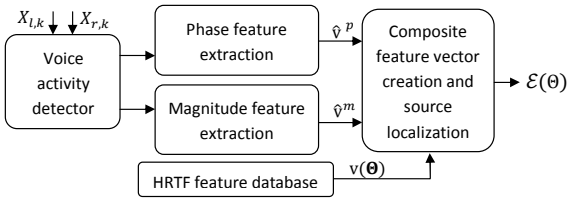


Fig. 1. Operational block diagram of the localization system.

In this paper, we present a method to localize a speech source in 3-D space, by creating a new feature vector, for simultaneous estimation of both azimuth and elevation. First, the localization features are characterized from the low frequency interaural phase difference (IPD), and the mid-high frequency ILD/spectral cues. A new feature vector is constructed to include all these key localization features. Next, we develop the signal processing required to extract these features from the binaural received signals. For example, the IPD is obtained from the cross spectral density of the two-ear signals and a truncated cepstral transform is used to extract the ILD and spectral cues. Finally, the optimal frequency range of the phase and magnitude features are investigated, and the overall localization performance of the proposed feature vector based method is compared with the simple correlation and the two-step localization methods.

2. SYSTEM MODEL

Consider a binaural localization system, where receivers at the two ears sense the convolution of a speech source $s(t)$ and the corresponding head-related impulse response $h_i(t)$ ($i \in \{l, r\}$ representing the left and right ears respectively). For a single (simultaneously active) source localization scenario, the frequency domain representation of the received signal can be expressed as

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

where $X_{i,k}(f)$, $H_i(f, \Theta)$ and $S_k(f)$ represent the received signal, HRTF and source spectra at a frequency f. The source location in

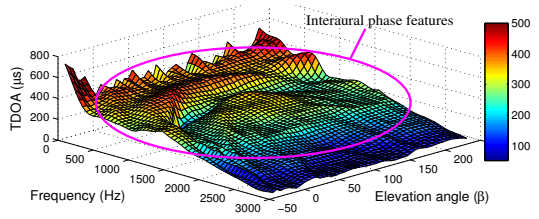


Fig. 2. Time difference of arrival of the HRTFs with respect to the elevation β in the sagittal plane $\alpha = 30^\circ$ of CIPIC 'subject_003'.

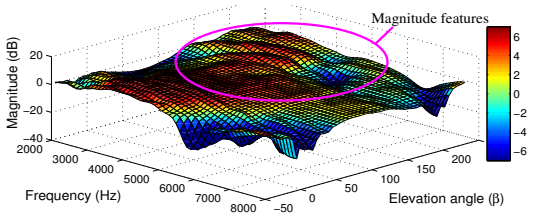


Fig. 3. Left ear HRTF magnitude response indicating monaural magnitude features in the sagittal plane $\alpha = 30^\circ$ of CIPIC 'subject_003'.

3-D is denoted by $\Theta = (\alpha, \beta)$, where α, β correspond to the azimuth and elevation, respectively, in a sagittal coordinate system. $k = 1 \dots K$ represents the time frame number of the speech signal separated into K frames, where the frame length is less than the stationary time duration of the signal (typically 10–50 ms for voiced speech [11]), and $N_{i,k}(\omega)$ represents the additive noise component.

Fig. 1 illustrates the functionality of the different modules in the proposed localization system. In order to localize the speech source in 3-D, the localization features that exist within the HRTF must first be described and stored in a feature database. The main operations of defining the interaural phase and magnitude features (i.e., level difference and spectral information), and thereafter creating a composite feature vector for localization are described below.

2.1. Interaural Phase Features

The ITD that arises as a natural consequence of the spatial separation of the ears, is commonly used for estimating the azimuth of an incoming sound source. Localization techniques such as the Generalized Cross-Correlation Phase Transform (GCC-PHAT) [12] method exploit this fact to estimate the broadband time difference of arrival (TDOA) between the signals and estimate the source azimuth. This however loses much of the subtle differences in phase introduced (which are both frequency and elevation dependent as shown in Fig. 2) by the head, torso and pinna through the effects of scattering and reflections. Thus, incorporating the change in IPD with frequency, for both azimuth and elevation localization, could lead to greater localization accuracy. We propose that this localization information be extracted as an interaural phase feature, and express it as a normalized cross power spectral density given by

$$\mathcal{V}^p(f, \Theta) = \frac{H_r(f, \Theta) H_l(f, \Theta)^*}{|H_r(f, \Theta)||H_l(f, \Theta)^*|}, \tag{2}$$

where * denotes the conjugation operation. Thus, the feature vector of interaural phase information for a source located in the direction Θ can be expressed as

$$\mathbf{v}^p(\Theta) = \left[\; \mathcal{V}^p(f^p_{\min}, \Theta), \quad \cdots, \quad \mathcal{V}^p(f^p_{\max}, \Theta) \; \right], \tag{3}$$

in the range of frequencies $f \in [f^p_{\min}, f^p_{\max}]$.

2.2. Interaural Magnitude Features

The magnitude features of the HRTFs are primarily comprised of the ILD and monaural spectral cues (shown in Fig. 3), and represent a location dependent modulation of the received signal amplitude. Together, the ILD as well as the left and right ear spectral cues (which are commonly used in elevation localization), are therefore interaural magnitude features that can be exploited for 3-D source localization.

In order to extract these features, we adopt a modified cepstral processing method [13], previously proposed to extract the HRTF magnitude response for binaural localization in the median plane. Here, the HRTFs are first transformed into the the cepstral domain, truncated to a finite order to remove any rapid fluctuations in the frequency domain, and finally transformed back into the frequency domain as a smoothed HRTF magnitude response. We therefore express the magnitude feature vector of the signal received at a particular ear as [13]

$$\mathbf{v}^m_i(\Theta) = \mathcal{C}^{-1}\Big\{ \mathcal{T}\big[\, \mathcal{C}\left\{ \mathbf{h}_i(\Theta) \right\} \big] \Big\} \Big| \quad f \in [f^m_{\min}, f^m_{\max}], \tag{4}$$

where $\mathcal{C}$ and $\mathcal{C}^{-1}$ represent the cepstral and inverse cepstral transforms, respectively. $\mathcal{T}$ describes the cepstral truncation operation in [13] and $\mathbf{h}_i(\Theta) = \left[\; H_i(0, \Theta), \quad \cdots, \quad H_i(F_s/2, \Theta) \; \right]$ for a sampling rate of F_s. $f^m_{\min}$ and $f^m_{\max}$ demarcates the range of frequencies whose magnitude features are of interest to us.

2.3. The Composite Feature Vector for 3-D Localization

In order to simultaneously determine α and β in Θ for 3-D localization, we define a new composite feature vector as a non-linear combination of the inteaural phase and interaural magnitude features described in the previous subsections. Mathematically, this feature vector can be expressed as

$$\mathbf{v}(\Theta) \triangleq \mathbf{v}^p(\Theta) \odot \left\{ \mathbf{v}^m_r(\Theta) \oslash \mathbf{v}^m_l(\Theta) \right\}, \tag{5}$$

where $\odot$ and $\oslash$ represent the element-wise multiplication and division of vectors, respectively. The advantage of this non-linear combination is to enlarge the differences of the feature vectors between closely spaced source positions, thus making it especially suitable for 3D localization. It should also be noted that $\mathbf{v}^p(\Theta)$ and $\mathbf{v}^m_i(\Theta)$ must be of similar length; thus, one feature vector may require finer sampling in the frequency domain (in this work, a simple interpolation of $\mathbf{v}^p(\Theta)$ is adopted for this purpose).

3. FEATURE EXTRACTION FROM RECEIVED SIGNALS

The received signals at the two ears in (1), although containing directional information, are both time variant due to the effects of speech and are corrupted by noise. Hence, these signals must be processed further to extract the features discussed in the previous section.

3.1. Received Signal Processing

3.1.1. Feature Extraction: Interaural Phase

In order to extract the interaural phase feature vector, we define the estimated interaural phase as the mean normalized cross power spectral densities of the two received signals in (1), given by

$$\hat{\mathcal{V}}^p(f) \triangleq E\left\{ \frac{X_{r,k}(f) X_{l,k}(f)^*}{|X_{r,k}(f)||X_{l,k}(f)^*|} \right\}, \tag{6}$$

where $E\{\cdot\}$ represents the expectation operator over time. Approximating (6) for K voiced speech frames obtained through the voice activity detector in Fig. 1 (i.e., $|S_k(f)| \neq 0$),

$$\hat{\mathcal{V}}^p(f) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_{r,k}(f)X_{l,k}(f)^*}{|X_{r,k}(f)||X_{l,k}(f)^*|} \approx \frac{H_r(f,\Theta)H_l(f,\Theta)^*}{|H_r(f,\Theta)||H_l(f,\Theta)^*|}, \tag{7}$$

for large received signal-to-noise ratios (SNRs), i.e., $|N_{i,k}(f)| \ll |H_i(f,\Theta)S_k(f)|$, at either ear. Hence, the estimated interaural phase feature vector can be expressed as

$$\hat{\mathbf{v}}^p = \left[\ \hat{\mathcal{V}}^p(f^p_{\min}),\quad \cdots,\quad \hat{\mathcal{V}}^p(f^p_{\max})\ \right]. \tag{8}$$

3.1.2. Feature Extraction: Interaural Magnitude

We extract the interaural magnitude features from the received signal using a modified version of the cepstral preprocessing method proposed in [13]. Exploiting the properties of the cepstral domain signal processing, the time averaged and cepstral truncated signal becomes

$$\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{x}_{i,k}\right\}\right] = \mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{h}_i(\Theta)\right\}\right] + \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{s}_k\right\}\right] + \mathbf{n}_i, \tag{9}$$

where $\mathbf{n}_i$ represents the time averaged noise cepstrum, $\mathbf{x}_{i,k} = [X_{i,k}(0),\cdots,X_{i,k}(F_s/2)]$ and $\mathbf{s}_k = [S_k(0),\cdots,S_k(F_s/2)]$. Assuming sufficiently long observations and same statistical properties for $\mathbf{n}_l$ and $\mathbf{n}_r$ (i.e., $\mathbf{n}_r - \mathbf{n}_l \to 0$), the inverse cepstral transform of the difference in (9) between the two ears represents the magnitude feature vector. That is, we extract the feature

$$\begin{aligned}\hat{\mathbf{v}}^m \ &\triangleq\ \mathcal{C}^{-1}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{x}_{r,k}\right\}\right] - \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{x}_{l,k}\right\}\right]\right\}\\ &\approx\ \mathcal{C}^{-1}\left\{\mathcal{T}\left[\,\mathcal{C}\left\{\mathbf{h}_r(\Theta)\right\} - \mathcal{C}\left\{\mathbf{h}_l(\Theta)\right\}\right]\right\},\end{aligned} \tag{10}$$

which is roughly equivalent to $\mathbf{v}^m_r(\Theta) \oslash \mathbf{v}^m_l(\Theta)$ in (5).

Combining (8) and (10), the estimated composite feature vector can therefore be expressed as

$$\hat{\mathbf{v}} \triangleq \hat{\mathbf{v}}^p \odot \hat{\mathbf{v}}^m. \tag{11}$$

3.2. Source Location Estimation

Ideally, the estimated composite feature vector in (11) is coincident only with the composite feature vector of the true source location given by (5). Thus, the magnitude of the Euclidean distance between these quantities can be used to estimate the source location. We express this as a localization error metric $\forall\Theta$, given by

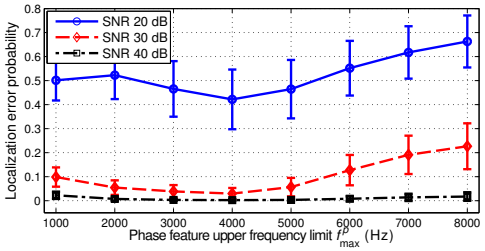
$$\mathcal{E}\big(\Theta \equiv (\alpha,\beta)\big) = 20\log_{10}\left\|\left\{\frac{\hat{\mathbf{v}}}{\|\hat{\mathbf{v}}\|} - \frac{\mathbf{v}(\Theta)}{\|\mathbf{v}(\Theta)\|}\right\}\right\|, \tag{12}$$

the minimum of which yields the estimated source location in 3-D.

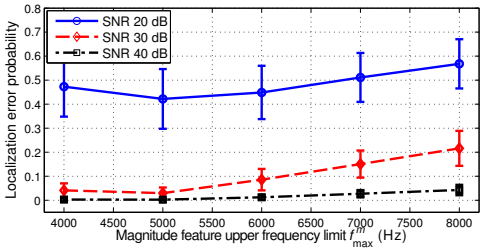
4. EVALUATION

4.1. Simulation Configuration

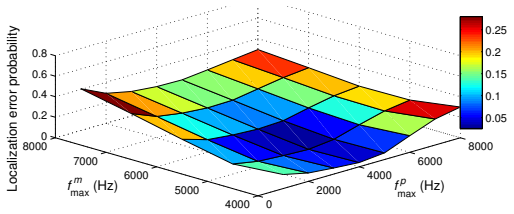
The proposed binaural localization technique in 3-D space is evaluated through simulations using the CIPIC HRTF database [14]. Here, the HRTFs of 45 subjects, each with 950 different locations (the azimuth angle varies from -45° to 45° with a 5° interval and the elevation varies from -45° to 230.625° degree with a 5.625° interval), are utilized. The clean speech signals are obtained from the recordings used in the "PASCAL 'CHiME' Speech Separation and Recognition Challenges" [15]. The database contains 34 speakers, each with 500 utterance segments sampled at 16 kHz. The received binaural signals are simulated by convoluting the HRTF with speech samples, and considering the variability of the speech spectrum, a short-time approach, i.e., a short-time Fourier transform, is applied. A voice activity detector identifies the voiced speech frames, while both the Fourier and cepstral transforms employ a 20 ms Hamming window with 10 ms overlap. For the 16 kHz sample rate, this implies that a window length of 320 is used throughout the simulation. The optimum feature frequency ranges are selected by comparing the localization performances for different range combinations.



(a) Interaural phase feature frequency range selection



(b) Interaural magnitude feature frequency range selection



(c) Joint phase and magnitude feature frequency range selection

Fig. 4. 3-D localization error probability for different frequency ranges of phase and magnitude features with respect to SNR. (a) Localization error probability with respect to the phase feature frequency range 0–$f^p_{\max}$ kHz and a magnitude feature range of 3–5 kHz. (b) Localization error probability with respect to the magnitude feature frequency range 3–$f^m_{\max}$ kHz and the phase feature range of [0, 4] kHz. (c) Localization error probability with respect to upper frequency limits of phase and magnitude features at 30 dB SNR.

The performance of the proposed localization technique is compared with the two-step [2,10] and simple correlation approaches [8]. The two-step method, uses ITD/ILD information to narrow down the potential source locations to a specific cone-of-confusion, prior to estimating the elevation using the binaural magnitude spectra. The simple correlation approach on the other hand, simply computes the correlation between the received binaural signals and the complete HRTF dataset. In both cases, the best match to the HRTF dataset identifies the estimated source location. We compare the localization performance in terms of the 'localization error probability', i.e., the likelihood of a localization error across all 950 source locations over multiple trials (lower values imply better performance). The results indicate the mean localization performance for all 45 subjects in the HRTF dataset, while the error bars indicate the standard deviation.

4.2. Impact of Bandwidth on the Feature Extraction Process

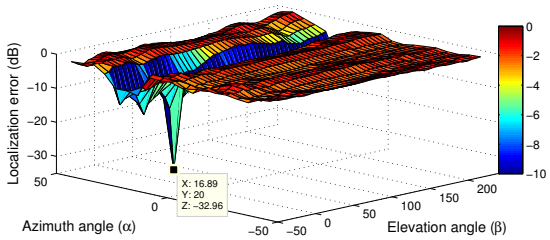
A significant issue when constructing the composite feature vector is the selection of the appropriate frequency bands for feature extraction. This requires that only the most relevant localization information is included, so as to reduce the complexity and minimize the noise influence. Traditionally, the phase features are extracted from a relatively low frequency range (i.e., 0–1.5kHz) to calculate the ITD/IPD [16]. However, in the proposed feature vector, the phase features are not only used to estimate the azimuth but also to localize the elevation; and hence, requires much higher frequencies to be included [7]. In the simulations, we modify the upper frequency limit of the phase feature, $f^p_{\max}$, in order to determine the optimal frequency range. Here, the lower frequency limit $f^p_{\min}$ is fixed at 0, since the low frequency phase information is essential for azimuth estimation. As for the magnitude features, it is necessary to find the frequency band that includes the key features, subject to minimal distortion by the speech spectrum, within the speech bandwidth. Since the formants and a majority of the speech energy exists below 3 kHz [17], the lower frequency limit, $f^m_{\min}$, is therefore fixed at 3 kHz and the upper limit $f^m_{\max}$ is varied during the simulation.

The simulation results of the localization error probability (for all 45 subjects in the HRTF database) for feature range selection are illustrated in Fig. 4 at different SNRs. The results indicate that the optimum frequency bands for phase and magnitude feature extraction are $[0, 4]$ kHz and $[3, 5]$ kHz; a region where the localization error probability is a minimum for all SNRs. Figs. 4(a) and (b) illustrate the impact of the upper frequency limits for the phase and magnitude features, respectively. The joint impact of both parameters is illustrated in Fig. 4(c). They show that the interaural phase features, up to a relatively high frequency range, do actively contribute to elevation estimation, as expected intuitively from Fig. 2. As for the magnitude feature frequency range, when the upper frequency limit is greater than 6 kHz, the received signals are more strongly affected by noise (due to the rapid decay of speech energy with frequency), and results in a degradation of the localization performance.

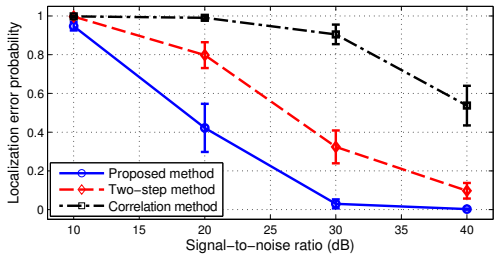
4.3. Overall 3-D Localization Performance

Fig. 5(a) illustrates an example spectra of the localization error metric (described in Section 3.2), for a source located at $\Theta \equiv (20^\circ, 16.875^\circ)$ at 30 dB SNR. As shown, the estimated source location corresponds to the location that minimizes the error metric $\mathcal{E}(\Theta)$. Furthermore, the distinctive spike suggests that elevation ambiguity is also minimal. In order to evaluate the 3-D localization performance, we use these estimates for all possible source locations and subjects, and analyse the overall localization error probability.

Fig. 5(b) illustrates the overall localization error probability and compares the performance of the proposed method and the two comparison methods. The presented data illustrates the mean performance for all 45 CIPIC subjects, where the error bar indicates the standard deviation. In general, as expected, the localization error probability decreases with increasing SNRs. The proposed method has the best localization performance with the lowest probability of localization errors under all noise conditions. This suggests that the proposed method is more robust to the effects of noise. Furthermore, the error bars of the proposed method are generally smaller than the other two methods, especially at higher SNR, which implies that the proposed method has more consistent performance for different listeners' HRTFs. In addition, comparing the proposed method and the two-step method, the crucial difference is in the use of the phase features for elevation estimation. Here, the two-step method utilizes the interural phase feature only for azimuth localization, while the



(a) Localization error metric $\mathcal{E}$ for a source located at $\Theta \equiv (20^\circ, 16.875^\circ)$



(b) Overall 3-D localization performance comparison

Fig. 5. (a) Localization error metric of the proposed method for a source located at α=20° and β=16.875°. (b) Comparison of the overall 3-D localization error probability of the proposed method, two-step method and correlation method for SNRs from 10–40 dB.

proposed method does so for joint azimuth and elevation estimation. The superior localization performance of the proposed method therefore proves that the phase features do indeed contain elevation localization information, and should not be neglected in 3-D binaural localization of speech sources.

5. CONCLUSION

In this paper, we propose a novel composite feature vector based method for binaural localization of speech sources in 3-D space. We describe how interaural phase and magnitude feature vectors can be derived from the HRTFs, and describe the process of extracting and combining these features from the received signals. A method to estimate the azimuth and elevation simultaneously is introduced afterwards. We evaluate its performance through simulations, and show that optimum bandwidths for both phase and magnitude features exist. The overall performance of the proposed method is compared with two other methods, and is shown to produce a more consistent, noise-robust source location estimate. Furthermore, the results show that interaural phase information is critically important for accurate elevation estimation using sources such as speech. Future work will extend this feature vector concept to multi-source localization.

6. RELATION TO PRIOR WORK

This work presents a method for binaural localization of speech sources in 3-D using a feature vector of the localization information in the HRTFs. In the literature, elevation estimation has predominantly relied on spectral cues [1–3, 7, 8, 13], which may be eclipsed by noise when receiving sources such as speech with less energy at higher frequencies. The present study attempts to extract a set of phase and magnitude features [13] in the HRTFs, free from the effects of noise, within the speech bandwidth. It combines these features into a single composite feature for simultaneous azimuth and elevation estimation, which has not been considered in previous studies.

7. REFERENCES

[1] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2098–2107, Sep. 2014.

[2] H. Liu and J. Zhang, "A binaural sound source localization model based on time-delay compensation and interaural coherence," in *Proc. 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1424–1428.

[3] H. Nakashima and T. Mukai, "3D sound source localization system based on learning of binaural hearing," in *Proc. IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Waikoloa, Hawaii, USA, Oct. 2005, vol. 4, pp. 3534–3539.

[4] 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, Jan. 2010.

[5] 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, Jul. 2012.

[6] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *Journal of the Audio Engineering Society*, vol. 49, no. 4, pp. 231–249, Apr. 2001.

[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, Aug. 2007.

[8] F. Keyrouz and K. Diepold, "An enhanced binaural 3D sound localization algorithm," in *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Vancouver, BC, Canada, Aug. 2006, pp. 662–665.

[9] K. D. Martin, "Estimating azimuth and elevation from interaural differences," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 1995, pp. 96–99.

[10] 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, Oct. 2006, pp. 1170–1176.

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

[12] C. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[13] 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 *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, Sep. 2014, pp. 1–5.

[14] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Paltz, NY, USA, Oct. 2001, pp. 99–102.

[15] 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, May 2013.

[16] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Trans. Syst., Man, Cybern., Part B*, vol. 36, no. 5, pp. 982–994, Oct. 2006.

[17] D. S. Deepawale and R. Bachu, "Energy estimation between adjacent formant frequencies to identify speaker's gender," in *Proc. 5th International Conference on Information Technology: New Generations (ITNG)*, Las Vegas, NV, USA, Apr. 2008, pp. 772–776.