SPATIAL FEATURE LEARNING FOR ROBUST BINAURAL SOUND SOURCE LOCALIZATION USING A COMPOSITE FEATURE VECTOR

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

The performance of binaural speech source localization systems can be significantly impacted by an imperfect selection of spatial localization cues, due to the limited bandwidth of speech, and the effects of noise. In order to mitigate these impacts, this paper presents a novel method that combines a deterministic localization approach with a spatial feature learning process. Here, we (i) obtain a composite feature vector derived from analysing the mutual information between different spatial cues and (ii) estimate the optimum feature combination that minimizes the angular localization error in three dimensional space. The performance of the proposed mutual information based feature learning approach is evaluated for a range of speech inputs and noise conditions. We also demonstrate that the proposed approach improves the localization accuracy and its robustness, with respect to traditional localization algorithms, especially in the relatively low signal-to-noise ratio localization scenarios.

Index Terms— Binaural localization, generalized cross correlation, head related transfer function (HRTF), mutual information.

1. INTRODUCTION

In the process of localizing a sound source in three dimensional (3-D) space, the human auditory system must contend with and adapt to a wide variety of input signals and noise conditions. Although it is understood that the head-related transfer function (HRTF) plays a major role in determining the localization accuracy [1–4], precisely how the brain utilizes the knowledge of the HRTF and adapts to these different environments remains unclear [5–7]. Mimicking this ability, for example by learning the most robust spatial features contained within the HRTF, could therefore greatly benefit the development of a range of artificial hearing systems from robotic platforms to hearing aids.

It is widely believed that people exploit three sets of spatial cues embedded in the HRTF for localization. Namely, the interaural time difference (ITD), the interaural level difference (ILD) and spectral cues [8–10]. Although each spatial cue dominates at a particular range of frequencies [11] and predominantly contributes to either azimuth or elevation resolution, it has been shown that combinations of these cues can be used to localize a source in 3-D space [12, 13]. However, improving the localization accuracy (the elevation accuracy in particular) requires the use of broadband sources at high signal-to-noise-ratios (SNRs), which represent conditions that may not always be satisfied in practical applications such as speaker localization for the hearing impaired. Furthermore, in the case of speech sources, the spectral cues that help determining elevation are constrained further due to the limited bandwidth of speech, making the characterization of the spatial cue variations even more challenging.

Although the simplest localization techniques are effectively based on the analysis of the ITD between the signals received at the two ears [12], more complex techniques that evaluate combinations of spatial cues (i.e., spatial features) in the received signals have been proposed in the literature [13–17]. In each case, the selection of the appropriate feature set is a crucial factor that determines the azimuth localization performance. Considering the changing environment, research has been done to dynamically assess the importance of spatial features, where the interaural coherence (IC) was analysed for selecting reliable binaural cues in horizontal plane [18, 19].

This paper presents a learning-based approach to select spatial features for estimating speech sources in 3-D space, based on the concept of assessing the mutual information (MI) [20–22] that exists between each spatial cue and the corresponding source location. We focus on estimating two parameters, azimuth and elevation. The paper is organized as follows. First, the spatial cues described in [13] are adopted as a basic inputs to a feature learning process and are used to evaluate the MI content between the spatial cues and source locations in the space-frequency domain for a variety of speech and noise conditions. Next, a feature ranking process is utilized to rearrange the spatial cues in terms of their learned relative importance to the overall localization performance, and a subset of these are extracted to create a composite feature vector for localization. Finally, the performance of the resulting learning-based algorithm is evaluated for a range of speech and noise conditions, and is compared with the generic feature-based algorithm [13] and the traditional correlation-based approach.

2. SYSTEM MODEL

The received signals of a single source binaural localization system consist of several components; a speech source $s(t)$ convolved with a direction-dependent head-related impulse response (HRIR) $h_i(t, \Theta)$ ($i \in \{l, r\}$ indicates the left and right ears), where $\Theta \equiv (\alpha, \beta)$ describes the source location in a 3-D sagittal coordinate system, and an additive uncorrelated diffuse background noise component. Expressed in the frequency domain, this relationship is given by

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

where $X_{i,k}(f)$, $S_k(f)$ and $N_{i,k}(f)$ represent the received signal, speech source and noise spectra respectively at a frequency f.

This work was supported under the Australian Research Councils Discovery Projects funding scheme (project no. DE150100363).

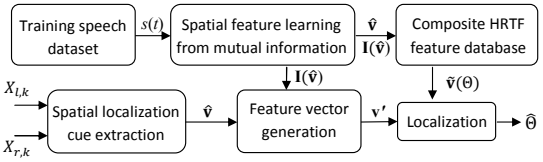


Fig. 1. Functional diagram of the proposed localization approach.

$H_i(f, \Theta)$ denotes the HRTF, i.e., the frequency domain expression of the HRIR $h_i(t, \Theta)$, and $k = 1, \ldots, K$ indexes the K speech frames [13, 23] that exist in $s(t)$.

Fig. 1 illustrates the functional components of the proposed binaural localization system that estimates Θ utilizing the two signals in (1) and a database of pre-measured HRTFs applicable to a particular system. Here, in order to minimize the impact of varying speech and noise conditions on the estimation accuracy, the concept of MI is applied to learn the appropriate spatial cues for different noise profiles. Afterwards, the knowledge gained is used to generate a composite spatial feature vector database for localization. The spatial cues used in the process, as well as the concept of MI, are briefly described in the following subsections.

2.1. Spatial cues: Interaural phase and magnitude

Following from the work in [13], two spatial localization cues known as interaural phase and magnitude cues are used as a basis for the localization process. First, the spectrum of each HRTF is decomposed into L subbands, where each subband corresponds to $f_\mu \in \{f_1, \ldots, f_L\}$. Next, the interaural phase cue v_μ^P and interaural magnitude cue v_μ^M for each subband are extracted. Thus, from [13],

$$v_\mu^P(\Theta) = \frac{H_r(f_\mu, \Theta) H_l(f_\mu, \Theta)^*}{|H_r(f_\mu, \Theta)||H_l(f_\mu, \Theta)^*|} \tag{2}$$

and

$$v_\mu^M(\Theta) = \mathcal{C}^{-1}\Big\{\mathcal{T}\big[\,\mathcal{C}\left\{H_l(f_\mu, \Theta)\right\} - \mathcal{C}\left\{H_r(f_\mu, \Theta)\right\}\big]\Big\}, \tag{3}$$

where the $(\cdot)^*$ denotes the conjugation operation, and $\mathcal{C}$, $\mathcal{C}^{-1}$ and $\mathcal{T}$ represent the spectrum to cepstrum transformation, its inverse transformation and the cepstral truncation operation [17], respectively.

2.2. MI between location and spatial cues

Although the feature vectors described above can be used to localize speech sources [13], the impact of noise cannot be neglected. For example, the additive noise distorts the characteristics of both phase and magnitude features, especially at the higher frequencies (above 3 kHz) where the speech energy is comparatively lower. This is exacerbated as most elevation localization cues being generated by reflections and diffraction of human body and pinna are above 3 kHz [24]. An intelligent frequency bin selection mechanism that maximizes the direction-dependent information and minimizes the impact of noise, could therefore lead to superior performance when localizing speech sources as per [13]. Thus, the MI that exists between each spatial localization cue and the source location for a particular frequency bin could be used as a criteria to evaluate both the effectiveness and the robustness of the feature vectors extracted for the localization process.

2.2.1. MI computation

As one of the most common measures to evaluate the dependency between variables, MI is widely used to estimate the maximally relevant feature selection that corresponds to a particular outcome [20, 22]. Thus, in order to evaluate the relationship between the features and the location (including the impact of speech), the MI contained in a spatial cue can be expressed as [20]

$$I(\hat{v}_\mu^j; \Theta) = \int\int p(\hat{v}_\mu^j, \Theta) \log \frac{p(\hat{v}_\mu^j, \Theta)}{p(\hat{v}_\mu^j) p(\Theta)} \mathrm{d}\hat{v}_\mu^j \mathrm{d}\Theta, \tag{4}$$

where $j \in \{P, M\}$ and $\hat{v}_\mu^j$ denotes the μ^{th} spatial cue extracted from the received binaural signals (the details of extracting $\hat{v}_\mu^j$ are described in Sec. 3.1). Computed using the Parzen windows density estimation method [21], $p(\Theta)$, $p(\hat{v}_\mu^j)$ and $p(\hat{v}_\mu^j, \Theta)$ represent the individual and joint probability densities of Θ and $\hat{v}_\mu^j$.

2.2.2. Analysis of MI in spatial cues

Figure 2 illustrates the variations in MI that exist between the elevation angle of the source location and the spatial cues for different azimuths and SNRs. From Figs. 2(a) and (b), it can be observed that in high SNR scenarios the MI in the high frequency range becomes dominant for elevation localization. However, with the decreasing SNRs, the high-frequency cues no longer provide reliable localization information unlike the mid-frequency spatial cues. Furthermore, the distribution (in frequency) of the most effective cues varies with different azimuths, and illustrates both the difficulty of decoupling the localization process into separate azimuth and elevation estimation problems, as well as the challenge of localization in the median plane [24].

Furthermore, comparing the behaviour of the two types of spatial cues, we can observe that the importance of each changes with the SNR. For example, where no or low noise is present, the interaural magnitude cues become dominant, while in a comparably higher noise environment, the interaural phase cues show more robustness. Collectively, these observations imply that the selection of spatial cues for the creation of a feature vector for localization must be more nuanced than the simple selection of a fixed frequency range; thus, an adaptive noise-dependant feature selection and extraction process becomes a necessity for any noise-robust binaural localization system. A spatial feature learning algorithm that is aware of the MI contained in each spatial cue can satisfy this requirement and provides superior performance to the former approach, as illustrated in the following sections.

3. FEATURE EXTRACTION AND LOCALIZATION

This section describes this process, the learning algorithm that creates the spatial feature vectors based on the MI in analysis cues, and the final localization method based on the learned spatial feature vectors.

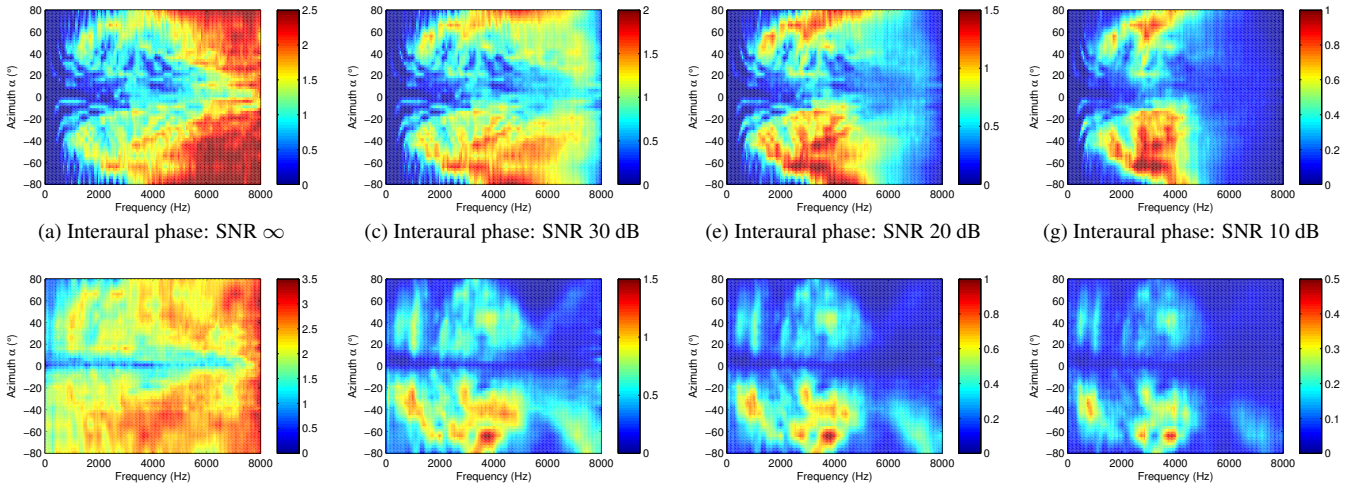
3.1. Spatial cue extraction from binaural signals

The interaural phase and magnitude cues from K voiced speech frames in (1) can be extracted and can be expressed as [13]

$$\hat{v}_\mu^P \triangleq \frac{1}{K}\sum_{k=1}^{K} \frac{X_{r,k}(f_\mu) X_{l,k}(f_\mu)^*}{|X_{r,k}(f_\mu)||X_{l,k}(f_\mu)^*|} \approx \frac{H_r(f_\mu, \Theta) H_l(f_\mu, \Theta)^*}{|H_r(f_\mu, \Theta)||H_l(f_\mu, \Theta)^*|} \tag{5}$$

and

$$\hat{v}_\mu^M \triangleq \mathcal{C}^{-1}\Big\{\frac{1}{K}\sum_{k=1}^{K} \mathcal{T}\big[\,\mathcal{C}\left\{X_{l,k}(f_\mu)\right\} - \mathcal{C}\left\{X_{r,k}(f_\mu)\right\}\big]\Big\}. \tag{6}$$



(a) Interaural phase: SNR ∞ (c) Interaural phase: SNR 30 dB (e) Interaural phase: SNR 20 dB (g) Interaural phase: SNR 10 dB

(b) Interaural magnitude: SNR ∞ (d) Interaural magnitude: SNR 30 dB (f) Interaural magnitude: SNR 20 dB (h) Interaural magnitude: SNR 10 dB

Fig. 2. MI between the spatial cues and the elevation for a range of azimuths, frequencies and noise conditions.

Combining the spatial cues in (5) and (6), a composite feature vector given by $\hat{\mathbf{v}} \triangleq [\hat{v}_1^P, \ldots, \hat{v}_L^P, \hat{v}_1^M, \ldots, \hat{v}_L^M]$, where $\hat{\mathbf{v}} \in \mathbb{R}^{2L}$, can be created. The MI relevant to each element in $\hat{\mathbf{v}}$ and an arbitrary Θ can also be computed (described in Algorithm 1) from (4), resulting in a vector of MI given by

$$\mathbf{I}(\hat{\mathbf{v}};\Theta) \triangleq [I(\hat{v}_1^P;\Theta), \ldots, I(\hat{v}_L^P;\Theta), I(\hat{v}_1^M;\Theta), \ldots, I(\hat{v}_L^M;\Theta)], \tag{7}$$

where $\mathbf{I}(\hat{\mathbf{v}};\Theta) \in \mathbb{R}^{2L}$. The following subsection describes the MI aided spatial feature learning process that can be applied to $\hat{\mathbf{v}}$ aided by the knowledge of $\mathbf{I}(\hat{\mathbf{v}};\Theta)$, and the localization process based on the resulting selected spatial feature vector.

3.2. Spatial feature learning and localization

Algorithm 1 describes the MI-based spatial feature learning mechanism used in the remainder of this work. Given an estimated noise level, such as using the method proposed in [25], the spatial feature vector $\hat{\mathbf{v}}$ and its corresponding MI in (7) are computed for a set of training speech signals. Figs.3 illustrate the variation of the MI between spatial cues and elevation angle with and without noise, and indicates the changing nature of the importance of individual spatial cues. The error bar reflects the MI variation on different azimuth planes and the mean MI obtained for the training speech dataset is used thereafter to create a rearranged spatial feature vector $\tilde{\mathbf{v}}$.

To arrive at $\tilde{\mathbf{v}}$, an optimal number of spatial cues to be used in the localization process $\tilde{l}$ is computed. The average angular localization error, obtained from the estimated source location Θ and its estimate $\hat{\Theta}$, is used as a metric to determine the optimal $\tilde{l}$. This results in a set of spatial feature vectors $\tilde{\mathbf{v}}(\Theta)$ applicable to the specified noise level (in the case of some practical applications, it is also possible to pre-train the system for a set of known, approximate noise conditions). The spatial cues extracted from the received signals are rearranged similarly, and the resultant feature vector $\mathbf{v}'$ and the $\tilde{\mathbf{v}}(\Theta)$ reference features are used to localize by applying (12) in [13].

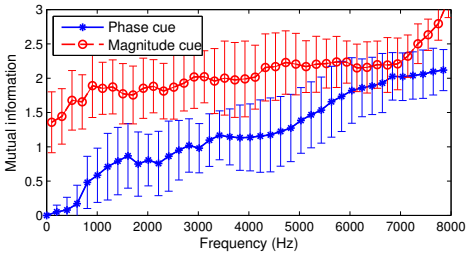
Step 1: Estimate the noise power $\mapsto \sigma^2$.
Step 2: Evaluate $I(\hat{v}_\mu^j;\Theta)$; MI of spatial cues.
  foreach Θ *in the HRTF dataset* **do**
    foreach $s(t)$ *in the training speech dataset* **do**
      Compute $X_{i,k}(f_\mu)$ for a simulated noise power σ^2.
      Calculate the corresponding $\hat{v}_\mu^j$ and $\hat{\mathbf{v}}$.
      Estimate $\mathbf{I}(\hat{\mathbf{v}};\Theta)$.
    end
  end
  Result: Obtain the set of $\mathbf{I}(\hat{\mathbf{v}};\Theta)$ for a noise power σ^2.
Step 3: Learn the optimal combination of the spatial cues in $\hat{\mathbf{v}}$.
  Calculate a mean MI $\forall\Theta$ from $\mathbf{I}(\hat{\mathbf{v}};\Theta)$.
  for $\tilde{l} \leftarrow 1$ **to** $2L$ **do**
    Rearrange $\hat{\mathbf{v}}$ in descending order of MI.
    foreach $\hat{\mathbf{v}}$ *derived from the training speech dataset* **do**
      Estimate the source location $\hat{\Theta}$ from a truncated composite spatial feature vector $\hat{\mathbf{v}}$ of length $\tilde{l}$.
      Calculate the angular localization error of $\hat{\Theta}$.
    end
  end
  Result: Obtain $\tilde{\mathbf{v}}$; the rearranged and truncated spatial feature vector from $\hat{\mathbf{v}}$, that corresponds to the minimum mean angular localization error.

Algorithm 1: Spatial feature learning for robust localization.

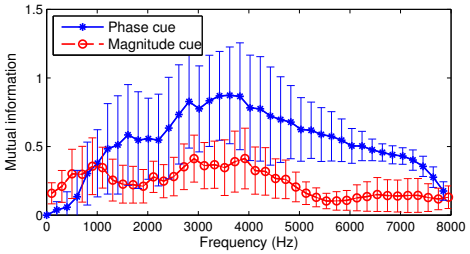
4. EVALUATION

4.1. Simulation configuration

The proposed approach is used to evaluate 950 source locations, ranging from azimuth $\alpha = -45^\circ$ to 45° in 5° intervals and elevation $\beta = -45^\circ$ to 230.625° in 5.625° increments, for the first 10 subjects' HRTF measurements in the CIPIC database [26]. The speech samples from the "PASCAL 'CHiME' Speech Separation and Recognition Challenge" [27] (34 males and females each with 500 utterances sampled at 16 kHz) are used as inputs; 340 randomly picked utterances are used for the learning process, and a separate 200 utterances are used to evaluate the localization performance. The binaural signals are simulated by convolving the HRTFs of different locations with the uncorrupted speech and introducing the independent additive white Gaussian noise with three different SNRs of 10 dB, 20 dB and 30 dB, then a short-time Fourier transform (using a 20 ms Hamming window) is applied afterwards to obtain (1).



(a) SNR: ∞



(b) SNR: 20 dB

Fig. 3. MI variation in spatial cues with respect to SNR.

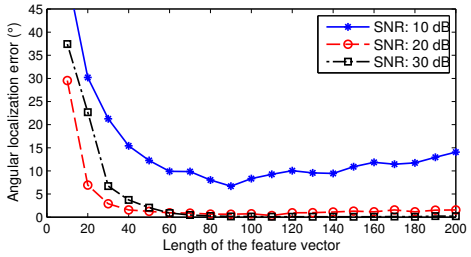


Fig. 4. Localization error with respect to feature vector length.

The localization performance of the proposed method is compared with the generic composite feature-based localization approach [13] and a simple correlation-based method [12]. The frequency range for generic composite feature-based approach is selected empirically, where [0,4] kHz and [3,5] kHz are the phase and magnitude feature regions for the feature-based method, while the full-band signal is used for the correlation approach. During the comparison, the mean angular error is employed as a metric to assess the localization performance. The angular error denotes the angular distance between the estimated and actual source directions in a 3-D coordinate system, therefore the estimation errors of both the azimuth and elevation (α and β) are implicitly included in the performance assessment.

4.2. Performance impact of the feature vector length

From Figs. 2 and 3, it becomes apparent that the length of the feature vector $\tilde{\mathbf{v}}$ can directly influence the localization performance. For example, a length smaller than some optimum will result in insufficient spatial information (especially in the case of the median plane), while a greater length could result in an increased ambiguity due to the effects of noise. In both cases, the mean angular localization error will be impacted; thus, an optimum length for $\tilde{\mathbf{v}}$ that minimizes this error must be computed at the noise power level observed in a particular localization scenario. Hence, the training process described in Algorithm 1 is applied to a range of simulated speech inputs, and the optimum feature vector length and spatial cue combination is obtained dynamically based on their MI content.

Figure 4 illustrates the relationship between the mean angular errors and the length of the composite feature vector at different noise levels. The results are presented for three different SNRs, where the selected number of spatial cues is varied from 10 to 200 in intervals of 10. The result for the 10 dB SNR case clearly illustrates the general behaviour discussed above (similar behaviour is observed at other noise levels as well), indicating an optimum feature vector length of approximately 90 elements. Notice that the angular localization error for 30 dB scenario is larger than that for 20 dB scenario when the feature vector length is less than 60. This suggests that a short feature vector may lead to unstable localization performance thus a minimum length of the feature vector should be guaranteed.

Localization approach	Mean angular localization error		
	10 dB	20 dB	30 dB
Proposed learning	5.63°	0.89°	0.14°
Composite feature [13]	24.30°	5.11°	0.85°
Correlation [12]	67.65°	58.55°	51.58°

Table 1. 3-D space localization performance comparison.

4.3. Localization performance

The performance of the proposed method is presented and compared with two other approaches in Table 1. Here, the received binaural inputs are obtained from 90 uniformly sampled source locations of the 950 locations in the HRTF dataset, and the resulting localization error is averaged across multiple untrained speech inputs and source locations. The results indicate a significant improvement in the performance over the generic composite feature-based localization approach in [13], especially in the low SNR configurations. It was notable that the improvement predominantly stemmed from a reduction in front-to-back confusions. This suggests that the approach overcomes the lack of spectral cues located beyond the mid-high frequency ranges [28] that are less robust to the effects of noise.In general, the results suggest that the MI-based feature learning and rearrangement of the spatial cues in the feature vector can both improve the localization performance and overcome the negative impact of the dynamic truncation of the feature vector to achieve a greater robustness to noise.

5. CONCLUSION

In this paper, we propose a novel method to assess the importance and robustness of spatial cues contained in the HRTFs for binaural speech source localization. First, we apply the concept of MI to determine the significance of spatial cues in noisy environments, and obtain a feature vector for localizing a source in 3-D space, using a rearrangement of the spatial cues in terms of their MI content. A learning process is introduced afterwards that dynamically generates an optimum composite feature vector based on the analysis of MI results. Finally, we evaluate the performance of the proposed method and compare it with a generic feature-based method and a traditional correlation-based approach. A significant performance improvement in the low SNR scenarios is observed, which suggests that the approach may be extendible to more complex localization environments.

6. REFERENCES

[1] J. Blauert, "Sound localization in the median plane," *Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

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

[3] Constantine Trahiotis, Leslie R. Bernstein, Richard M. Stern, and Thomas N. Buell, "Interaural correlation as the basis of a working model of binaural processing: An introduction," in *Sound source localization*, A. N. Popper and R. R. Fay, Eds., chapter 7, pp. 238–271. Springer, New York, 2005.

[4] H. Steven Colburn and Abhijit Kulkarni, "Models of sound localization," in *Sound source localization*, A. N. Popper and R. R. Fay, Eds., chapter 8, pp. 272–316. Springer, New York, 2005.

[5] P. M. Hofman, J. G. A. van Riswick, and A. J. van Opstal, "Relearning sound localization with new ears," *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421, Sep. 1998.

[6] P. Zahorik, P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam, "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 343–359, Jul. 2006.

[7] V. Best, S. Carlile, C. Jin, and A. van Schaik, "The role of high frequencies in speech localization," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 353–363, Jul. 2005.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Binaural localization of speech sources in 3-D using a composite feature vector of the HRTF," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 2654 – 2658.

[14] 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.

[15] 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.

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

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

[18] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, Jul. 2004.

[19] M. Dietz, S. D. Ewert, and V.Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, pp. 592–605, 2011.

[20] P. Hanchuan, L. Fuhui, and D. Chris, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 27, no. 8, pp. 1226 – 1238, Aug. 2005.

[21] N. Kwak and C. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 24, pp. 1667 – 1671, Dec. 2002.

[22] L. T. Vinh, S. Lee, Y. T. Part, and B.J. Auriol, "A novel feature selection mehtod based on normalized mutual information," *Springer Appl. Intell.*, vol. 37, pp. 100 – 120, Jul. 2012.

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

[24] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am*, vol. 109, no. 3, pp. 1110–1122, Mar. 2001.

[25] A.H. Kamkar-Parsi and M. Bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 521–533, May 2009.

[26] 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.

[27] 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.

[28] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *J. Acoust. Soc. Am*, vol. 118, no. 1, pp. 364–374, Jul. 2005.