ONLINE UNSUPERVISED OVERLAPPING SPEAKER DETECTION USING ENHANCED CLASSIFICATION HISTORY-BASED FEATURES

Youssef Oualil, Rahil Mahdian Toroghi, Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany

youssef.oualil@lsv.uni-saarland.de

ABSTRACT

Overlapping speaker localization approaches 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 firstly introduce an online implementation of a previously developed offline multiple speaker detector. This classifier is then extended to include new detection features. More precisely, the proposed approach uses the classified location estimates as labelled data to train new classification models for different potential features. The resulting models are then integrated into the online classifier to improve the classification performance. In particular, this paper investigates three different classification history-based models, namely, the location, the kurtosis and the probabilistic steered response power features. Experiments conducted on the AV16.3 corpus show the effectiveness of the proposed approach.

Index Terms— Multiple speaker detection, unsupervised Bayesian classifier, steered response power.

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].

A good overlapping 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, Nilesh et *al.* [9] proposed to use the distance separating the estimates as a criterion to extract the number and location of the sources, whereas Do et *al.* [10, 11] 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.* [12] proposed an unsupervised threshold selection technique to control the false alarm rate.

We have recently proposed an offline unsupervised classifier [13], which estimates the optimal Maximum Likelihood (ML) boundary between the noise and speaker classes directly from the data. This approach uses the Cumulative Steered Response Power (CSRP) and the ML Error (MLE) introduced at each location estimate as classification features, and trains two different 3-component mixture distributions that separate the noise from speakers estimates in each feature space separately. The resulting distributions are then combined using a Naive Bayesian Classifier (NBC). We have also introduced a theoretical and brief view of how an online implementation of the proposed approach can be achieved.

The proposed classifier, however, is a memoryless detector which performs the detection on a frame level using only two features. This paper follows the line of thoughts in [13] by 1) investigating the online counterpart of the proposed offline classifier in a first step, 2) and then extending the feature space by training new classification models based on the recent classification history. More particularly, we will investigate new history-based classification models for the location, kurtosis and the probabilistic SRP features. These models are directly integrated into the online classifier to improve the detection.

We proceed in this paper by briefly reviewing the offline classifier proposed in [13]. Then, we introduce the online implementation of this detector in Section 3. Section 4 presents the new history-based classification features, whereas Section 5 shows the performance of the proposed approach including the new features in comparison with the offline classifier. Finally, we conclude in Section 6.

2. OFFLINE NAIVE BAYESIAN CLASSIFIER

In this section, we review the offline NBC proposed in [13]. More precisely, we will briefly introduce the multiple speaker localization approach, which is used to estimate the potential speaker location and to calculate the classification features. Then, we will review the classification models, followed by a short introduction to the NBC which combines the likelihood distributions of all features.

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 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 Section 2.1

presents a brief introduction to the approach proposed in [14], which is used in this work as a detector.

Formally, let M and Q denote the number of microphones and corresponding pairs, respectively, and let $\mathbf{m}_h, h = 1, \ldots, M$, denote the positions of the microphones. 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:

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

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, respectively. The probabilistic SRP (PSRP) of a given location $\mathbf{s}$ is calculated according to [14]:

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

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 $\mathbf{s}_e$ is dominant. Then, 2) calculating the restriction of (3) on the space region $\mathcal{S}_e$ where $\mathbf{s}_e$ is dominant. Finally, 3) the optimal location estimate is obtained via numerical optimization (see [14, 15] for more details).

2.2. Cumulative Steered Response Power Feature

The first detection feature considered in [13] is the Cumulative SRP (CSRP). This feature does not simply consider the power coming from a single location, it rather considers the cumulative power emerging from the region of dominance associated to the location estimate. Formally, the cumulative SRP $C_{\mathbf{s}_e}$ introduced at the location estimate $\mathbf{s}_e$ is calculated according to:

$$C_{\mathbf{s}_e} = \int_{\mathcal{S}_e} \text{PSRP}(\mathbf{s}) \cdot \text{d}\mathbf{s} \approx \sum_{q=1}^{Q} w_{\mathbf{s}_e}^q \quad (4)$$

$\mathcal{S}_e$ represents the space region where the acoustic event that generated $\mathbf{s}_e$ is dominant. The equation (4) is obtained by mapping $\mathcal{S}_e$ 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. The CSRP classification model is obtained by training a 3-component mixture distribution on 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 algorithm [16]. Formally, the EM algorithm estimates a mixture distribution of the form

$$f^{csrp}(\mathbf{s}) = w_n^{csrp} \cdot \mathcal{G}_n^{csrp}(c) + w_s^{csrp} f_s^{csrp}(c) \quad (5)$$

where $\mathcal{G}_n^{csrp}(.)$ is a Gaussian distribution approximating the likelihood distributions of the noise, whereas $f^{csrp}(.)$ is a "Gaussian+Uniform" distribution approximating the likelihood of the speakers. w_n^{csrp} and w_s^{csrp} denote the noise and source priors, respectively (see example in Fig. 1). The reader is referred to [13] for more details.

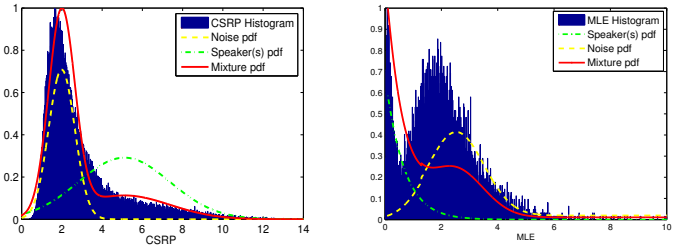

Fig. 1: *Example of the ML mixture distributions approximating the CSRP and the MLE distributions, respectively.*

2.3. Maximum Likelihood Error Feature

The second classification feature is the Maximum Likelihood Error (MLE) given by:

$$err(\mathbf{s}_e) = \sum_{q=1}^{Q} \left(\frac{\tau^q(\mathbf{s}_e) - \mu_{\mathbf{s}_e}^q}{\sigma_{\mathbf{s}_e}^q} \right)^2 \quad (6)$$

This feature is correlated with the nature of the acoustic sources. More precisely, the MLE is expected to be large for diffuse noise, but low for "point" sources (see [13] for more details).

The noise and source likelihood distributions are estimated using the same approach presented in Section 2.2, 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. The likelihood is approximated by a 3-component mixture distribution:

$$f^{mle}(\mathbf{s}) = w_s^{mle} \cdot \Gamma_s^{mle}(err) + w_n^{mle} \cdot f_n^{mle}(err) \quad (7)$$

where $\Gamma_s^{mle}(.)$ is a Gamma distribution approximating the likelihood distribution of the source MLE, whereas $f_n^{mle}(.)$ is a "Gaussian+Uniform" distribution approximating the likelihood of the noise (see [13] for more details).

2.4. Naive Bayesian Classifier

The Naive Bayesian Classifier (NBC) is an alternative solution to classification problems where a good estimation of the likelihood distribution in the joint feature space is difficult to obtain. In this case, a 1-dimension distribution can be estimated in each feature space, followed by their combination, under independence assumption, using the NBC (see [13, 17] for more details). Formally, if α is the classifier decision, $\alpha \in \{\text{source,noise}\}$. The NBC calculates the likelihood of the estimate $X = (\mathbf{s}, c, err)$ given the decision α according to:

$$p(X|\alpha) = \prod_{k=1}^{2} p(X_k|\alpha) = p(c|\alpha) \times p(err|\alpha) \quad (8)$$

Replacing the terms in (8) by their expressions in (5) and (7) and using Bayes' rule leads to the following posterior distributions:

$$p(source|X) \propto f_s^{csrp}(c) \cdot \Gamma_s^{mle}(err) \cdot w_s^{csrp} \cdot w_s^{mle} \quad (9)$$

$$p(noise|X) \propto \mathcal{G}_n^{csrp}(c) \cdot f_n^{mle}(err) \cdot w_n^{csrp} \cdot w_n^{mle} \quad (10)$$

X is considered to be generated by an actual source if $p(source|X) \geq p(noise|X)$.

The extension of this NBC to include more features is straightforward. The likelihood distribution of each new feature will be used to augment the likelihood product (8), whereas the prior distribution of each class is given by the product of the corresponding priors calculated in different feature spaces.

3. ONLINE NAIVE BAYESIAN CLASSIFIER

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 distribution parameters from Section 2.2 and 2.3. The proposed algorithm takes into account any

Algorithm 1 : Online Parameter Estimation

1. Initialize the distributions parameters using K-means
2. Let T be the re-estimation period.

for t multiple of T **do**

3. Set the initial parameters to the current parameters.
4. Use the last N estimates as training set.
5. Re-estimate the parameters using the EM algorithm.

end for

possible changes in the distance, number of speakers and noise conditions, which might affect the decision boundary. Therefore, only the last N feature estimates are used to re-estimate the parameters.

4. CLASSIFICATION HISTORY-BASED FEATURES

This section shows how the classification history can be used to augment the feature space and thereby improve the detection performance. More particularly, the classified location estimates and their corresponding PSRP and kurtosis values are investigated as three potential new features. The idea here is to use the classified estimates as labeled data to train separately new speaker and noise models for different features. These models are then incorporated into the NBC.

4.1. Location Feature

The location estimates are widely used as a main feature in speaker clustering and classification approaches. This is mainly due to the high density of the estimates originated from the same speaker in the location space, whereas the noise estimates are assumed to be randomly distributed. The main problem, however, is to identify which clusters of estimates represent actual speakers. This is mainly solved using speech cues or cluster variance-based discrimination [18]. We propose to overcome this identification problem here by training separately speaker(s) and noise models using the late classification history. Each model has the form of a GM with a number of components given by the minimum Bayesian Information Criterion (BIC). Algorithm 2 shows the proposed online training approach.

Algorithm 2 : Training of the location-based classification models

1. Let T_{loc} be the re-estimation period.

for t multiple of T_{loc} **do**

2. Use the last N_{loc} speaker/noise estimates as two separate training sets.

for $k = 1 \dots K$ **do**

3. train a speaker GM model $\mathcal{M}^{k}_{s,loc}$ (k components)
4. train a noise GM model $\mathcal{M}^{k}_{n,loc}$ (k components)

end for

5. return the speaker and noise models $\mathcal{M}^{k_s}_{s,loc}$ and $\mathcal{M}^{k_n}_{n,loc}$: $k_s = \text{argmin}_k \text{ BIC}(\mathcal{M}^{k}_{s,loc})$, $k_n = \text{argmin}_k \text{ BIC}(\mathcal{M}^{k}_{n,loc})$.

end for

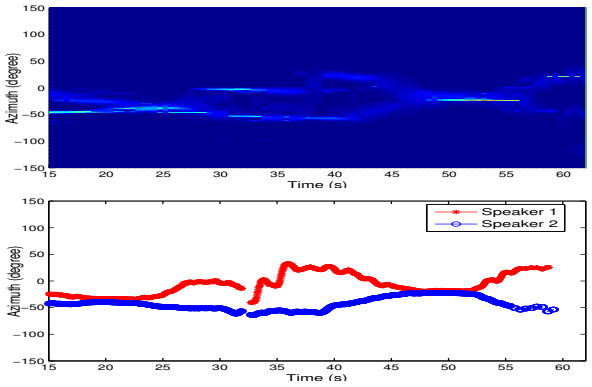

Fig. 2: Illustration of a history-based classification model: top figure illustrates the time evolution of the GM speaker model for the location feature. The figure in the bottom shows the location ground truth of the corresponding 2 speakers.

4.2. Kurtosis Feature

High order statistics of signals have been widely investigated to solve speech and signal processing problems. More particularly, the kurtosis of a signal is typically used in Blind Source Separation (BSS) to separate speech sources [19, 20]. The kurtosis of a zero mean random variable x is calculated according to:

$$kurt(x) = \frac{\mathcal{E}\{x^4\}}{\{\mathcal{E}\{x^2\}\}^2} \tag{11}$$

where $\mathcal{E}\{.\}$ is the expectation operator.
Speech signals are generally assumed to follow super-Gaussian distributions, whereas noise signals are mostly modeled as Gaussians or mixture of Gaussians. Therefore, speech and noise signals are expected to have different kurtosis. Based on this difference, we propose to use the kurtosis as a new detection feature. More precisely, the signal coming for each location estimate is calculated using a superdirective Beamformer (BF), followed by the calculation of the kurtosis according to (11) as shown in Fig. 3. The likelihood of each kurtosis model is approximated by a Gaussian distribution.

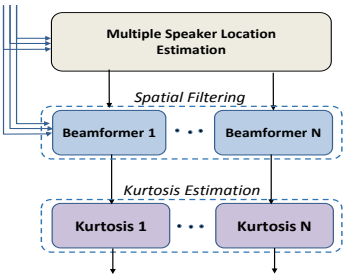

Fig. 3: Pipeline for calculating the kurtosis at each location estimate.

4.3. Probabilistic SRP Feature

The probabilistic SRP feature PSRP($\mathbf{s}$) of a location estimate $\mathbf{s}$ (calculated according to (3)) is a probabilistic interpretation of the signal power at that particular location. We have shown in [13] that the SRP is highly correlated with the variance of the location estimate. More precisely, sharp SRP peaks representing point speech sources (mouth) are more likely to have a small variance, and therefore small MLE, contrary to noise sources, which are expected to span over wider space regions resulting in a larger variance and flat SRP peaks.

Table 1: Speaker/Noise Classification Results

Sequences	seq18-2p-0101			seq24-2p-0111			seq40-3p-0111			seq45-3p-1111			seq37-3p-0001		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
Offline MLE+CSRP	0.81	**0.72**	0.76	0.75	**0.71**	**0.73**	0.56	0.85	0.67	0.65	0.48	0.55	0.70	0.65	0.67
Online MLE+CSRP	0.84	0.68	0.75	0.76	0.68	0.72	0.63	0.81	0.70	0.68	0.49	0.57	0.76	**0.71**	0.73
+PSRP (only)	0.84	**0.72**	**0.78**	0.77	0.70	**0.73**	0.60	0.86	0.71	0.65	**0.51**	**0.58**	0.75	0.67	0.70
+LOC (only)	**0.88**	0.67	0.76	**0.80**	0.57	0.67	**0.65**	0.86	**0.74**	**0.73**	0.41	0.53	**0.80**	0.67	0.73
+KURT (only)	0.83	0.65	0.73	0.80	0.60	0.68	0.60	0.81	0.69	0.70	0.44	0.54	0.73	0.59	0.65
+PSRP+LOC	0.87	0.71	**0.78**	0.78	0.68	**0.73**	0.64	**0.89**	**0.74**	0.68	0.50	**0.58**	0.79	0.70	**0.74**
+PSRP+LOC+KURT	0.87	0.71	**0.78**	0.78	0.66	0.72	**0.65**	0.87	**0.74**	0.68	0.50	**0.58**	0.79	0.70	**0.74**

This correlation is extended here to include the CSRP feature. More particularly, we expect the PSRP feature to help the classifier discriminating between estimates based on the signal power at these locations, whereas the CSRP classifies the estimates based on the cumulative signal power coming from the region surrounding the location. Therefore, we expect the classifier to be able to correctly classify distributed noise source, such as projector, which generally have a high CSRP value but low PSRP. The main goal here is to use this complimentary and redundant information provided by the MLE, CSRP and PSRP features to increase the robustness of the classifier. Similarly to the location and kurtosis features, the speaker and noise PSRP classification models are trained separately using the recently classified speech and noise estimates, with the only exception of using a 3-component mixture similar to the one used to train the CSRP.

5. 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 $\approx$ 1.2cm [21]. The AV16.3 corpus has a variety of scenarios, such as stationary or quickly moving speakers and varying number of simultaneous speakers. The multiple speaker sequences are highly overlapping recordings with speakers talking simultaneously during the complete audio sequences. The source localization experimental setup used in these experiments is similar to that proposed in [15], whereas the multiple speaker detection setting is the same as the one used in the offline classifier [13]. More particularly, the re-estimation period $T = 3s$ for CSRP, MLE, PSRP and kurtosis, whereas $T_{loc} = 1s$. This differences aims at modeling any possible fast changes in the speakers location. Moreover, the number of the classification history-based estimates $N = 1000$, whereas $N_{loc} = 300$. The first $20s$ of each recording were used to initialize the models. The signal was divided into frames of 512 samples (32ms), and the GCCs were calculated using PHAT [22] weighting. The robustness of the proposed approach to noise is evaluated by introducing a high noise rate. More precisely, the multiple speaker localization approach provides 6 estimates per frame (N_{max} in [14, 15]). Given that the number of simultaneous speakers varies between 1 and 3, this leads to a noise rate of $\geq 50\%$ in the best case.

We have shown in [13] that the offline NBC outperforms the classical Support Vector Machine (SVM) [17, 23] when they are applied to the CSRP and MLE features. In this paper, however, we will conduct an experimental comparison between the proposed online and the previously developed offline classifiers when they are applied to the CSRP and MLE features. The results are reported in Table 1. This table also reports the results of the multiple speaker detection experiments when the NBC is augmented with the proposed history-based classification features. The results are reported in terms of the Recall (R), Precision (P) and F-measure (F). These measures are given by:

$$P = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (12)$$

$$R = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (13)$$

$$F = 2 \cdot \frac{R \cdot P}{R + P} \quad (14)$$

The higher these measures are, the better the classification is. The recall represents the fraction of actual speaker estimates that is correctly classified, whereas the precision reports the fraction of estimates which are correctly classified. Finally, the F-measure is the harmonic mean which is used to evaluate the overall performance.

The results reported in Table 1 show that the proposed online classifier combined only with the CSRP and MLE features perform better than its offline counterpart. This improvement is mainly due to the online adaptation of the classification model parameters, which is necessary in the case of changes in the speaker environment. This improvement appears clearly in the sequences 40 and 37, where the number of simultaneous speakers and distance to the array changes over time. We can also conclude that the online classifier adapts quickly to the environment, as it requires only $20s$ as initial duration to estimate the models. We can also see that adding more features increases the robustness of the online classifier. More precisely, we can see that augmenting the online classifier with the PSRP or the location (LOC) features alone lead to an "unstable" improvement of the performance. This is mainly due to the non-convergence of the classification models in few parameters re-estimation steps or due to long segments of intended silence. This instability of the performance appears also in the unbalanced P and R results. Combining more features, however, provides more information to the classifier, which successfully increases the F-measure of all sequences using the MLE+CSRP+PSRP+LOC features. We can also conclude that adding the kurtosis feature to this mixture did not improve the performance. This is mainly due to the BF step which introduces distortions in the speech signal, leading to non-reliable classification models in many of the re-estimation steps. This confirms a similar conclusion regarding the MFCC features that was reported in [18].

6. CONCLUSION

We have proposed an online unsupervised Bayesian classifier to the multiple speaker detection task. The proposed approach uses an adaptive online learning approach to re-estimate the classification models. This classifier was further extended to include three different classification history-based features. This approach is flexible and can be easily extended to integrate more speech features.

7. 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] J. Dmochowski, J. Benesty, and S. Affes, "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. thesis, Brown University, 2000.

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

[10] H. Do and H.F. 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.

[11] 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.

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

[13] Y. Oualil, F. Faubel, and D. Klakow, "An unsupervised Bayesian classifier for multiple speaker detection and localization," in *Proc. INTERSPEECH*, Aug. 2013.

[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] Y. Oualil, M. Magimai.-Doss, F. Faubel, and D. Klakow, "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)*, Wiley-Interscience, 2 edition, Mar. 2008.

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

[18] G. Lathoud, *Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, Dec. 2006.

[19] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P.N. Garner, and Weifeng Li, "Beamforming with a maximum negentropy criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 994–1008, July 2009.

[20] J.P. LeBlanc and P.L. De Leon, "Speech separation by kurtosis maximization," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, May 1998, vol. 2, pp. 1029–1032 vol.2.

[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.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 1st ed. 2006. corr. 2nd printing 2011 edition, Oct. 2007.