AN UNSUPERVISED BAYESIAN CLASSIFIER FOR MULTIPLE SPEAKER DETECTION AND LOCALIZATION

Youssef Oualil, Friedrich Faubel, Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany

youssef.oualil@lsv.uni-saarland.de

Abstract

Multiple speaker localization algorithms generally require a binary detector, which performs the source/noise classification of the location estimates. This is mainly due to the unknown time-varying number of sources, and to the presence of noise and reverberation. In this paper, we propose an unsupervised learning approach based on a naive Bayesian classifier. The proposed approach couples two speaker location features, namely, 1) the steered response power introduced at the location estimate, and 2) the corresponding maximum likelihood error, which characterizes the variance of the estimate. The latter is experimentally shown to be highly correlated with the steered power at the location estimate. The proposed method is further extended to control the misclassification rate through the use of a loss function. This approach is general, and can be easily extended to integrate more speaker/speech features. Experiments on the AV16.3 corpus show the effectiveness of the proposed approach.

Index Terms: microphone arrays, multiple speaker localization, source detection, Bayesian classification.

1. Introduction

Microphone arrays have become an essential tool for a large number of signal processing problems. Their area of application includes speech separation/enhancement, acoustic source localization and tracking, but also more advanced approaches such as camera steering for teleconference systems and audio-visual tracking. Among these applications, the detection and localization of multiple concurrent speakers from a short segment of speech remains a difficult and open task; and that although an abundance of localization methods have been proposed in the literature: multi-channel cross correlation (MCCC) [1], adaptive eigenvalue decomposition (ED) [2, 3, 4], time difference of arrival (TDOA)-based techniques [5, 6, 7] and steered response power (SRP)-based techniques [8, 9], just to name a few.

A good multiple speaker localization performance cannot be achieved without a source detector, which classifies the obtained estimates to speaker/noise. This is mainly due to 1) the presence of noise and/or reverberation, which introduces secondary peaks, and to 2) the unknown time-varying number of sources per frame. Few attempts have been made to overcome this problem, the authors of [10] proposed to use the distance separating the estimates as a criterion to extract the number and location of the sources, whereas Do et *al.* [11, 12] proposed to combine the signal power with a double clustering technique to estimate the number of speakers. In a more advanced approach, Lathoud et *al.* [13] proposed an unsupervised threshold selection technique to control the false alarm rate.

Following a line of thought similar to [13], we propose to estimate the optimal boundary between the noise and speaker classes, using an unsupervised Bayesian classifier. Contrary to the approaches taken in [12, 13], where a *single* power-based feature is used, we propose in this work to augment the feature space with the Maximum Likelihood Error (MLE), introduced at each location estimate. In doing so, the classification boundary between the two classes becomes more obvious. This property is of most interest in low SNR/SRR environments, as well as in the multiple speaker case, where the signal power emerging from the secondary speakers becomes comparable to the noise/reverberation power.

In this framework, we first estimate the likelihood distribution and the prior of each class. This is done by fitting a 3-components mixture to each feature space. Then, the posterior distribution of each class is obtained using a Naive Bayesian Classifier (NBC), which combines the two features. The choice of the mixtures is dependent on the used features. Experiments conducted on the AV16.3 corpus show that 1) combining the features improves the detection performance, and that 2) the proposed unsupervised classification approach performs better than a supervised Support Vector Machine (SVM) classifier.

We proceed in this paper by introducing the classification features. Then, we show how these features can be used to estimate the likelihood distributions and the priors (Section 2). The unsupervised Bayesian classifier is presented in Section 3. Section 4 shows the performance of the proposed approach in comparison with SVM. Finally, we conclude in Section 5.

2. Features Extraction And ML Estimation

In this section, we proceed by reviewing the multiple speaker localization approach used to estimate the source(s) location, and thereby extract the classification features. Then, we show how the mixture distributions can be used to characterize each feature space. Finally, we propose an online algorithm to estimate the parameters of these mixtures.

2.1. Multiple Speaker Localization Approach

In a recent work [14, 15], we have proposed a novel approach to the multiple source localization problem. This framework interprets each normalized Generalized Cross Correlation function (GCC) as a probability density function (pdf) of the Time Difference of Arrival (TDOA). This pdf is then approximated by a Gaussian mixture (GM) distribution using either the Weighted Expectation Maximization (WEM) algorithm from [15] or its practical approximation in [14]. The resulting TDOA Gaussian mixtures are mapped to the location space using the location-TDOA mapping, given by (1). The approach proposed in [14] combines the GMs using a probabilistic interpretation of the Steered Response Power (pSRP), whereas the approach proposed in [15] maximizes the TDOA joint pdf in the location

space. The rest of this section presents a brief introduction to the mathematical formulation of these two frameworks.

Formally, let M and Q denote the number of microphones and corresponding pairs, respectively, and let $\mathbf{m}_h$ denote the positions of the microphones, $h = 1, \ldots, M$. The location-TDOA mapping between the location $\mathbf{s}$ and the TDOA $\tau_q(\mathbf{s})$, introduced by the source $\mathbf{s}$ at the microphone pair $q = \{\mathbf{m}_g, \mathbf{m}_h\}$ is given by

$$\tau_q(\mathbf{s}) = (\|\mathbf{s} - \mathbf{m}_h\| - \|\mathbf{s} - \mathbf{m}_g\|) \cdot c^{-1} \quad (1)$$

where c denotes the speed of sound in the air.
The GM approximating the normalized GCC function (interpreted as a pdf of the TDOA) of the q-th microphone pair, is given by

$$p(\tau^q) = \sum_{k=1}^{K^q} w_k^q \cdot \mathcal{N}_k^q(\tau_q, \mu_k^q, (\sigma_k^q)^2) \quad (2)$$

where μ_k^q, σ_k^q and w_k^q denote the mean, standard deviation and mixture weight of the k-th component. The probabilistic SRP SRP_{prob} of a given location $\mathbf{s}$ is then given by [14]

$$SRP_{prob}(\mathbf{s}) \propto \sum_{q=1}^{Q} \sum_{k=1}^{K^q} w_k^q \cdot \mathcal{N}_k^q(\tau_q(\mathbf{s}), \mu_k^q, (\sigma_k^q)^2) \quad (3)$$

whereas the ML approach maximizes the location likelihood distribution given by [15]

$$p(\mathbf{s}) \approx \prod_{q=1}^{Q} \sum_{k=1}^{K^q} w_k^q \cdot \mathcal{N}_k^q(\tau^q(\mathbf{s}), \mu_k^q, (\sigma_k^q)^2) \quad (4)$$

The source location estimate $\mathbf{s}_e$ is obtained by 1) extracting from each GM distribution the Gaussian component $(w_{\mathbf{s}_e}^q, \mu_{\mathbf{s}_e}^q, \sigma_{\mathbf{s}_e}^q)$ where the source is dominant. Then, 2) calculating the restriction of (3) and (4) on the space region where $\mathbf{s}_e$ is dominant. Finally, 3) the optimal location estimate is obtained using a numerical optimization algorithm. These two approaches however use two different detection methods to classify a location estimate $\mathbf{s}_e$ to source/noise estimate. In [14], the decision is based on the probabilistic power coming from that particular location, that is

$$\mathbf{s}_e \text{ is a source if } SRP_{prob}(\mathbf{s}_e) > P_{noise} \quad (5)$$

where P_{noise} is a predefined threshold, whereas [15] accomplishes this task by comparing the MLE $\epsilon(\mathbf{s})$ to a predefined threshold Γ. This is done according to

$$\mathbf{s}_e \text{ is a source if } \epsilon(\mathbf{s}_e) = \sum_{q=1}^{Q} \left(\frac{\tau^q(\mathbf{s}_e) - \mu_{\mathbf{s}_e}^q}{\sigma_{\mathbf{s}_e}^q} \right)^2 < \Gamma \quad (6)$$

The difficulty with these two detection approaches lies in the choice of the decision thresholds. The latter is dependent on the environment and the distance to the microphone array. Therefore, a static threshold might not be well suited to the location changes, as it might poorly perform in unseen environments. In this work, we propose an unsupervised learning approach, which improves the detection performance by combining these two approaches. This approach is easy to adapt to possible location changes, using an online learning process, and provides different decision boundaries in different environments.

2.2. Cumulative Steered Response Power Feature

Similarly to the approach taken in [14], we propose to use the steered power as the first detection feature. This approach however, does not simply consider the power coming from a single location, it rather considers the cumulative power emerging from the estimate region of dominance. This cumulative steered response power (CSRP) is calculated according to

$$CSRP_{\mathbf{s}_e} = \int_{\mathcal{S}_e} SRP_{prob}(\mathbf{s}) \cdot \mathrm{d}\mathbf{s} \quad (7)$$

$$\approx \int_{\mathcal{S}} \sum_{q=1}^{Q} w_{\mathbf{s}_e}^q \cdot \mathcal{N}_{\mathbf{s}_e}^q(\tau_q(\mathbf{s}), \mu_{\mathbf{s}_e}^q, (\sigma_{\mathbf{s}_e}^q)^2) \cdot \mathrm{d}\mathbf{s} \approx \sum_{q=1}^{Q} w_{\mathbf{s}_e}^q \quad (8)$$

$\mathcal{S}_e$ represents the space region where the acoustic event generating $\mathbf{s}_e$ is dominant. The equation (8) is obtained by mapping $\mathcal{S}$ to the different TDOA spaces (see [15] for more details).

Let $\{(\mathbf{s}_i, c_i)\}_{i=1}^{N_T}$ denote the set of N_T location estimates $\mathbf{s}_i$ and their corresponding CSRP values c_i, obtained in T frames. We propose to separate the source from the noise by fitting a 3-components mixture distribution to the data in the CSRP space. This mixture is obtained by maximizing the likelihood of the CSRP estimates $\{c_i\}_{i=1}^{N_T}$ using the Expectation-Maximization (EM) algorithm [16].
Formally, the EM algorithm estimates a mixture distribution of the form

$$f^{csrp}(\mathbf{s}) = w_n^{csrp} \cdot \mathcal{G}_n^{csrp}(c) + w_s^{csrp} \cdot f_s^{csrp}(c) \quad (9)$$

$\mathcal{G}_n^{csrp}(.)$ is a Gaussian distribution approximating the likelihood distributions of the noise, whereas $f_s^{csrp}(.)$ is a "Gaussian + uniform" mixture distribution approximating the likelihood distribution of the sources (Figure 1-b). $f_s^{csrp}(.)$ is given by

$$f_s^{csrp}(c) \propto \mathcal{G}_s^{csrp}(c) + \mathcal{U}_s^{csrp}(c) \quad (10)$$

where $\mathcal{G}_s^{csrp}(.)$ is a Gaussian distribution and $\mathcal{U}_s^{csrp}$ is a uniform distribution. w_n^{csrp} and w_s^{csrp} denote the noise and source priors, respectively. The uniform distribution $\mathcal{U}_s^{csrp}$ is introduced to model the high CSRP values, which are poorly modeled by $\mathcal{G}_s^{csrp}$.

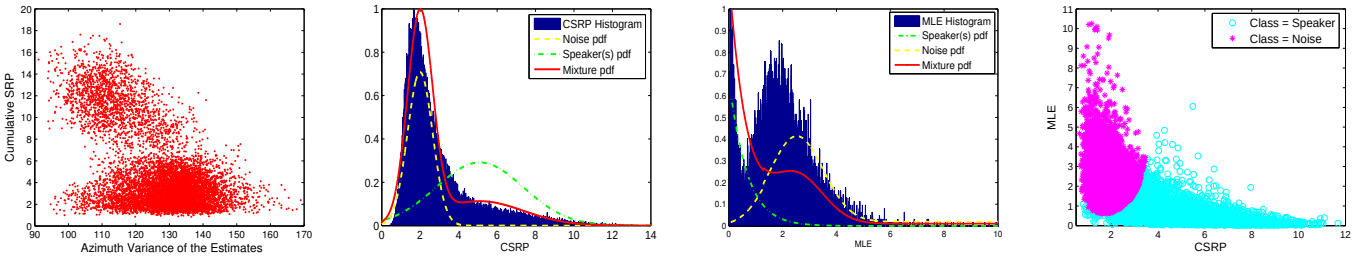
2.3. Maximum Likelihood Error Feature

The second classification feature is the Maximum Likelihood Error (MLE) given by eq (6). This feature is correlated with the nature of the acoustic sources. More precisely, we expect the MLE to be large for diffused noise, but low for "point" sources. Actually, the SRP shows that the diffused sources are characterized by flat peaks, whereas the point sources map to sharp peaks. This property is mainly due to the nature of the GCC peaks, representing the same source but in different microphone pairs. For a diffused noise, the peaks are generally flat, and might map to different peaks in the location space. As a result, the variance of the estimate is expected to be large and the MLE tends to increase, and vice versa (Figure 1-a).

We propose to use the approach presented in Section 2.2 to estimate the noise and source likelihoods, with the exception of using different distributions. Formally, let $\{(\mathbf{s}_i, err_i)\}_{i=1}^{N_T}$ denote the set of N_T location estimates $\mathbf{s}_i$ and their corresponding MLE values err_i, obtained in T frames. We propose to use a 3-components mixture distribution :

$$f^{mle}(\mathbf{s}) = w_s^{mle} \cdot \Gamma_s^{mle}(err) + w_n^{mle} \cdot f_n^{mle}(err) \quad (11)$$

where $\Gamma_s^{mle}(.)$ is a Gamma distribution approximating the likelihood distribution of the source MLE, whereas $f_n^{mle}(.)$ is a



(a) Estimates variance vs power (b) Mixture fit of the CSRP feature (c) Mixture fit of the MLE feature (d) Classification using NBC

Figure 1: *The figure in (a) illustrates the high correlation between the variance of the SRP peaks generating the estimates and the cumulative SRP, this figure shows clearly that the estimates map to two distinct classes. The graph in (b) and (c) show an example of the maximum likelihood mixture distributions approximating the CSRP distribution and the MLE distribution, respectively. The graph in (d) shows an example of a classification boundary obtained with the NBC.*

"Gaussian + uniform" mixture distribution approximating the likelihood distribution of the noise. $f_n^{mle}(.)$ is given by

$$f_n^{mle}(err) \propto \mathcal{G}_n^{mle}(err) + \mathcal{U}_n^{mle}(err) \tag{12}$$

Similarly to eq. (10), $\mathcal{U}_n^{mle}$ is introduced to model the high MLE values, which are poorly modeled by $\mathcal{G}_n^{mle}$. The CSRP and the MLE features are combined in Section 3 to improve the detection performance.

2.4. Online Parameter Estimation

Acoustic source localization applications, such as camera steering and audio-visual tracking, often require an online localization performance. Therefore, the source/noise classification should be also performed online. Algorithm 1 proposes an approach that accomplishes an online estimation of the distributions parameters from Section 2.2 and 2.3. The proposed algorithm takes into account any possible changes in the distance, number of speakers and noise conditions, which might affect the detection performance. Therefore, only the last N frames are used to re-estimate the parameters. It is worth mentioning that N should not be too small as well.

Algorithm 1 : Online Parameter Estimation

1. Initialize the distributions parameters randomly
2. Let T be the re-estimation period.

for each time t multiple of T **do**

3. Set the initial parameters to the current parameters.
4. Keep the estimates from the last N frames.
5. Re-estimate the parameters using the EM algorithm.

end for

3. Unsupervised Bayesian Classifier

3.1. Naive Bayesian Classifier

The detection task can be improved by fitting a mixture distribution to the joint 2-D feature space, formed by the CSRP and the MLE features. Such an approach is beneficial, because it incorporates the correlation between the two features, which would lead to a more realstic model. The distribution of the 2-D data however narrows the possible choices of the mixture distribution (Figure 1-d), which can efficiently maximize the likelihood, and thereby accurately models the estimates. This problem can be solved by maximizing the likelihood of the data in each feature space (Section 2.2 and 2.3), and then combining the resulting distributions using a Naive Bayesian Classifier (NBC) [17]. Formally, let $\{X_i = (\mathbf{s}_i, c_i, err_i)\}_{i=1}^{N_T}$ be the set of augmented estimates, and let α be the classifier decision, $\alpha \in \{\text{source,noise}\}$. The posterior probability of the decision α given an estimate $X = (\mathbf{s}, c, err)$ is given by

$$p(\alpha|X) = \frac{p(X|\alpha) \cdot p(\alpha)}{p(X)} \tag{13}$$

The NBC assumes the independence of the features [17], and expresses the likelihood distribution according to

$$p(X|\alpha) = \prod_{k=1}^{2} p(X_k|\alpha) = p(c|\alpha) \times p(err|\alpha) \tag{14}$$

Replacing the terms in (13) and (14) by their expressions from (9), (10), (11) and (12) leads to the following unsupervised classifier

$$p(source|X) \propto f_s^{csrp}(c) \cdot \Gamma_s^{mle}(err) \cdot w_s^{csrp} \cdot \cdot w_s^{err} \tag{15}$$

$$p(noise|X) \propto \mathcal{G}_n^{csrp}(c) \cdot f_n^{mle}(err) \cdot w_n^{csrp} \cdot w_n^{mle} \tag{16}$$

The decision α is independent of the probability of the estimate X. Therefore, $p(X)$ is ignored in eq. (15) and (16). X is considered to be generated by an actual source if $p(source|X) \geq p(noise|X)$.

3.2. Loss Function For Noise Control

Acoustic source localization approaches are generally combined with a large number of applications, some of which may require a reduced noise rate, such as beamforming techniques [18], whereas other applications, such as the audio-visual tracking approaches [7, 19], are more robust against noise, and expects a high frequency of correct estimates, even if that leads to an increasing noise rate. The variety of these approaches require more flexibility in the acoustic source classification. This idea is successfully implemented using the loss function [17, 20]. Formally, let $\lambda(\alpha|g)$, be the loss incurred for deciding α knowing that g is the true class, with $\alpha, g \in \{\text{source,noise}\} = \{\mathcal{S}, \mathcal{N}\}$. The risk associated with taking the decision α given the estimate X is calculated according to

$$\mathcal{R}(\alpha|X) = \lambda(\alpha|\mathcal{S}) \cdot p(\mathcal{S}|X) + \lambda(\alpha|\mathcal{N}) \cdot p(\mathcal{N}|X) \tag{17}$$

The classification according to the minimum-risk decision rule is obtained by deciding $\mathcal{S}$ when $\mathcal{R}(\mathcal{S}|X) \leq \mathcal{R}(\mathcal{N}|X)$ and vise versa. This rule is equivalent to

Table 1 : Source/Noise Classification Results

Sequences	SVM + CSRP			SVM + MLE			SVM + CSRP + MLE			NBC + CSRP + MLE		
	R	P	F	R	P	F	R	P	F	R	P	F
seq18-2p-0101	0.46	**0.85**	0.60	**0.94**	0.33	0.49	0.83	0.67	**0.74**	0.86	0.62	0.72
seq24-2p-0111	0.42	**0.80**	0.55	**0.94**	0.27	0.45	0.83	0.56	0.66	0.82	0.60	**0.69**
seq40-3p-0111	0.26	**0.92**	0.41	**0.81**	0.56	0.67	0.58	0.82	0.68	0.63	0.79	**0.70**
seq45-3p-1111	0.30	**0.55**	0.40	**0.89**	0.26	0.40	0.70	0.42	0.52	0.71	0.42	**0.53**
seq37-3p-0001	0.10	**0.91**	0.17	**0.77**	0.49	0.60	0.72	0.61	**0.66**	0.76	0.57	**0.66**

$$\frac{\lambda(\mathcal{N}|\mathcal{S}) - \lambda(\mathcal{S}|\mathcal{S})}{\lambda(\mathcal{S}|\mathcal{N}) - \lambda(\mathcal{N}|\mathcal{N})} \geq \frac{p(\mathcal{N}|X)}{p(\mathcal{S}|X)} \tag{18}$$

$\lambda(\mathcal{S}|\mathcal{S})$ and $\lambda(\mathcal{N}|\mathcal{N})$ represent the loss incurred for making the right decision. Therefore, these two parameters are generally set to 0. On the other hand, setting $\lambda(\mathcal{N}|\mathcal{S}) = \lambda(\mathcal{S}|\mathcal{N}) = 1$ leads to the NBC (Section 3.1). The noise rate can be then controlled by adapting the ratio of these two parameters.

4. Experiments and Results

We evaluate the proposed approach using the AV16.3 corpus [21], where human speakers have been recorded in a smart meeting room (approximately 30m^2 in size) with a 20cm 8-channel circular microphone array. The sampling rate is 16 kHz and the real mouth position is known with an error $\leq$ 5cm [21]. The AV16.3 corpus has a variety of scenarios, such as stationary or quickly moving speakers and varying number of simultaneous speakers. The source localization experimental

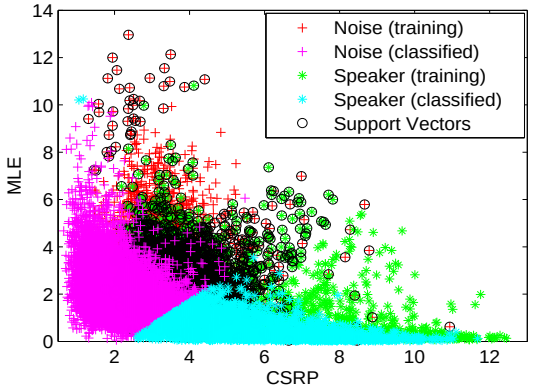


Figure 2: *Example of classification with SVM.*

setup used in these experiments is the similar to that proposed in [15]. More precisely, the signal was divided into frames of 512 samples (32ms); the GCCs were calculated using PHAT [22] weighting; and a voice activity detector was used in order to suppress silence frames. The multiple speaker localization approach provides 6 estimates per frame (N_{max} in [14, 15]), whereas the number of simultaneous speakers varies between 0 and 3. The proposed approach is compared to the classical Support Vector Machine (SVM) [17, 20] approach with quadratic kernel (Figure 2). Using different kernels did not improve the results. The SVM training data is obtained by calculating the MLE and CSRP features for all locations given by the sequences ground truth, as well as for noise locations selected randomly. The reported results were obtained with a training on the audio sequence *seq02-1p-0000*, and then testing on the remaining multiple speaker sequences from the corpus. The results are reported in terms of the Recall (R), Precision (P) and F-measure (F). These measures are given by

$$P = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{19}$$

$$R = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{20}$$

$$F = 2 \cdot \frac{R \cdot P}{R + P} \tag{21}$$

The values of these measures are between 0 and 1. The higher they are, the better the classification is. The recall represents the fraction of actual source(s) estimates that is correctly classified, whereas the precision reports the fraction of estimates which are correctly classified. Finally, the F-measure is a weighted harmonic mean of precision and recall. This measure is very relevant in assessing the overall performance.

Table 1 presents the results of the multiple source detection task using the SVM approach, when it is combined with each feature separately, as well as when the features are jointly used. These results show that combining the MLE and CSRP features leads to better classification results. More precisely, we can see that using the MLE feature alone, leads to a good recall performance but very poor precision. On the other hand, using the CSRP feature alone results in a good precision performance but a poor recall. Combining these two features, however provides more information to the SVM classifier, which successfully increases the F-measure of all sequences. We can also see that, contrary to the "MLE only" and "CSRP only" results, the recall and precision performance of the joint features experiments are balanced. We can conclude from these results that combining the MLE and CSRP features increases the detection performance.

Table 1 also reports the results of the proposed unsupervised Bayesian classifier. These results show clearly that the proposed classifier performs slightly better than SVM. This is mainly due to the dependency of the features on the source location and the number of speakers. These two factors highly affect the signal power level and the SNR. Therefore, using a single training data to classify the different scenarios proposed by the AV16.3 corpus leads to a sub-optimal performance. The proposed classifier however adapts easily to these changes. This is due to the self-learning approach, which uses the data itself to infer the best boundary that explains the two classes.

5. Conclusion

We have proposed an unsupervised Bayesian classifier to the multiple speaker detection task. The proposed approach uses the maximum likelihood error and the cumulative SRP as classification features, and uses a naive Bayesian technique to combine their distributions. This approach also provides a flexible framework to control the noise rate, and can be easily extended to integrate more speaker/speech features.

6. References

[1] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 11, no. 6, pp. 549–557, 2003.

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

[3] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1327 –1339, May 2007.

[4] ——, "The generalization of narrowband localization methods to broadband environments via parametrization of the spatial correlation matrix," in *Proc. EUSIPCO*, Sep. 2007, pp. 763–767.

[5] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661 – 1669, Dec. 1987.

[6] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 7, no. 1, pp. 45–50, Jan. 1997.

[7] Y. Oualil, F. Faubel, , and D. Klakow, "A multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking," in *13th International Workshop on Acoustic Signal Enhancement*, Sep. 2012, pp. 233–236.

[8] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[9] J. P. Dmochowski, J. Benesty, and S. Affes, "Fast steered response power source localization using inverse mapping of relative delays," in *Proc. ICASSP*, 2008, pp. 289–292.

[10] M. Nilesh and M. Rainer, "A scalable framework for multiple speaker localization and tracking," in *Proc. IWAENC*, 2008.

[11] H. Do and H. Silverman, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. ICASSP*, Apr. 2008, pp. 301 –304.

[12] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. ICASSP*, 2010, pp. 125–128.

[13] G. Lathoud, M. Magimai.-Doss, , and B. Hervé, "Threshold Selection for Unsupervised Detection, with an application to Microphone arrays," in *Proc. ICASSP*, Toulouse, France, May 2006.

[14] Y. Oualil, M. Magimai.-Doss, F. Faubel, and D. Klakow, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Statistical and Perceptual Audition Workshop*, Sep. 2012.

[15] ——, "A probabilistic framework for multiple speaker localization," in *Proc. ICASSP*, May 2013.

[16] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*, 2nd ed. Wiley-Interscience, Mar. 2008.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, 2nd ed. Wiley-Interscience, Nov. 2000.

[18] H. L. Van Trees, *Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)*, 1st ed. Wiley-Interscience, Mar. 2002.

[19] A. Levy, S. Gannot, and A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Acoust., Speech, Signal Process.*, 2010.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, Oct. 2007.

[21] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI 04 Workshop*, May 2006, pp. 182–195.

[22] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.