INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Subband selection for binaural speech source localization

Girija Ramesan Karthik, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bengaluru-560012, India
grkarthik@ee.iisc.ernet.in, prasantg@ee.iisc.ernet.in

Abstract

We consider the task of speech source localization using binaural cues, namely interaural time and level difference (ITD & ILD). A typical approach is to process binaural speech using gammatone filters and calculate frame-level ITD and ILD in each subband. The ITD, ILD and their combination (ITLD) in each subband are statistically modelled using Gaussian mixture models for every direction during training. Given a binaural test-speech, the source is localized using maximum likelihood criterion assuming that the binaural cues in each subband are independent. We, in this work, investigate the robustness of each subband for localization and compare their performance against the full-band scheme with 32 gammatone filters. We propose a subband selection procedure using the training data where subbands are rank ordered based on their localization performance. Experiments on Subject_003 from the CIPIC database reveal that, for high SNRs, the ITD and ITLD of just one subband centered at 296Hz is sufficient to yield localization accuracy identical to that of the full-band scheme with a test-speech of duration 1sec. At low SNRs, in case of ITD, the selected subbands are found to perform better than the full-band scheme.

Index Terms: gammatone filters, interaural time difference, interaural level difference

1. Introduction

Machine localization of sound sources is necessary for a wide range of applications, including human-robot interaction, surveillance and hearing aids. Robot sound localization algorithms have been proposed using microphone arrays with varied number of microphones [1–6]. Adding more microphones helps increase the localization performance as more spatial cues can be obtained based on the number and arrangement of the microphones. However, humans have an incredible ability to localize sounds with just two ears. The major cues that help in localization are interaural time difference (ITD), interaural level difference (ILD) and spectral coloration. These cues can be captured by the head-related transfer function (HRTF) [7]. Its equivalent in the time domain is the head-related impulse response (HRIR). An algorithm inspired by binaural localization of humans would extract these features from the input signals [8–20].

High frequencies, unlike low frequencies, fail to diffract around the head to reach the contralateral ear. This causes a prominent intensity difference between the two ears making ILD more robust for localization at high frequencies compared to ITD. Due to phase wrapping, many ITDs correspond to a single phase difference. For a given distance between the two ears, there exists a range of plausible ITDs depending on the maximum delay between the two ears. At low frequencies, only one possible ITD lies in this range. But as the frequency increases, many ITDs start falling in the plausible range causing ambiguity in ITD estimation. Hence, ITD is more robust for localization at low frequencies [8].

To account for the time and frequency variability of these cues, time-frequency representations of the binaural signals are used. One of the most common time-frequency representations is the Short-Time Fourier Transform (STFT) [10, 12, 14, 16, 18] which assumes uniform subband width and spacing in the frequency domain. Another approach is to use gammatone filters [21] where the subband width and spacing are not uniform [8, 11, 13, 15]. 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.

May et al. [13] use Gaussian mixture models (GMMs) to model ITD, ILD and their joint distribution (ITLD) for each gammatone subband in each direction. Then, for a test-speech, log-likelihoods are calculated 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. In this work, we investigate the localization accuracy of each subband rather than combining the likelihoods of all the subbands (full-band scheme). DoA estimation can be treated as a classification problem where, given a feature (ITD or ILD), the latent class (DoA) needs to be inferred. For a subband, each direction will have its own distribution of ITD and ILD. Higher discrimination among distributions of different directions results in better classification accuracy. Hence, subbands with a high level of discrimination are more reliable than the ones with a low level of discrimination. Addition of noise could decrease this discrimination and lead to a decrease in localization accuracy. We hypothesize that choosing the most reliable subbands and discarding the rest can improve localization accuracy. Using localization error as a measure of discrimination, we select a set of subbands from the clean training data and examine how they perform in noisy conditions. Experiments with Subject_003 from the CIPIC database [22] reveal that ITD and ITLD of one subband centered at 296Hz yield a localization accuracy identical to that of the full-band scheme for a test binaural speech of duration 1sec. Such a subband selection also reduces the computational complexity.

2. Binaural Cue Extraction and Localization

DoA estimation consists of the following steps. First, the binaural speech is processed through a set of gammatone filters followed by frame-level ITD and ILD computation in each subband. These binaural features are then processed through GMMs trained on each subband for each direction. The direction with the maximum likelihood is the DoA estimate. We provide the details of these steps in the following subsections.

2.1. Gammatone Filters

The binaural signals are processed through N=32 fourth order gammatone filters. Their center frequencies are equally dis-

http://dx.doi.org/10.21437/Interspeech.2017-954

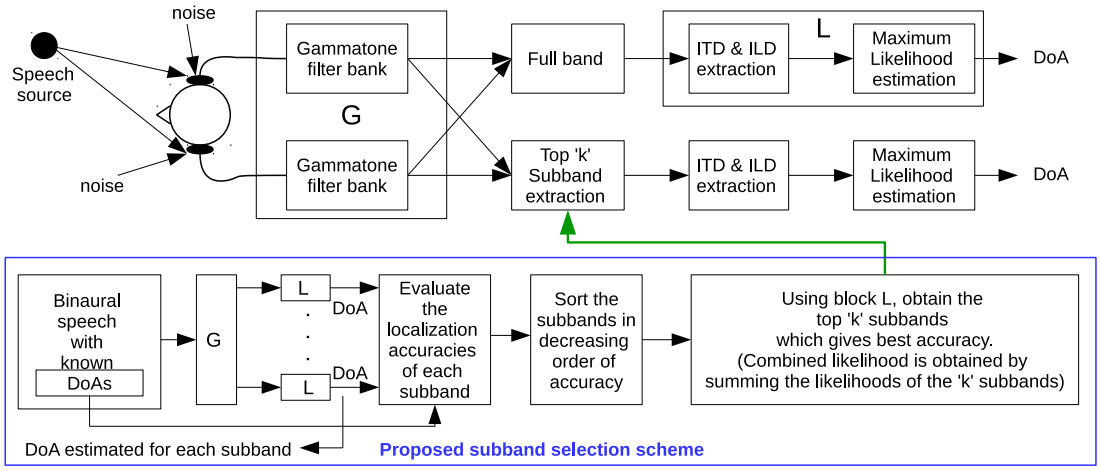

Figure 1: *Localization setup and the proposed subband selection scheme*

tributed 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 [13]. The resulting outputs of the left and right channels of the i^{th} subband are denoted by l_i and r_i.

2.2. ITD and ILD

Frame-level ITD in each subband is then calculated using normalized cross correlation (NCC) [8, 13] 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 [13]. ILD is obtained by taking the ratio of the energies of the signals in each gammatone subband as follows:

$$L_{i,j} = 20\log_{10}\left(\frac{\sum_{n=0}^{W-1} {l_i}^2[W_s\cdot(j-1)+n]}{\sum_{n=0}^{W-1} {r_i}^2[W_s\cdot(j-1)+n]}\right), \tag{2}$$

where $L_{i,j}$ is the ILD of i^{th} subband in the j^{th} frame.

2.3. GMM Parameter Estimation

GMMs are trained on the binaural cues of each subband in each direction. Separate GMMs are trained for ITD, ILD and their joint distribution (ITLD). Binaural features of different subbands span different regions in the ITD-ILD space. Information theoretic criteria such as Akaike Information Criterion (AIC) [23] and Bayesian Information Criterion (BIC) [24] are used to evaluate the optimum number of components in each GMM. The optimal number of components is obtained using AIC as well as BIC. The lower number between the two is chosen as the optimal number of components.

Given a GMM with K components, the probability density function (pdf) of input feature $\mathbf{x}$ is given by

$$p(\mathbf{x}|\lambda) = \sum_{c=1}^{K} w_c p_c(\mathbf{x};\mu_c;\mathbf{\Sigma}_c), \tag{3}$$

where $\lambda = \{w_c, \mu_c, \mathbf{\Sigma}_c : c = 1, ..., K\}$ is the set of all parameters of the GMM. w_c μ_c and $\mathbf{\Sigma}_c$ are the weight, mean and covariance matrix of the c^{th} component. p_c is the pdf given the c^{th} component.

2.4. Likelihood and Localization

Suppose we consider D directions. Then $\lambda_{i,d}$ is the set of GMM parameters for the i^{th} subband in the d^{th} direction where d ranges from 1 to D. Let $\mathbf{x}_{i,j}$ denote the binaural feature (ITD/ILD/ITLD) of the i^{th} subband in the j^{th} frame. Then, $p(\mathbf{x}_{i,j}|\lambda_{i,d})$ is calculated $\forall i, d$. May et al. [13] combined the likelihoods of all the subbands (full-band scheme), for each d to obtain a single likelihood for each direction. And then, the direction with the maximum likelihood is chosen as the DoA estimate in the j^{th} frame.

$$DoA_j = \underset{d\in\{1,..,D\}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\mathbf{x}_{i,j}|\lambda_{i,d}). \tag{4}$$

The DoAs from multiple frames are pooled to obtain the DoA with the maximum frequency of occurrence.

3. Proposed Subband Selection

Interestingly, DoA can also be estimated from individual subbands.

$$DoA_{i,j} = \underset{d\in\{1,..,D\}}{\operatorname{argmax}} \log p(\mathbf{x}_{i,j}|\lambda_{i,d}), \tag{5}$$

where $DoA_{i,j}$ is the DoA estimate of the i^{th} subband in the j^{th} frame. Different subbands, in general, would have different localization accuracies. Adding the log-likelihoods of all subbands may include unreliable subbands which can degrade the localization performance. So, we want to select the least set of k subbands that achieves the best accuracy. The steps of the proposed subband selection are summarized in Figure 1. We use the subband specific localization accuracies to sort the subbands in the decreasing order of their accuracy, as shown in Figure 1. Using the first k sorted subbands, localization accuracy is evaluated. The log-likelihoods of these k subbands are added to obtain the combined log-likelihood in each frame. This is done for k varying from 1 to 32. Finally, from the obtained accuracies, the least value of k is chosen for which the accuracy is maximum. This is done using clean binaural speech.

4. Experiments and Results

4.1. Database

Speech from TIMIT database [25] is used for all evaluations. To simulate binaural speech, HRIRs from CIPIC database [22] have been used. All experiments have been performed using the HRIRs of Subject_003.

4.2. Experimental Setup

4.2.1. Binaural speech data preparation

Localization experiments are performed only in the frontal horizontal plane. The CIPIC database consists of HRTFs of 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.

4.2.2. ITD, ILD extraction & GMM parameter estimation

Frame-level ITDs and ILDs are calculated using eqn. (1) and (2) respectively. This is done using a frame duration of 20msec (W = 882) with a shift of 10msec (W_s = 441). To train the GMMs, frame-level ITDs and ILDs are computed using a training binaural speech of duration 10sec. This provides 1000 frames to train each of the 800 (25 directions $\times$ 32 subbands) GMMs for ITD, ILD and ITLD. As the natural clusters of the binaural features in the ITD-ILD space are elongated in the direction of the ILD (vertical) axis, K-means can split the data horizontally to form clusters. Hence, for GMMs of ITLD features, EM algorithm [26] with random initialization is used for parameter estimation. Diagonal covariance matrix is used since the clusters are oriented parallel to the ILD axis [13]. As described in Section 2.3, AIC and BIC are used to compute the optimal number of Gaussian components. However, the maximum number of components is restricted to 20.

4.2.3. Localization error

Let ϕ be the actual azimuthal angle and $\hat{\phi}$ be the estimated angle. Then the localization error is

$$e = |\phi - \hat{\phi}|. \tag{6}$$

Average localization error is obtained by taking the mean of the localization errors from ns different binaural speech segments. So, the average localization error of a particular subband is

$$e_{av}(i) = \frac{1}{ns}\sum_{j=1}^{ns} e_{i,j}, \tag{7}$$

where $e_{i,j}$ is the error in the i^{th} subband for the j^{th} speech segment. In our experiments we consider ns = 45. The localization errors are calculated in degrees.

4.3. Results and Discussion

The trained GMMs, on an average, have approximately 5 components with a maximum of 11 and minimum of 1.

4.3.1. Subband selection

Clean binaural speech of 1sec, i.e., 100 frames with known DoAs is passed through the subband selection block shown in Figure 1. DoA is estimated for every subband in each frame using eqn. (5). The DoA estimates of the 100 frames are pooled to obtain the DoA with maximum frequency of occurrence. From the obtained DoAs, the average localization error of each subband is computed. Figure 2 shows the localization errors of ITLD, ITD and ILD separately for all the subbands. The subbands are then sorted in the decreasing order of their accuracies. The sorted subband indices are provided within each subplot in Figure 2. It can be seen that, in case of ITD and ILD, the subbands with the highest accuracy are in the low and high frequency ranges respectively. For ITD, subbands with center frequencies from 296 Hz to 1.04 kHz have zero localization error indicating that one of these subbands might be sufficient to achieve the least localization error. The best k subbands are then selected using the subband selection scheme provided in Section 3. As shown in Table 1, in case of ITD and ITLD, we see that one subband is sufficient to achieve localization accuracy identical to that of the full-band scheme (k = 32). This corresponds to the subband with center frequency 296Hz, as shown in Figure 2. Similarly, ILDs of the top six subbands are sufficient to achieve the full-band accuracy. They are subbands with center frequencies 4.56, 2.59, 2.85, 4.15, 2.35 and 3.78kHz. The set of k optimal subbands is denoted by S_k. Similarly, the full-band set is denoted by F.

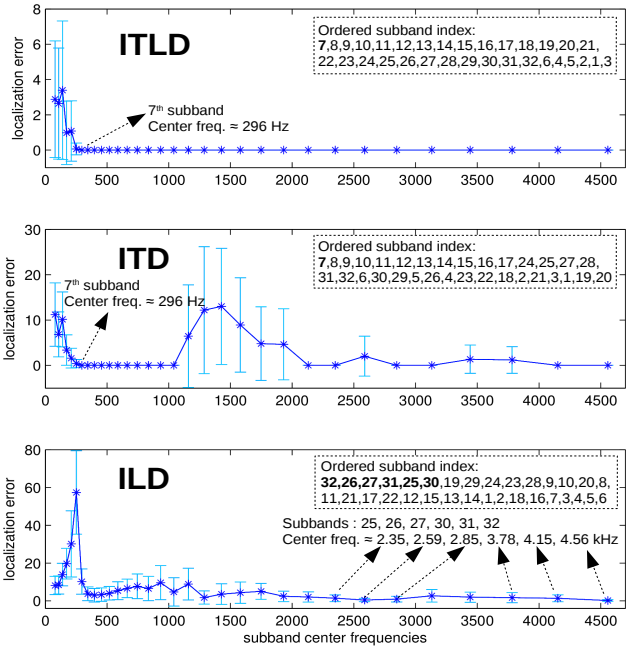

Figure 2: *Localization error (in degrees) across subbands using ITLD, ITD and ILD. Ordered subband index: Subbands are sorted in the increasing order of localization error.*

Table 1: *Localization error (in degrees) vs number of ordered subbands (k) for clean speech of duration 1sec.*

	no. of ordered subbands (k)								
	1	**2**	**5**	**6**	**10**	**15**	**20**	**25**	**32**
ITLD	0	0	0	0	0	0	0	0	0
ITD	0	0	0	0	0	0	0	0	0
ILD	0.21	0.04	0.03	0	0	0	0	0	0

4.3.2. Localization error on test binaural speech

We evaluate the localization performance of the full-band and the selected subbands on the test set using the trained GMMs. We consider noisy test conditions at different SNRs namely, 40, 20, 10 and 5dB by adding white Gaussian noise. We also

Table 2: *Localization error (in degrees) comparison between the selected-subband scheme and the full-band scheme for various SNRs and test-speech durations.*

		Clean				SNR = 40				SNR = 20				SNR = 10				SNR = 5			
no. of Frames (*nf*)		**1**	**10**	**50**	**100**	**1**	**10**	**50**	**100**	**1**	**10**	**50**	**100**	**1**	**10**	**50**	**100**	**1**	**10**	**50**	**100**
ITLD	S_1	0.7	0.1	0.0	0.0	5.3	2.7	0.1	0.0	16.6	11.7	1.7	0.0	20.3	17.4	5.0	0.9	26.0	21.1	9.4	4.7
	F	0.0	0.0	0.0	0.0	0.8	0.2	0.0	0.0	6.5	3.7	0.0	0.0	10.2	6.3	1.3	0.0	12.7	10.6	3.9	1.7
ITD	S_1	0.7	0.2	0.0	0.0	5.4	3.3	0.1	0.0	18.1	14.5	2.7	0.0	21.3	21.2	7.2	1.3	28.0	27.2	13.8	7.0
	F	0.0	0.0	0.0	0.0	2.4	1.1	0.0	0.0	13.4	9.1	2.1	0.0	22.0	16.7	8.7	3.2	27.5	23.7	17.9	12.5
ILD	S_6	0.2	0.0	0.0	0.0	1.6	0.5	0.0	0.0	6.5	4.8	3.0	1.9	9.9	8.8	8.4	8.6	11.8	11.0	11.0	11.3
	F	0.0	0.0	0.0	0.0	1.7	0.9	0.0	0.0	7.1	5.5	1.6	0.6	12.1	10.3	6.0	4.5	14.3	13.7	10.2	8.4

Table 3: *Localization error (in degrees) vs SNR for different values of k. (nf = 100)*

		ITLD					ITD					ILD				
SNR		**Clean**	**40**	**20**	**10**	**5**	**Clean**	**40**	**20**	**10**	**5**	**Clean**	**40**	**20**	**10**	**5**
no. of ordered subbands (*k*)	**1**	0.00	0.00	0.00	0.91	4.69	0.00	0.00	0.02	1.34	7.03	0.27	0.60	5.45	13.54	15.78
	2	0.00	0.00	0.00	0.29	1.28	0.00	0.00	0.00	0.08	1.30	0.03	0.17	3.97	9.43	11.14
	6	0.00	0.00	0.00	0.16	1.80	0.00	0.00	0.00	0.86	3.52	0.00	0.00	1.89	8.58	11.32
	10	0.00	0.00	0.00	0.13	0.77	0.00	0.00	0.00	0.06	0.53	0.00	0.00	0.72	6.57	10.07
	15	0.00	0.00	0.00	0.23	1.86	0.00	0.00	0.00	0.01	2.44	0.00	0.00	0.60	5.65	9.37
	20	0.00	0.00	0.00	0.00	2.51	0.00	0.00	0.00	0.00	3.23	0.00	0.00	0.47	4.66	8.47
	25	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00	1.05	8.84	0.00	0.00	0.56	4.40	7.95
	28	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	2.12	10.22	0.00	0.00	0.67	4.04	7.13
	32	0.00	0.00	0.00	0.00	1.70	0.00	0.00	0.00	3.17	12.47	0.00	0.00	0.60	4.53	8.44

consider different durations (denoted in number of frames, nf) of the test speech. Below, we report the localization performance on the test speech in three experimental conditions - Experiment-1) comparison of full-band and selected subband schemes with nf=100 at different SNRs, Experiment-2) repeating experiment-1 for nf=1, 10, 50, Experiment-3) localization performance by varying the top k ordered subbands at different SNRs.

Experiment-1: For nf = 100, the localization performance of the full-band scheme and the selected subband scheme obtained using ITLD, ITD and ILD are shown as green columns in Table 2. It can be seen that, at all SNRs, ITD based localization using just the selected single subband performs as good as or even better than the full-band scheme. For SNR$\geq$20 in case of ITLD and for SNR$>$20 in case of ILD, the performance using the selected subbands is the same as that of the full-band scheme.

Experiment-2: The localization performance for this experiment is shown in Table 2. In case of ITLD and ITD, at SNR = ∞ (clean signal), one subband is enough to achieve localization performance close to that of the full-band scheme for nf = 1, 10 and equal to the full-band scheme for nf = 50, 100. In case of ITD and ILD, at all other SNRs, the performance of the full-band and the selected subbands are comparable. The selected subbands sometimes perform better than the full-band scheme. Hence, using a very few subbands we are able to achieve good localization accuracy with reduced computation. In case of ITLD, on an average, the performance of the full-band is better than the selected subbands.

Experiment-3: To understand the effect of choosing more number of subbands, we evaluate the localization performance of the first k subbands from the ordered list for different values of k. Table 3 shows this analysis at different SNRs using nf = 100. The row with k = 32 corresponds to the full-band scheme. The best number of subbands at different SNRs is shown in green in every column of Table 3. As seen in Table 3, every column has a value of k $<$32 whose localization performance is equal to or better than that of the full-band scheme. In case of ITD, the performance of the first subband alone, i.e., 296Hz is better than the full-band scheme for almost all SNRs. Another important observation is that, as SNR decreases in the case of ITD, choosing more number of subbands degrades the performance (Table 3: red cells). In case of ILD, increase in the number of subbands increases the performance till 28 subbands, after which there is a slight degradation. However, the performance of ITD of the first subband is better than the performance of ILD with the full-band scheme. Incorporating subband selection reduces computational complexity, as the amount of preprocessing (gammatone filtering) and likelihood computations get considerably reduced. In cases where a single subband is sufficient to achieve the best accuracy, the number of computations reduces by a factor of 32.

5. Conclusions

We present a localization error based subband selection method and experiments conducted using this method reveal that subband selection is useful for improving binaural speech localization accuracy and reducing the computational complexity. It would be interesting to understand the order in which the subbands get affected by noise and also the variability of the most reliable subbands across different subjects. Also, to improve the accuracy for ITLD based localization, the best subbands of ITD and ILD could be combined. During training, many subbands have zero localization error in the clean case using ITD and ITLD. A quantified measure of discrimination between the distributions of ITD for all directions in each subband could be used to sort these subbands.

6. Acknowledgement

Financial support for this paper by the Robert Bosch Centre for Cyber-Physical Systems at the Indian Institute of Science, Bengaluru, is gratefully acknowledged.

7. References

[1] S. Argentieri, P. Danes, and P. 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, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, 2003, pp. 1228–1233.

[3] H. Do, H. F. Silverman, and Y. 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)*, vol. 1, 2007, pp. 121–124.

[4] J. 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] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, and T. 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] D. Pavlidi, A. Griffin, M. Puigt, and A. 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] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Transactions on Speech and Audio processing*, vol. 6, no. 5, pp. 476–488, 1998.

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

[9] C. Faller and J. 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.

[10] H. Viste and G. Evangelista, "Binaural source localization," in *Proc. 7th International Conference on Digital Audio Effects (DAFx-04), invited paper*, no. LCAV-CONF-2004-029, 2004, pp. 145–150.

[11] V. Willert, J. Eggert, J. Adamy, R. 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.

[12] M. Raspaud, H. Viste, and G. 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.

[13] T. May, S. van de Par, and A. 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.

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

[15] J. Woodruff and D. 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.

[16] F. 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.

[17] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1207–1217, 2014.

[18] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[19] X. Zhong, L. Sun, and W. Yost, "Active binaural localization of multiple sound sources," *Robotics and Autonomous Systems*, vol. 85, pp. 83–92, 2016.

[20] M. Zohourian and R. 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.

[21] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," 2006.

[22] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.

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

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

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. 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] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.