BINAURAL SPEECH SOURCE LOCALIZATION USING TEMPLATE MATCHING OF INTERAURAL TIME DIFFERENCE PATTERNS

Girija Ramesan Karthik, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bengaluru-560012, India

ABSTRACT

In this paper we present a template based algorithm for localizing speech sources from a binaural recording. Binaural recordings are associated with head related transfer functions (HRTFs) for each direction which are specific to the object, say head, in between the two microphones. So, using these HRTFs and time-frequency representations of the binaural signals, we learn direction specific two dimensional reference templates using histograms of interaural time difference (ITD) in each frequency subband. These are called ITD pattern templates (IPTs). Test templates are then compared with each of the reference IPTs. The reference IPT, that matches best with the test template, provides the estimated direction of arrival for the test speech source. Experimental results obtained using subject_003 from the CIPIC database show that IPT based localization performs better than existing methods where the ITD distribution is modeled using Gaussian mixture model. Given n time-frequency points, we also present a method with complexity $O(n)$ to compute the IPT, thus making it computationally efficient.

Index Terms— interaural time difference, gammatone filters, binaural localization.

1. INTRODUCTION

Speech source localization is essential for a wide range of applications, including human-robot interaction, surveillance and hearing aids. Several localization algorithms have been proposed using microphone arrays with varied number of microphones [1–6]. However, humans have an incredible ability to localize sounds with just two ears. We use two key differences/cues to localize sounds. They are called interaural time difference (ITD) and interaural level difference (ILD). These are the time difference and the intensity difference between the signals reaching the two ears whose values depend on the position of the source. Algorithms inspired by the binaural localization ability of humans would extract these cues from the two input signals [7–20].

If we have two omni-directional microphones without an object, like the human head in between, then the ITD and ILD will be frequency independent. However, most applications like robots and hearing aids will have an object between the microphones. With the inclusion of an object , the interaural cues become frequency dependent due to the diffractions and reflections caused by this object. These frequency dependent characteristics of the interaural cues can be captured by the head-related transfer function (HRTF) [21]. This frequency dependence motivates the use of time-frequency representations. One of the most common time-frequency representations is the Short-Time Fourier Transform (STFT) [9, 11, 13, 15] which assumes uniform subband width and spacing in the frequency domain. Another approach is to use gammatone filters [22] where the subband width and spacing are not uniform [7, 10, 12, 14]. The use of gammatone filters is inspired by the filter structure of the cochlea in human ears. In this work, we use gammatone filters to preprocess the binaural signals.

The gammatone filtered signals are then processed to obtain the interaural parameters for each frame in each gammatone subband. We consider only ITDs in this paper. May et al. [12] and Woodruff et al. [14] use Gaussian mixture models (GMMs) to model the interaural parameters for each subband in every direction. Then, for a test-speech, May et al. calculate the log-likelihoods on a frame by frame basis. In each frame the log-likelihood is obtained by adding the log-likelihoods of all the subbands. The direction with the maximum likelihood is then picked as the direction of arrival (DoA) for each frame. If there are nf frames, then the mode of the histogram obtained by accumulating the DoA estimates of these frames is declared to be the DoA of the source. Woodruff et al. introduced a method called Binaural-ML where, instead of finding the DoA estimates for each frame, the log-likelihoods of T-F points of all the nf frames are added to obtain a single log-likelihood for each direction. The direction with the maximum log-likelihood is then picked as the DoA.

In this paper, we introduce a template matching approach, where each direction is associated with a unique template. Previously, Nandy et al. [20] used HRTF magnitude spectrum and Zhang et al. [19] used learnt Interaural Matching Filters as the templates corresponding to each direction. In this work, to incorporate the uncertainity associated with the frame-level ITDs obtained for each direction, we use histograms of interaural time difference (ITD) in each frequency subband to generate the templates, as shown in Figure 1. These templates are learnt during the training phase as they are dependent on the object between the two microphones. The motivation is to obtain a representation that is invariant across many environmental scenarios rather than using multi-conditional training [12] or diffuse noise [14] to train the models. In this work, we restrict our study to the localization of speech sources corrupted by additive white Gaussian noise (AWGN). It can be seen in Figure 1 that for a particular direction, as SNR decreases, although the patterns get smeared out they are similar to the patterns obtained under clean condition (SNR = ∞). So, we hypothesize that the templates obtained at SNR = ∞ could be used as the reference templates as the patterns captured in them are invariant to AGWN. In Section 2.2 we show that this hypothesis is indeed valid. Now, given a test speech, the corresponding template is computed and then compared with each of the reference templates. The direction corresponding to the template with maximum match is chosen as the DoA. In the remainder of this paper we will refer to these templates as ITD pattern templates (IPTs). IPT generation and matching are described in detail in Section 2.2. Experiments with Subject_003 from the CIPIC database [23] reveal that the proposed IPT matching performs better than the above described GMM based methods. It should be noted that IPT is obtained by stacking the histograms of ITDs obtained in

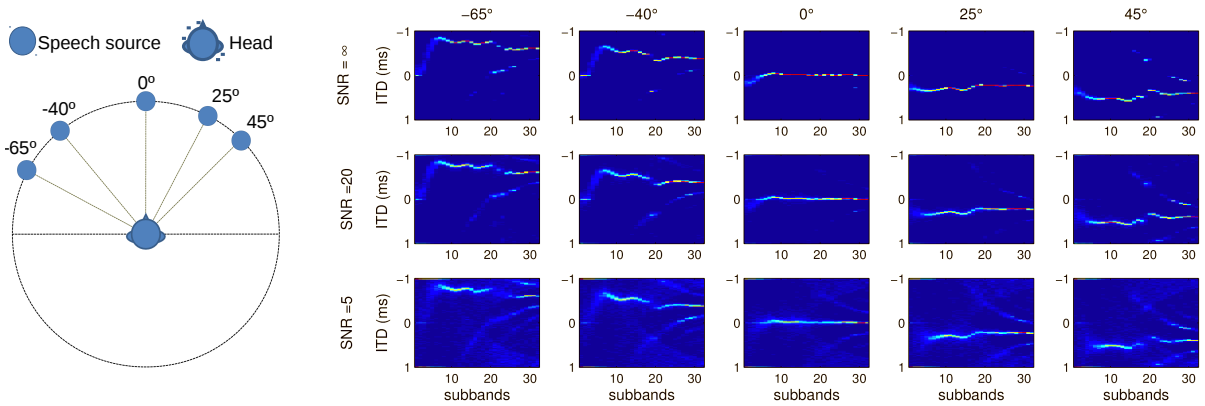

Fig. 1. ITD histograms across subbands (IPTs) for 5 directions at 3 different SNRs (AWGN) for Subject_003 of the CIPIC HRTF database.

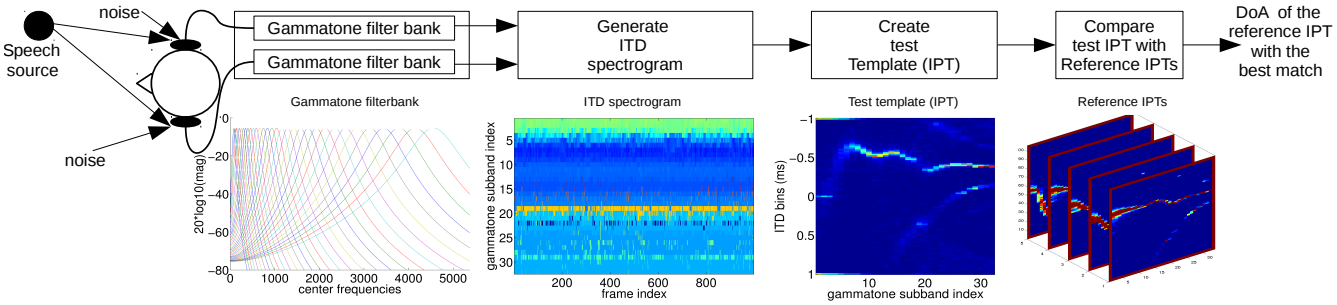

Fig. 2. Block diagram of the proposed IPT matching based localization scheme.

each subband. General histogram algorithms have a complexity of $O(n \times nb)$ where n is the number of T-F points and nb is the number of bins. However, in section 2.3, we show that IPTs can be obtained using a technique with a complexity of $O(n)$.

2. PROPOSED TEMPLATE BASED LOCALIZATION

A given test binaural speech is processed through a set of gammatone filters followed by frame-level ITD computation in each subband as shown in Figure 2. These ITDs are used to generate IPT which is then compared with the reference IPTs. The direction corresponding to the reference IPT with the maximum match is chosen as the DoA estimate. The details of these steps are discussed in the following subsections.

2.1. Gammatone Filters and ITD estimation

The binaural signals are processed through 32 fourth order gammatone filters. Their center frequencies are equally distributed with respect to the equivalent rectangular bandwidth (ERB) scale between 80Hz and 5kHz, starting with 80Hz and ending with 4.6kHz. This range primarily covers the entire speech spectrum. To approximate the neural transduction process of the inner hair cells, the outputs of the gammatone filters are halfwave rectified and square-root compressed [12]. The resulting outputs of the left and right channels of the i^{th} subband are denoted by l_i and r_i. Frame-level ITD in each subband is calculated using normalized cross correlation (NCC) [7,12] between l_i and r_i with a rectangular window of length W and shift of length W_s. $\tau_{i,j}$ is the ITD of i^{th} subband in the j^{th} frame and is given by

$$\tau_{i,j} = \underset{\tau}{\operatorname{argmax}}\, C_{i,j}(\tau), \tag{1}$$

where $C_{i,j}$ is the NCC function. In addition to this, exponential interpolation is used to obtain fractional delays [12].

2.2. IPT Generation, Matching and Localization

***Generation*:** Following ITD estimation, we have the ITD spectrogram as shown in Figure 2 i.e., a matrix of size $ns \times nf$ where ns is the number of gammatone subbands and nf is the number of frames. It contains the ITDs obtained from each subband over all frames. We then generate the IPT, which is obtained by stacking the ITD histograms obtained from each gammatone subband. Hence, IPT is an $nb \times ns$ matrix $T(b,i),\ 1 \le b \le nb, 1 \le i \le ns$, where nb denotes the number of ITD bins as shown in Figure 2.

***Matching and Localization*:** Consider d directions $\theta_1 \cdots \theta_d$ and the corresponding d reference IPTs $T_1 \cdots T_d$ obtained from clean (SNR = ∞) binaural training data. Let T_{test} be the test IPT obtained from test binaural speech. The similarity of T_{test} with a reference template is obtained by taking the sum of all the elements of their Hadamard product [24]. The direction corresponding to the reference template with the largest sum is chosen as the DoA. Therefore, the DoA estimate, $\hat{\theta}$, of the test speech is given by

$$k^* = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{ns} \sum_{b=1}^{nb} (T_{test}(b,i) \times T_k(b,i)) \implies \hat{\theta} = \theta_{k^*} \tag{2}$$

To motivate the use of Hadamard product we consider the clean IPTs of -65° and 25° shown in Figure 3(a) & 3(b). From the colour bars it can be seen that most of the values in the IPTs are close to 0. So, if the non-zero regions of the two IPTs are almost non-overlapping, the Hadamard product of the two templates will be a highly sparse matrix as seen in Figure 3(c). Similarly, if the Hadamard product of every pair of IPTs corresponding to two different directions is highly sparse, then all these templates are non-overlapping. If all the the templates are non-overlapping then it also means that they are unique. Hence, the similarity between a pair

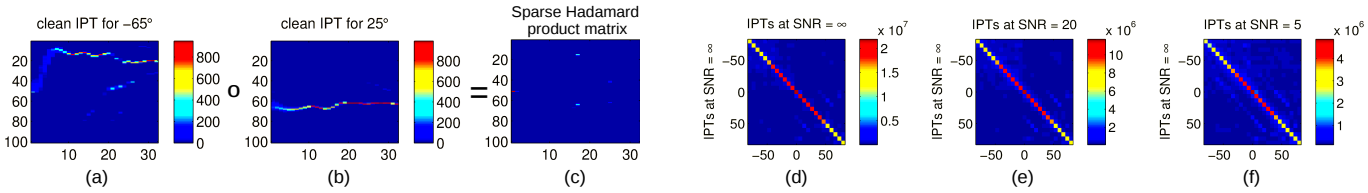

Fig. 3. (a) & (b) are the clean IPTs corresponding to -65° and 25°. (c) is their Hadamard product. (d) is the similarity matrix for all pairs of clean IPTs. (e) and (f) are the similarity matrices between clean IPTs, IPTs at SNR = 20dB and clean IPTs, IPTs at SNR = 5dB respectively. The similarity matrices are computed for the 25 frontal azimuthal directions available in the CIPIC HRTF database for Subject_003.

of IPTs can be obtained by computing the sum of all the elements of their Hadamard product, as defined in eqn. (2).

Consider the clean IPTs obtained for the 25 frontal azimuthal directions, ranging from -80° to 80°, of Subject_003 of the CIPIC database. Let us define S_{SNR_1,SNR_2} as the similarity matrix between a set of 25 IPTs obtained at SNR_1 and SNR_2. The 25×25 matrix, $S_{\infty,\infty}$, is shown in Figure 3(d). It can be seen that, for every row, the maximum occurs on the diagonal and is much higher than the off-diagonal entries. This indicates that the clean IPTs are almost non-overlapping and hence unique for each direction. Since we also want to verify if the clean IPTs can be used as reference templates under AWGN conditions, the similarity matrices between the clean and noisy IPTs are computed. Figures 3(e) & 3(f) show $S_{\infty,20}$ and $S_{\infty,5}$ respectively. $S_{\infty,\infty}$, $S_{\infty,20}$ and $S_{\infty,5}$ are computed using nf=1000. Let us define $r(SNR_1, SNR_2)$ as the ratio of the sum of diagonal elements to the sum of all elements in S_{SNR_1,SNR_2}. We obtain $r(\infty,\infty)$ = 0.7811, $r(\infty,20)$ = 0.6187 and $r(\infty,5)$ = 0.3789. This indicates that, as SNR decreases, the diagonal values decrease with respect to the off-diagonal entires. However, it can be seen in Figures 3(e) & 3(f) that though the the diagonal values have reduced, in every row the maximum is still on the diagonal and their values are much higher than the off-diagonal elements. This indicates that with drop in SNR, the locations of the IPT patterns do not alter much suggesting that the pattern locations are invariant to AWGN noise, although the patterns get smeared out. This, in turn, validates the use of clean IPTs as reference templates for localization under AWGN.

2.3. Complexity of IPT Generation

As seen in Section 2.2, an IPT is obtained by computing ITD histograms for each of the ns subbands. Each histogram contains nb bins of constant bin-width bw and forms each column of the IPT matrix. So, given an ITD from a subband, there could be several methods to assign it to one of the nb bins. One method is to compute the distance of the given ITD from each of the nb bin centers/edges and then find the bin center/edge to which it is the closest. Another method would be to check sequentially for the bin center whose distance from the given ITD is less than half the bin-width bw. Both these methods have a complexity of $O(n \times nb)$.

However, the complexity can be made independent of nb if, given an ITD v, its bin index is computed by

$$i(v) = \left\lceil \frac{v - l}{bw} \right\rceil, \tag{3}$$

where l lower limit of the total ITD range and bw is the bin-width. It can be seen that this method is independent of the number of bins. Hence the complexity of this method is $O(n)$. This histogram implementation makes IPT matching a computationally efficient scheme.

3. EXPERIMENTS AND RESULTS

3.1. Database

Speech from TIMIT database [25] is used for all evaluations. To simulate binaural speech, Head Related Impulse Responses (HRIRs) from the CIPIC database [23] have been used. All experiments have been performed using the HRIRs of Subject_003.

3.2. Experimental Setup

3.2.1. Data preparation

Localization experiments are performed only in the frontal horizontal plane. The CIPIC database consists of HRIRs for 25 directions in the frontal horizontal plane. Speech from the TIMIT database has a sampling frequency of 16kHz, whereas CIPIC HRIRs are sampled at 44.1kHz. Therefore, speech is upsampled to 44.1kHz and then filtered through the HRIRs to obtain binaural speech corresponding to each direction. Frame-level ITDs are calculated using eqn. (1). This is done using a frame duration of 20msec (W = 882) with a shift of 10msec (W_s = 441). While using the NCC function to estimate the time delay, we restrict the maximum delay to 44 samples i.e, a delay of ∼1ms. This is because, in general, any delay > 1ms is not plausible around the human head.

3.2.2. Learning reference IPTs

To train the reference IPTs, we use clean binaural speech (SNR=∞) of duration 10sec. Considering the frame length and shift mentioned in the previous section, this provides 1000 frames to train each of the 25 IPTs corresponding to the 25 frontal azimuthal directions. Each IPT is a matrix where the i^{th} column is a histogram of the ITDs from the i^{th} subband. To obtain each of these histograms, we consider 101 bins of equal width over the ITD interval -1 to 1ms. The histograms are obtained using eqn. (3). Then, we concatenate all the 32 histograms to obtain an IPT of size 101×32. The reference IPTs $T_k, 1 \leq k \leq 25$ for all the 25 directions are thus learnt separately. So, given a test binaural speech, the corresponding IPT, T_{test}, is computed and compared with these reference IPTs, using eqn. (2) to obtain the DoA. It should be noted that there is no need to normalize the reference IPTs as long as each of them are trained using the same number of frames.

3.2.3. Baseline systems

We compare our method with the methods of May et al. [12] and Woodruff et al. [14]. These will be referred to as Binaural Hist and Binaural ML respectively as mentioned in [14]. Both these methods use GMMs to model the ITD distributions in each subband. We train these GMMs using the same 10s of clean binaural speech used for preparing reference IPTs. This provides 1000 frames to train each of the 800 (25 directions × 32 subbands) GMMs for ITD. Expectation Maximization algorithm [26] with random initialization is used for parameter estimation. AIC (Akaike Information Criterion) [27]

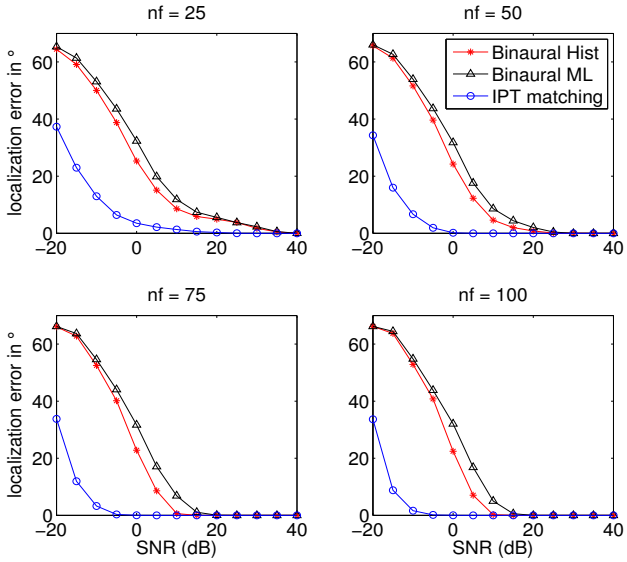

Fig. 4. Localization error vs. SNR for different durations of test speech.

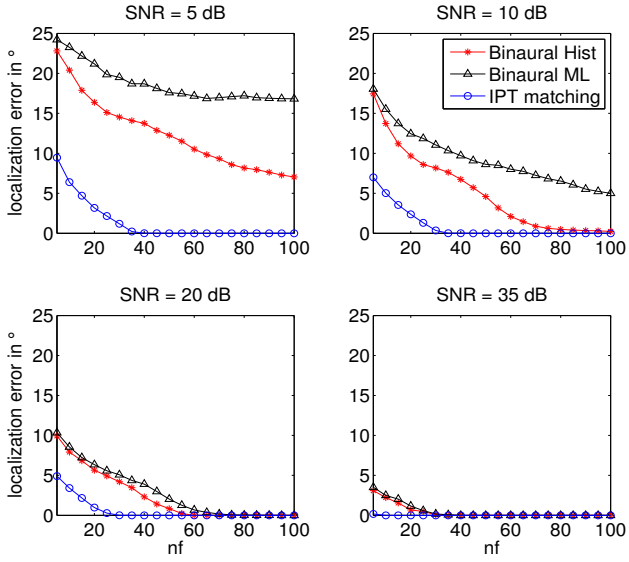

Fig. 5. Localization error vs. duration of test speech for different SNRs.

and BIC (Bayesian Information Criterion) [28] are used to compute the optimal number of Gaussian components. The lower number between the two is chosen as the optimal number of components. However, the maximum number of components is restricted to 20.

The trained GMMs are then used for localization. Binaural Hist estimates the direction in each frame according to

$$\hat{\theta}(j) = \underset{\theta_k}{\operatorname{argmax}} \sum_{i=1}^{ns} \log(P(\tau_{i,j}|\theta_k, i)), \tag{4}$$

where $\hat{\theta}(j)$ is the direction estimate of the j^{th} frame and $P(.|\theta_k, i)$ is the GMM likelihood given the direction θ_k and subband i. It then generates a histogram of the frame-level DoA estimates, $\hat{\theta}(j)$, and picks the mode as the estimated DoA, $\hat{\theta}$, of the source.

Binaural ML, however, sums the likelihood over all the frames and picks the direction with the maximum likelihood according to

$$\hat{\theta} = \underset{\theta_k}{\operatorname{argmax}} \sum_{j=1}^{nf} \sum_{i=1}^{ns} \log(P(\tau_{i,j}|\theta_k, i)), \tag{5}$$

3.3. Results and Evaluation

For evaluating the localization performance of each algorithm, we generate 180 different instances of binaural test speech for each of the 25 directions at every SNR (SNR is varied from -20dB to 40dB with a step of 5dB) using AWGN. Experiments are performed using different number of test speech frames (nf). Average localization error for a particular pair of SNR and nf is then evaluated using

$$e(SNR, nf) = \sum_{k=1}^{25} \sum_{p=1}^{180} |\theta_k - \hat{\theta}(k, p)|. \tag{6}$$

where $\hat{\theta}(k, s)$ is the angle estimated for the p^{th} instance of test binaural speech generated from angle θ_k.

When localizing speech sources, the amount of speech data available to localize the source depends on the scenario. Ideally, we need the algorithm to localize the source with as less data as possible. In our experiments we test the algorithms using speech of durations ranging from 60 ms to 1.01 s. This corresponds to nf = 5 to 100.

3.3.1. *Experiment 1 - Localization error vs. SNR*

For a fixed nf we vary the SNR and measure the average localization accuracy at each SNR. Figure 4 shows the results for nf = 25, 50, 75, 100. It can be seen that IPT matching performs better than the baseline methods. For SNR$\geq$0dB IPT matching achieves zero localization error with nf = 75, 100 and very close to zero for nf = 25, 50.

3.3.2. *Experiment 2 - Localization error vs. number of frames (nf)*

For a fixed SNR, we vary nf from 5 to 100 in steps of 5 and measure the average localization accuracy at each nf. Figure 5 shows the results for SNR = 5,10,20,35 dB. As expected, we can see that the performance of the algorithms gets better with increase in nf. At SNR = 5dB, IPT matching approaches zero localization error at nf = 40. While the other methods do not approach zero error even at nf = 100 for SNR = 5dB.

In all experimental conditions, though the GMMs were also trained on the ITDs of clean binaural speech, it was found that IPTs outperform GMMs.

4. CONCLUSIONS

We presented a new template based localization algorithm which uses templates (IPTs) generated from ITDs to localize the speech sources. Our motivation for this work was to find representations that are invariant to different environmental scenarios rather than training the models using multi-conditional training where the model might fail for untrained scenarios. We observed that the location of the patterns in clean IPTs was well preserved under additive white Gaussian noise (AWGN). This validates the use of clean IPTs for localization at different SNRs of AWGN. We have also presented an $O(n)$ method to compute IPTs which makes it computationally efficient. As part of further analysis, we would like to extend the use of IPTs to reverberant and multiple speech source scenarios.

5. ACKNOWLEDGEMENT

Authors thank the Pratiksha Trust for their support.

6. REFERENCES

[1] Sylvain Argentieri, Patrick Danes, and Philippe Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[2] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, vol. 2, pp. 1228–1233.

[3] Hoang Do, Harvey F Silverman, and Ying Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 1, pp. 121–124.

[4] Jacob Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[5] Yuki Tamai, Satoshi Kagami, Hiroshi Mizoguchi, Yutaka Amemiya, Koichi Nagashima, and Tachio Takano, "Real-time 2 dimensional sound source localization by 128-channel huge microphone array," in *13th IEEE International Workshop on Robot and Human Interactive Communication*, 2004, pp. 65–70.

[6] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[7] Nicoleta Roman, DeLiang Wang, and Guy J Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[8] Christof Faller and Juha Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[9] Harald Viste and Gianpaolo Evangelista, "Binaural source localization," in *Proc. 7th International Conference on Digital Audio Effects (DAFx-04), invited paper*, 2004, number LCAV-CONF-2004-029, pp. 145–150.

[10] Volker Willert, Julian Eggert, Jürgen Adamy, Raphael Stahl, and E Korner, "A probabilistic model for binaural sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 5, pp. 982–994, 2006.

[11] Martin Raspaud, Harald Viste, and Gianpaolo Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[12] Tobias May, Steven van de Par, and Armin Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language processing*, vol. 19, no. 1, pp. 1–13, 2011.

[13] Antoine Deleforge and Radu Horaud, "2D sound-source localization on the binaural manifold," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[14] John Woodruff and DeLiang Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[15] Fakheredine Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, 2014.

[16] Xuan Zhong, Liang Sun, and William Yost, "Active binaural localization of multiple sound sources," *Robotics and Autonomous Systems*, vol. 85, pp. 83–92, 2016.

[17] Mehdi Zohourian and Rainer Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 430–434.

[18] Girija Ramesan Karthik and Prasanta Kumar Ghosh, "Subband selection for binaural speech source localization," *Proc. Interspeech 2017*, pp. 1929–1933.

[19] Jie Zhang and Hong Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4771–4783, 2015.

[20] Dibyendu Nandy and Jezekiel Ben-Arie, "Estimating the azimuth of a sound source from the binaural spectral amplitude," *IEEE transactions on speech and audio processing*, vol. 4, no. 1, pp. 45–55, 1996.

[21] C Phillip Brown and Richard O Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 5, pp. 476–488, 1998.

[22] DeLiang Wang and Guy J Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," 2006.

[23] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

[24] Roger A Horn, "The hadamard product," *Proc. Symp. Appl. Math*, vol. 40, pp. 87–169, 1990.

[25] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[26] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[27] Hirotugu Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[28] Gideon Schwarz et al., "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.