An Online Expectation-Maximization Algorithm for Tracking Acoustic Sources in Multi-Microphone Devices during Music Playback

Daniele Giacobello
Sonos Inc.,
Santa Barbara, CA, USA
daniele.giacobello@sonos.com

***Abstract*—In this paper, we propose an expectation-maximization algorithm to perform online tracking of moving sources around multi-microphone devices. We are particularly targeting the application scenario of distant-talking control of a music playback device. The goal is to perform spatial tracking of the moving sources and to estimate the probability that each of these sources is active. In particular, we use the expectation-maximization algorithm to capture the statistical behavior of the feature space representing the ensemble of sources as a Gaussian mixture model, assigning each Gaussian component to an individual acoustic source. The features used exploit a wide range of information on the sources behavior making the system robust to noise, reverberation, and music playback. We then differentiate between desired and interfering sources. The spatial information and activity level is then determined for each desired source. Experimental evaluation of a real acoustic source tracking problem with and without music playback shows promising results for the proposed approach.**

I. Introduction

Tracking multiple moving sources is an essential component of modern multi-microphone speech enhancement systems [1]. Examples on how to use the tracking information can be seen in, e.g., adaptive beamforming [2] or post-filtering [3]. These solutions are generally based on classical spatial filters, computed as a closed-form or adaptive solutions of a specific optimization problem. These generally make the filter unable to adapt fast enough to ever changing acoustic scenarios [4]. Alternatively, parametric spatial filters rely on instantaneous estimates of model parameters, notably the direction of arrival (DOA) or, equivalently, time difference of arrival (TDOA), e.g., [5], [6]. These allow the system to adapt quickly to the changing scenario but violate the underlying signal model when multiple sources are active at the same time [7], [8]. It is, however, possible to localize and classify sources and move past the rigid assumptions of these two paradigms. This is what is usually referred to as *informed* spatial filtering [9], [10], [11].

Spatial localization identifies prominent sources in the current acoustic scene based on the set of features chosen, e.g., TDOA. In [10] and [12], Gaussian mixture models (GMMs), parametrized in a different feature set from the one used in this work, are used. In [13], by modeling the generalized cross-correlation (GCC) [14] with a GMM, they are able to track multiple sources avoiding possible multiple maxima. The spatial tracking error of the TDOA is generally assumed to be Gaussian. In this case, it is easy to see that the least-squares metric provides the maximum likelihood (ML) estimate of the speaker location [15]. However, methods moving away from this strict assumption exist, notably [16] and [17]. In these papers, the authors track sources using *particle filters* with likelihood models derived from both the GCC function and the time-delay estimate (TDE), respectively.

After spatial localization and tracking, classification is required to differentiate desired from interfering sources such as reverberation, background noise, interfering talkers, or music played back by the device. In [18], a ML approach was derived based on a probabilistic interpretation of the GCC function. This was scored against models representing speech and noise. This feature is then combined with the Steered Response Power (SRP), creating a naive Bayes classifier to identify active and interfering talkers.

In this paper, we propose a statistical framework to perform online source tracking in a multi-microphone *smart* loudspeaker device, e.g., for distant-talking control [19], [20]. The goal of the system is to track desired and interfering sources and to estimate the probability that each source is active. We define desired sources (DSs) as those which should eventually remain untouched by the spatial noise suppression system, therefore including the music playback feeding back in the microphones.

We use a GMM to capture the statistical behavior of an ensemble of sources where each source is seen as a multivariate Gaussian component. We will describe the online estimation of model parameters, which allows our statistical model to track non stationary processes. We will also discuss the design of feature vectors which promote separation between distributions. The use of a mixture model provides a low-complexity method of tracking the spatial location of acoustic sources, while also maintaining estimates of their online statistical behavior. We then evaluate the source membership using the GMM model obtained. After source classification, TDOAs of each DS can be extracted from the statistical model, along with the probability that each source is active.

This paper is organized as follows. Section II presents the first stage of the source tracking system, deriving the statistical modeling of acoustic sources. Section III presents the second

ISBN 978-90-827970-1-5 © EURASIP 2018

stage of the algorithm, where inference of the DS behavior is discussed. Experimental results are provided in Section IV, and conclusions are given in Section V.

II. Statistical Modeling of Acoustic Sources

A. Feature Vector Design

The first stage of our algorithm consist in extracting meaningful features from the acoustic scenario to model and track the DSs using the statistical framework provided by the GMMs.

- *Time Difference of Arrival (TDOA)*
  Source detection and tracking relies heavily on spatial information, which can be represented as the DOA or as the TDOA [1]. Numerous methods based on spectral cross-correlation exist for estimating the TDOA, e.g., the GCC function [14]. Such correlation-based approaches generally involve finding the TDOA which maximizes a cost function designed to capture similarity between signals observed at different microphones [18], [21].
- *Correlation Measures*
  The correlation-based cost functions discussed above can be leveraged for source tracking in an alternative way by measuring the maximum cost corresponding to the selected TDOA. Such measures can be expected to show small values for diffuse sources, but large values for point sources which are more consistent with DSs.
- *Predictors of Speech Activity*
  These can be used to discriminate between speech and non-speech acoustic sources. For example, voice activity detectors (VADs) can provide measures which convey the likeness of a particular acoustic signal to speech. Furthermore, certain discriminative features used in VADs, i.e., pitch information [22], can be leveraged for speaker identification [23].
- *Predictors of Loudspeaker Activity*
  The acoustic activity of the loudspeaker, i.e., if the loudspeaker playback is active, can be estimated from the residual of the acoustic echo cancellation, a necessary step to improve the signal at the microphone. The coherence between the music playback and the output of the AEC (residual echo) measures this activity [24].

B. Statistical Framework

The proposed statistical modeling assumes that a feature vector $\mathbf{x}_n$ is extracted from the observed acoustic signals in frame n. The vector $\mathbf{x}_n$ is designed to capture important characteristics about the current acoustic scene and its design follows the discussion in Section II-A. Here, it is assumed to be a generalized frame-specific sample so that statistical modeling is applied on the frame level.

A GMM is fully parametrized by the set:

$$\Lambda = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M, w_1, \ldots, w_M\}, \quad (1)$$

where $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$, and w_m are the mixture means, covariance matrices, and priors, respectively, and M is the number of mixtures. The GMM likelihood conditioned on Λ is expressed as [25]:

$$p(\mathbf{x}_n | \Lambda) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m). \quad (2)$$

where $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the evaluation of a single Gaussian distribution. In order to apply a GMM in statistical analysis, parameter estimation of Λ must be performed based on training data. Due to the uncertainty associated with mixture membership of training samples, the parameter estimation is typically performed using the EM algorithm [26].

C. Online/Recursive Parameter Estimation

Similarly to what proposed in other online spatial tracking methods, e.g., [10], [12], [27], we base the recursive estimation of the GMM parameters on the maximum a posteriori (MAP) criterion proposed in [28]. Let $\Lambda^{(N)}$ denote the set of parameters estimated from $\mathcal{X}_{[1,N]} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathcal{X}_{[N+1,N+K]} = \{\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+K}\}$ a new set of feature vectors measured in the acoustic space. Note that $K = 1$ corresponds to applying parameter estimation for every new feature vector. A MAP estimate of $\Lambda^{(N+K)}$ can be obtained recursively based on the data in $\mathcal{X}_{[N+1,N+K]}$. Assuming to know $p\left(\Lambda^{(N)} \middle| \mathcal{X}_{[1,N]}\right)$, the distribution of the parameters at instant N:

$$\begin{aligned}
\boldsymbol{\mu}_m^{(N+K)} =& a_m \mathbf{E}_{m,1}/E_{m,0} + (1 - a_m)\, \boldsymbol{\mu}_m^{(N)}, \\
\boldsymbol{\Sigma}_m^{(N+K)} =& a_m \mathbf{E}_{m,2}/E_{m,0} \\
&+ (1 - a_m) \left(\boldsymbol{\Sigma}_m^{(N)} + \boldsymbol{\mu}_m^{(N)} \boldsymbol{\mu}_m^{(N),T}\right) \\
&- \boldsymbol{\mu}_m^{(N+K)} \boldsymbol{\mu}_m^{(N+K),T}, \\
w_m^{(N+K)} =& a_m E_{m,0}/K + (1 - a_m)\, w_m^{(N)},
\end{aligned} \quad (3)$$

where sufficient statistics of $\mathcal{X}_{[N+1,N+K]}$ are given by:

$$\begin{aligned}
E_{m,0} =& \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right), \\
\mathbf{E}_{m,1} =& \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right) \mathbf{x}_{N+k}, \\
\mathbf{E}_{m,2} =& \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right) \mathbf{x}_{N+k} \mathbf{x}_{N+k}^T,
\end{aligned} \quad (4)$$

where $P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right)$ is the posterior probability of mixture m, and the adaptation step size is given by

$$a_m = E_{m,0} / \left(E_{m,0} + N w_m^{(N)}\right). \quad (5)$$

To reduce long-term data dependencies, we can simply applying a maximum limit to the term $N w_m^{(N)}$.

D. Merging and Splitting of Gaussian Components

Modeling non stationary processes using recursive estimation may require online merging or splitting of the GMM components. In the context of source tracking, recursive estimation may lead multiple components to represent a single

ISBN 978-90-827970-1-5 © EURASIP 2018

acoustic source. Thus, merging is applied to those mixtures which are statistically similar [29], [10]. Recursive estimation may also produce mixtures with low priors or highly peaked covariances, which effectively become marginalized during statistical analysis. Thus, the use of thresholds to delete mixtures which do not meet minimum prior values or covariance spreads is an efficient way to guarantee proper statistical modeling without adding complexity or storage burden [30], [31]. In the case where one component might be modeling multiple sources, generally if the prior, mean, and variance are above a certain threshold, the Gaussian distribution is split into two using the method described in [29].

III. Inference of Desired Source Behavior

The GMM obtained through the method presented in Section II implicitly embeds the spatial location of the sources, L_{DS}, and the posterior probability that the sources are active, P_{DS}. In order to track the DSs based on the GMM modeling of the acoustic space, the corresponding mixtures must be identified. The proposed solution allows for multiple DSs. For notational brevity, we discuss the system in the context of a single DS. We propose to use the minimum Mahalanobis distance [32] between a point $\mathbf{z}_n$ in the parameter space and the Gaussian mixtures representing the space at instant n:

$$\widehat{m}_{\text{DS}} = \arg\min_{m} \sqrt{(\mathbf{z}_n - \boldsymbol{\mu}_m)\, \boldsymbol{\Sigma}_m\, (\mathbf{z}_n - \boldsymbol{\mu}_m)}. \qquad (6)$$

Once $\widehat{m}_{\text{DS}}$, the distribution with smaller distance to $\mathbf{z}_n$, has been identified, the DS can be tracked and its level of activity can be inferred. The spatial location of the DS, L_{DS}, is determined as the TDOA element of the GMM mixture mean for $\widehat{m}_{\text{DS}}$ (then converted to DOA). The probability of DS activity is estimated as the posterior probability of the DS mixture conditioned on $\mathbf{z}_n$:

$$P_{\text{DS}} = P\left(\widehat{m}_{\text{DS}} \mid \mathbf{z}_n\right) = \frac{w_m \mathcal{N}\left(\mathbf{z}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)\big|_{m=\widehat{m}_{\text{DS}}}}{\sum_{j=1}^{M} w_j \mathcal{N}\left(\mathbf{z}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}. \qquad (7)$$

A. Extension to Frequency-Based Activity Level

The source tracking system produces time frame DS activity probabilities. If these values are used for controlling other processing, e.g., adaptive beamforming, greater spectral resolution of P_{DS} may be advantageous. Spectral resolution can be introduced by modeling specific frequency-based feature vectors, wherein a single frame n yields multiple features, $\mathbf{x}_{n,f}$, from various frequency bands where f denotes the band index. In this case, the MAP adaptation presented in II-C is applied so that $\Lambda^{(N+1)}$ is updated based on $\Lambda^{(N)}$ and the set $\{\mathbf{x}_{n,1}, \ldots, \mathbf{x}_{n,F}\}$, where K denotes the number of frequency bands.

IV. Experimental Evaluation

We evaluated the proposed algorithm in a rectangular meeting room of size 7.5m×6.2m×2.6m with $\text{RT}_{60} = 0.51$s calculated with [33]. The room contained a centrally located 5m×2m rectangular table where a Sonos™ One smart speaker was positioned on top (of size approx. 16cm×12cm×12cm). We use two microphones on the diameter of the 6-microphone circular array of 72mm diameter laying on top of the speaker.

The signal at the microphones was acquired at 16 kHz and the spectro-temporal representation was obtained by windowing the microphone signals in 512 samples using a Hamming window with 50% overlap with a 512 FFT. When the music is playing, we employed the STFT-domain echo canceller presented in [34] implemented using the robust adaptation proposed in [35] to avoid using double-talk detection. Furthermore, we applied the residual echo suppressor presented in [36] to cope with the possibility of echo leakage. The feature vector, calculated to estimate the posterior probability of source in (7), is a 5-dimension vector. This comprised of the TDOA and associated correlation measure (CDOA) obtained with the algorithm presented in [14], a VAD measure obtained through a combination of spectral entropy and energy [37], the result of the likelihood ratio (LR) tests of the residual echo canceller (RES) [36], and a autocorrelation-based pitch estimate, obtained using [38], done after the RES algorithm. The adaptation term N in (5) was set to give the system a forgiving factor of 250ms (tuned accordingly to the experiment at hand). The parameter space was modeled initially using a GMM with three Gaussian components empirically initiated. For each new frame, we applied the recursive estimation of mean, covariance, and priors presented in Section II-C. To avoid local maxima in fitting the GMM model to our parameter space, we used the Gaussian splitting and merging criterion presented in [29].

In the first experiment, we considered a static talker (DS) located 60cm away. A static interfering talker (IS) was located 120cm away, with an angle of incidence of 90° relative to the DS. To simulate an interaction with the device, the DS was active in 10s intervals, while music (PB) was played back on from the loudspeaker in 5s intervals. The IS was active in 15s intervals. In the primary microphone signal, the DS was observed to be 5dB louder than the IS, the music was picked randomly from a TOP40 playlist and kept at a SPL level of moderate listening, giving approximately a 20dB music-to-DS ratio (roughly 5dB after AEC). Fig. 1 illustrates a 36s segment from the captured primary microphone signal. The highlighted part correspond to acoustic source activity, where green, red, and black correspond to DS, IS, and PB (after AEC and RES), respectively. Fig. 2 provides a snapshot of model parameters from the GMM during the test signal at 10s. The panels show the projections of the multivariate mixture distributions onto individual feature subspaces. Here, mixture 1 (green) is tracking the DS, mixture 2 (red) is tracking the IS, and mixture 3 (black) is capturing diffuse background noise and music playback (PB). Note that individual features, especially the TDOA, were not able to provide clear separation between mixtures during the source classification phase. The higher distance from the microphone of the IS makes the sound diffuse (top pane) but it is easier to discriminate using the CDOA (second pane). Music and speech are easier to classify using the VAD values (consistent with music/speech differences in spectral entropy). Thus, the use of multivariate

ISBN 978-90-827970-1-5 © EURASIP 2018

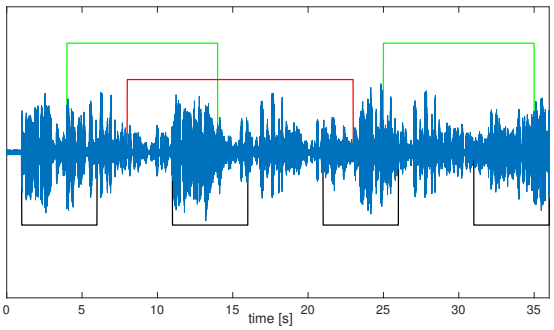

Fig. 1. Input audio signal after AEC. Highlighted is the presence of desired source (green), interfering talker (red), and music playback (black).

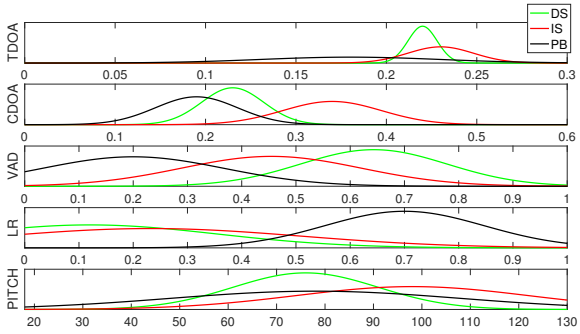

Fig. 2. An illustrative snapshot of the model parameters from the 5-dimension GMM with three mixtures during the signal segment from Fig. 1 at 10s. The panels show the projections of the multivariate Gaussian mixture distributions onto individual feature subspaces (from the top pane: TDOA, CDOA, VAD, LR, PITCH). The distribution modeled are the desired source (green), interfering talker (red), and downlink playback (black).

features introduced complementary information and increases mixture separation. The pitch measure helped particularly in discriminating among the two talkers, while the music mixture resembles a flat distribution. The LR, calculated at the RES, works really well in discriminating PB, while gives roughly same likelihood values for the DS and IS sources.

In the second experiment, a talker was reciting the alphabet while moving slowly through the room (one full rotation of the room in 45s). We then increased the speed of the speaker through the room, performing one full rotation around the table in 25s. We tracked the ground truth angle with a laser digital angle finder and we filmed the process with a camera to manually label the ground truth angle (similar to the AV16.3 dataset [39]). Since we were interested in the source tracking performance during music playback, the experiment was repeated with and without music playing from the device. Again, the music was picked randomly from a TOP40 playlist with approximately a -20dB speaker-to-music ratio at the microphones. We repeated these recordings 5 times with 10 different speakers, for a total of 100 trials. In both experiments, we considered the measurement bias of the ground truth angle to be negligible. We also neglected the front-back TDOA

TABLE I
LOCALIZATION ACCURACY IN TERMS OF ROOT-MEAN-SQUARE ERROR (RMSE) OF THE DOA.

	RMSE DOA [°]			
	Slow Moving		Fast Moving	
	Music Off	Music On	Music Off	Music On
GMM_{2D}	9.3	13.5	8.8	18.2
GMM_{4D}	5.4	6.9	5.8	6.7
GMM_{5D}	4.3	**4.1**	**4.2**	**4.9**
[16]	**4.2**	7.4	5.1	10.2
[17]	5.4	8.7	7.3	11.4

uncertainty of the measurement as not particularly problematic in our controlled scenario. We compared three GMM-based methods for tracking with a different number of features. In the first approach, GMM_{2D}, we used TDOA and CDOA, the most intuitively features to perform tracking. In GMM_{4D}, we added VAD and RES, and we then added the pitch to measure in GMM_{5D}. In Table I, the root-mean-square error (RMSE) of the resulting DOA estimated with the proposed method is shown. We also compared with two other popular nonlinear spatial tracking algorithms [16], [17]. The results of Table I outline the flexibility of the GMM approach and the choice of features. Differently from most spatial tracking methods, like [13] and [12], our feature vector includes terms different from the TDOA that help the estimation inside the EM procedure by providing speech activity information directly into the multivariate GMM, as well as pitch information, which has been known as a low-cost method to perform speaker identification [40]. In GMM_{4D}, we had a clear jump in performance. It is, however, very interesting noting that including the pitch to measure the coupling between loudspeaker and microphone in GMM_{5D} gave a more notable improvement over methods like [16] and [17].

V. CONCLUSION

We proposed a source tracking algorithm based on the GMM modeling of features extracted from the acoustic space. The algorithm, in particular, targets the case of voice-controlled playback device, i.e., a *smart* speaker. Carefully selecting a feature vector that allows for discriminating between residual echo, desired source and interfering sources allows for a flexible amount of separation of the acoustic sources, not possible with just the TDOA, and the overall complexity of the system. The use of voice activity metrics, residual echo likelihood information and pitch information has shown, in particular, to enrich the GMM model in the discrimination allowing for a relatively low RMSE in the DOA estimate during music playback at -20 dB of signal-to-music ratio with a talker moving fairly fast through the room.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, Springer, 2008.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on array processing and sensor networks*, pp. 269–302, 2008.

ISBN 978-90-827970-1-5 © EURASIP 2018

[3] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 684–699, 2003.

[4] O. Thiergart, M. Taseska, and E. A. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2182–2196, 2014.

[5] C. Y.-K. Lai and P. Aarabi, "Multiple-microphone time-varying filters for robust speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 233–236, 2004.

[6] X. Zhang and Y. Jia, "A soft decision based noise cross power spectral density estimation for two-microphone speech enhancement systems," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 813–816, 2005.

[7] O. Thiergart and E. A. Habets, "Sound field model violations in parametric spatial sound processing," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[8] O. Thiergart and E. A. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 659–663, 2013.

[9] D. T. Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2010.

[10] M. Taseska and E. A. P. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.

[11] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, 2017.

[12] M. Nilesh and R. Martin, "A scalable framework for multiple speaker localization and tracking," *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2008.

[13] Y. Oualil, F. Faubel, M. Doss, and D. Klakow, "A TDOA Gaussian Mixture Model for improving acoustic source tracking," *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1339–1343, 2012.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech ,and Signal Processing*, vol. 24(4), pp. 320–327, 1976.

[15] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Advances in Signal Processing*, 2006.

[16] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3021–3024, 2001.

[17] B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1777–1780, 2002.

[18] Y. Oualil, M. Doss, F. Faubel, and D. Klakow, "A probabilistic framework for multiple speaker localization," *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3962–3966, 2013.

[19] R. Pichevar, J. Wung, D. Giacobello, and J. Atkins, "Design and optimization of a speech recognition front-end for distant-talking control of a music playback device," *CoRR*, vol. abs/1405.1379, 2014. Available: http://arxiv.org/abs/1405.1379

[20] D. Giacobello and T. L. Jensen, "Speech dereverberation based on convex optimization algorithms for group sparse linear prediction," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[21] Y. Oualil, F. Faubel, and D. Klakow, "An unsupervised Bayesian classifier for multiple speaker detection and localization," *Proc. Interspeech*, pp. 2943–2947, 2013.

[22] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, 2009.

[23] Y. Shao and D. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. II–205–II–208 vol.2., 2003

[24] G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. II–1893–II–1896.

[25] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[26] R. O. Duda, P. E. Hart, D. G. Stork, *et al.*, *Pattern classification*, Wiley New York, 1973.

[27] Y. Oualil, F. Faubel, and D. Klakow, "A multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking," *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[28] J. Gauvin and C. Lee, "Maximum a Posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[29] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "Split and merge EM algorithm for improving gaussian mixture density estimates," in *Proc. Workshop Neural Networks for Signal Processing*, pp. 274–283, 1998.

[30] V. V. Digalakis, "Online adaptation of hidden markov models using incremental estimation algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 253–261, 1999.

[31] Y. Zhang and M. S. Scordilis, "Effective online unsupervised adaptation of gaussian mixture models and its application to speech classification," *Pattern Recognition Letters*, vol. 29, no. 6, pp. 735–744, 2008.

[32] S. Xiang, F. Nie, and C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

[33] B. Dumortier and E. Vincent, "Blind RT60 estimation robust across room sizes and source distances," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5187–5191, 2014.

[34] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1644, 2006.

[35] T. S. Wada and B.-H. Juang, "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 205–208, 2009.

[36] S. Y. Lee and N. S. Kim, "A statistical model-based residual echo suppression," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 758–761, 2007.

[37] D. Giacobello, M. Semmoloni, D. Neri, L. Prati, and S. Brofferio, "Voice activity detection based on the adaptive multi-rate speech codec parameters," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2008.

[38] D. Malah and R. Cox, "A generalized comb filtering technique for speech enhancement," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 7, May 1982, pp. 160–163.

[39] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. International Workshop on Machine Learning for Multimodal Interaction*, pp. 182–195, 2004.

[40] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

ISBN 978-90-827970-1-5 © EURASIP 2018