Source number estimation for multi-speaker localisation and tracking

Nilesh Madhu[1] *and Rainer Martin*[2]

[1]IDLab, Ghent University - imec, Belgium
[2]Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
nilesh.madhu@ugent.be, rainer.martin@rub.de

Abstract

Presented is an approach for source number estimation, applicable when performing multi-source localisation by a mixture-of-distributions model. The approach is derived from information-theoretic considerations and allows for estimating the number of sources both in non-competing and in concurrent talker situations. We also propose instrumental metrics for evaluating the performance of source localisation and tracking algorithms in a multi-source scenario. We implement and evaluate the proposed source number estimation approach in the framework of our previously proposed localisation algorithm based on a mixture-of-Gaussians model. We show that using the source number estimator reduces the number of 'ghost' localisations (i.e. spatial locations falsely deemed as containing an active acoustic source) without compromising on the accuracy of the localisation estimates.

Index Terms: source localisation, source separation, source number estimation, hearing devices.

1. Introduction

The aim of multiple talker localisation is to detect and localise a number of overlapping or competing speakers, under varying levels of background noise, by means of the spatial diversity afforded by microphone arrays. Active speaker localisation is an important part of the noise suppression chain in state-of-the-art communication systems and finds its use, for example, in steering the video camera towards the active speaker in video conferencing scenarios, for interference cancellation and noise suppression in hands-free systems, and hearing devices.

An important problem in this context is the adaptive detection of the number of active speakers. Detection can be done using Akaike's Information Criterion (AIC), Rissanen's Minimum Description Length (MDL), or the Bayesian Information Criterion (BIC) (see, e.g, [1] and references therein). But the formulation of these criteria is difficult for the broadband case, especially where disjoint sources like speech are concerned. Moreover, the detection problem is coupled with the localisation, requiring a multi-dimensional non-linear maximum likelihood optimisation, which adds to the complexity. The approach proposed in [2] decouples the localisation from source number detection and bases the estimate on the eigenvalue distribution. However, this approach requires, firstly, a sufficiently large number of microphones in order to evaluate the dynamics of the eigenvalues. Secondly, it requires sources to be simultaneously active during a period sufficient to gather the statistics. Lastly, it implicitly assumes sources have similar power and the interference is spatially white. These conditions are too restrictive for speech signals in a natural scenario. Such scenarios are, by nature, dynamic: a speaker may start, be active for a while, fall silent, and then start again. Even *within* active speaker segments, we have speech pauses. Speech signals also demonstrate sparsity and disjointness, which makes it difficult to adapt this method for speech. For these reasons, most applications either assume the number of concurrently active speakers to be known or implicitly assume a single dominant speaker.

We propose a method for source number estimation that imposes *neither* the constraint of constant multi-speaker activity (competing situation) *nor* that of single source dominance. We build upon the approach developed in [3], since this gives us an existing framework to evaluate our approach. However, we note that other mixture-of-distributions model can also be used such as those employing super-Gaussian kernels [4] or circular distributions [5]. We begin with a brief discussion of the signal model and the GMM-based localisation. We then present the source number estimator. Next we propose the two instrumental measures to evaluate the performance of localisation algorithms before testing the complete approach (localisation, source number estimation and tracking) on single- and multiple- speaker recordings made in a reverberant and noisy room.

2. Signal model

The signal model we consider is that of a compact array of M microphones at positions $\boldsymbol{r}_m = (x_m, y_m, z_m)^T$, capturing the signals emitted from Q sources at positions $\boldsymbol{r}_q = (x_q, y_q, z_q)^T$. We consider a spectral representation obtained from the K–point discrete Fourier transform (DFT) on overlapped, windowed segments of the discrete time domain signal [6, 7]. The signals recorded by any microphone m may then be formulated as:

$$X_m(k,b) = \sum_{q=1}^{Q} A_{0,mq}(k)\, S_{0,q}(k,b) + V_m(k,b), \quad (1)$$

where k and b are, respectively, the discrete frequency bin index and frame index; $A_{0,mq}(k)$ is the room transfer function from $\boldsymbol{r}_q$ to $\boldsymbol{r}_m$; $S_{0,q}(k,b)$ is the signal produced by source q; and $V_m(k,b)$, the noise at microphone m. Each $A_{0,mq}(k)$ may be further decomposed as

$$A_{0,mq}(k) = |A'_{0,mq}(k)\,|e^{-\jmath\,\Omega_k \tau_{mq}} + A''_{0,mq}(k)\ , \quad (2)$$

where $|A'_{0,mq}|$, represents the gain along the direct path and $A''_{0,mq} \in \mathbb{C}$ indicates the net gain and phase smearing caused by the reflections along the indirect paths. τ_{mq} represents the *absolute* time delay of the signal from source q to the microphone m along the direct path and $\Omega_k = 2\pi k f_s/K$ represents the k-th discrete frequency. Usually, the direct path is assumed dominant and the effect of the indirect paths is subsumed into the noise. Further, the model is usually simplified by considering the signals received at the first microphone through the direct path as the reference:

$$S_q(k,b) = |A'_{0,1q}(k)\,|e^{-\jmath\,\Omega_k \tau_{1q}} S_{0,q}(k,b) \quad (3)$$

This gives us the following compact representation:

$$\boldsymbol{X}(k,b) = \boldsymbol{A}(k)\,\boldsymbol{S}(k,b) + \boldsymbol{V}(k,b)\,, \tag{4}$$

with

$$A_{mq}(k) = \left|\frac{A'_{0,mq}(k)}{A'_{0,1q}(k)}\right| e^{\jmath\,\Omega_k \Delta\tau_{mq}}\,,$$

where $\Delta\tau_{mq} = \tau_{1q} - \tau_{mq}$ is the *time delay of arrival* (TDOA) with respect to the reference sensor. The signal vectors in (4) are defined as $\boldsymbol{X}(k,b) = \big(X_1(k,b), \ldots, X_M(k,b)\big)^T$, $\boldsymbol{V}(k,b) = \big(V_1(k,b), \ldots, V_M(k,b)\big)^T$, and $\boldsymbol{S}(k,b) = \big(S_1(k,b), \ldots, S_Q(k,b)\big)^T$, respectively.

In addition to the spatial diversity, a property of speech signals used frequently is their sparsity and *approximate disjointness* [8] in the short-time Fourier transform (STFT) domain. This means that the STFT spectra of any two speaker signals overlap at very few time-frequency (T-F) points (k,b). Consequently for a T-F point occupied by the source q, the signal model of (4) may be approximated as:

$$\boldsymbol{X}(k,b) \approx \boldsymbol{A}_q(k)\,S_q(k,b) + \boldsymbol{V}(k,b)\,. \tag{5}$$

For a given sampling frequency, the disjointness is dependent upon the resolution K of the DFT, the number of simultaneously active speakers Q, and the amount of reverberation (quantified by the reverberation time T_{60}) present. This influence of the parameters on the disjointness assumption has been evaluated in more detail in [9] for a sampling frequency of 16 kHz, where it is shown that the disjointness attains its maximum value for $K \in \{512,\, 1024,\, 2048\}$. Correspondingly, we fix our DFT resolution to lie in this range.

3. Source localisation

The sparsity and disjointness of speech in the time-frequency plane is exploited for multi-source localisation. The algorithm of choice is frequently a variant of the steered response power (SRP) [10, 11] algorithm due to its ease of implementation and scalability in terms of addition of new sensors and selection of the candidate locations (see e.g., [12]). Consequently, this algorithm is adopted in its narrowband form in [3, 4, 10] as the basic building block of the localisation framework.

For localisation along the azimuth direction of arrival (DOA) θ the narrowband SRP cost function at a frequency bin k and a time frame b is obtained as:

$$\mathcal{J}_{\mathrm{SRP}}(\theta,k,b) = \left|\boldsymbol{H}^H_{\mathrm{PHAT}}(\theta,k,b)\boldsymbol{X}(k,b)\right|^2\,, \tag{6}$$

where $\boldsymbol{H}_{\mathrm{PHAT}}(\theta,k,b)$ is a delay-and-sum beamformer, normalised by the amplitude of the signal at the respective microphone (SRP-PHAT). With the first microphone taken as the reference sensor, the SRP-PHAT beamformer expression is:

$$\begin{aligned}&\boldsymbol{H}_{\mathrm{PHAT}}(\theta,k,b) \\ &= \Big(\frac{1}{|X_1(k,b)|}, \frac{e^{\jmath\,\Omega_k\Delta\tau_2(\theta)}}{|X_2(k,b)|}, \ldots, \frac{e^{\jmath\,\Omega_k\Delta\tau_M(\theta)}}{|X_M(k,b)|}\Big)^T\,,\end{aligned} \tag{7}$$

where $\Delta\tau_m(\theta)$ is the relative TDOA of a source wavefront at microphone m, when the source is located along θ.

The $\mathcal{J}_{\mathrm{SRP}}(\theta,k,b)$ is computed for a subset of K' frequencies $k \in \{k_{\mathrm{low}}, k_{\mathrm{max}}\}$ over the pre-selected grid of search locations. Under the assumption of sparsity and disjointness of speech in the STFT domain, each time-frequency point (k,b) can be attributed to the dominant speaker at that point. Correspondingly, the localisation estimate, obtained as the maximum of the SRP cost function, is the location estimate of the dominant source at that T-F point:

$$\widehat{\theta}(k,b) = \underset{\theta}{\mathrm{argmax}}\; \mathcal{J}_{\mathrm{SRP}}(\theta,k,b). \tag{8}$$

The source location estimates $\big\{\widehat{\theta}(k_{\mathrm{low}},b),\ \ldots,\ \widehat{\theta}(k_{\mathrm{max}},b)\big\}$ obtained at frame b, are next *clustered* to obtain the multi-source location estimates. A mixture of Gaussians (MoG) model is employed for the clustering:

$$\widehat{\theta} \sim \sum_{i=1}^{\mathcal{I}} P_i\,\mathcal{N}\left(\theta_i, \sigma_i^2\right)\,, \tag{9}$$

where $\mathcal{I}$ is the model order, P_i is the weight (*a priori* evidence), and θ_i is the centroid of the i th element, with the corresponding variance σ_i^2 indicating the spatial spread of that component. The parameters are estimated using the Expectation-Maximisation (EM) [13] algorithm.

This clustering is done on a per-frame basis. In the following, we shall subsequently drop the frame index for convenience and reintroduce it when necessary. As the number of sources is not known *a priori*, we start with a pre-defined model order $\mathcal{I}$. The EM clustering on the set of $\widehat{\theta}(k)$ values then yields

the means : $\boldsymbol{\theta} = \big(\theta_1,\ \ldots,\ \theta_{\mathcal{I}}\big)^T$,
the variances: $\boldsymbol{\Xi} = \big(\sigma_1^2,\ \ldots,\ \sigma_{\mathcal{I}}^2\big)^T$, and
the weights/probabilities: $\boldsymbol{P} = \big(P_1,\ \ldots,\ P_{\mathcal{I}}\big)^T$

of the $\mathcal{I}$ components. Since the initial value of $\mathcal{I}$ is chosen to overestimate the underlying process, the model is iteratively shrunk and re-estimated such that mean-values lying within a shrink threshold Υ_θ of each other are merged and the model order is correspondingly recomputed.

4. Source number estimation

The model obtained after the iterative shrinkage might still contain clusters not belonging to any source. Such a situation occurs typically when the model utilises its degrees of freedom to model outliers. Such ghost components may be reduced by the following information-theoretic consideration. The weights P_i define a discrete probability distribution with *entropy* $\mathfrak{H}(\mathcal{I}_1)$,

$$\mathfrak{H}(\mathcal{I}_1) = \sum_{i=1}^{\mathcal{I}_1} P_i \log_2(P_i). \tag{10}$$

Using this value, we may estimate the number of *significant* components in the model as:

$$\mathcal{I}_2 = \underset{\mathcal{I}'}{\mathrm{argmin}}\ \left|\mathfrak{H}(\mathcal{I}_1) - \log_2(\mathcal{I}')\right|,\ \mathcal{I}' \in \{0,\ \ldots,\ \mathcal{I}_1\}. \tag{11}$$

If $\mathcal{I}_2 < \mathcal{I}_1$,

- select the $\mathcal{I}_1 - 1$ components with the highest weights,
- normalise weights to yield a well defined but reduced probability distribution, i.e.,

$$P_i \leftarrow P_i \Big/ \sum_{i=1}^{\mathcal{I}_1 - 1} P_i \quad \forall i \le (\mathcal{I}_1 - 1) \tag{12}$$

- $\mathcal{I}_1 \leftarrow \mathcal{I}_1 - 1$ and repeat (10)–(12) until $\mathcal{I}_2 = \mathcal{I}_1$.

Note that, irrespective of the difference between $\mathcal{I}_1$ and $\mathcal{I}_2$, the shrinkage in the first step only reduces the number of elements by 1. This conservative shrinkage strategy was chosen as it yielded the best results as compared to a direct shrinkage by $\mathcal{I}_1 - \mathcal{I}_2$ elements. The value of $\mathcal{I}_2$ after convergence is the source number estimate for that frame.

The rationale for the computation of $\mathcal{I}_2$ as in (11) derives from the information theoretic relationship between entropy and the optimal average symbol bitlength in source coding. If all the components of the estimated MoG representation are equally significant, we would have a uniform distribution on the P_i, and $\mathfrak{H} = \log_2(\mathcal{I}_1)$. If, however, this distribution is 'peaky', some components are more significant, and this information can be used to reduce the ghost clusters by the above iterative procedure. After the clustering per time-frame, the non-linear token-based source tracking discussed in [3] is applied, in order to preserve the source locations during short pauses and to track slowly moving sources.

5. Experimental evaluation

As described, the localisation approach consists of three stages: the parametric clustering of the bin-wise estimates, source-number estimation and model re-estimation and across-frame tracking. We now present a rigorous evaluation of the complete approach vis-à-vis a modified version of the traditional SRP-PHAT approach that can localise multiple sources. Principally this modification consists of selecting a *maximum* of $\mathcal{I}_{\mathrm{SRP}}$ local maxima from the spectrally-averaged SRP-PHAT:

$$\mathcal{J}_{\mathrm{SRP}}(\theta, b) = \sum_k \mathcal{J}_{\mathrm{SRP}}(\theta, k, b)\,. \tag{13}$$

yielding, *for each frame* b, an estimate of the location of sources $q = 1, \ldots, \mathcal{I}_{\mathrm{SRP}}$. This algorithm is denoted in the following as the multi-source SRP (M-SRP). For a fair evaluation, the M-SRP is compared to both the raw, frame-by-frame output obtained by MoG clustering (denoted as $\mathrm{MoG_r}$), and to the complete algorithm incorporating the tracking framework after the MoG clustering (subsequently denoted as MoG).

5.1. Experimental setup & parameter settings

The data was recorded with the 5-channel, linear microphone array (depicted in Figure 1) with the sources in the far-field. This leads to localization and separation models based on the DOA θ, measured with respect to the array axis. Throughout the simulations and the comparisons, the system parameters were set as in Table 1. Ten sentences (5 male, 5 female speakers) from the TIMIT database were used for the evaluation. To generate the test scenarios the speaker signals were individually played back through a loudspeaker (Genelec 2029BR) positioned at a distance of 1 m from the center of the array, for different angles of arrival, and recorded by the array. The recordings were made in a reverberant office room of dimensions $5.7\,\mathrm{m} \times 7.4\,\mathrm{m} \times 2.9\,\mathrm{m}$, a reverberation time T_{60} of 0.6 s and a critical distance [14] of 0.85 m thus placing the sources effectively *outside* the critical distance for our experiments. The competing speaker situation was created by additively mixing

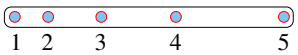


Figure 1: *Linear array used in this work, with inter-sensor spacings of* $d_{12} = 3\,cm$, $d_{23} = 5\,cm$, $d_{34} = 7\,cm$, *and* $d_{45} = 10\,cm$

the individual speaker recordings at the required azimuths, at 0 dB signal-to-interference ratio (SIR). Sources were recorded at azimuths of $\theta_q \in \{60^\circ, 90^\circ, 120^\circ\}$.

We test the localisation performance under two different types of background noise added to the mixtures at varying signal-to-noise ratio (SNR) ($\in \{0\,\mathrm{dB}, 10\,\mathrm{dB}, 20\,\mathrm{dB}\}$). The noise types considered were (a) white noise recorded using the microphone array in a *diffuse* environment; and (b) cafeteria babble recorded using the microphone array. Note that the noise signals are spatially *correlated* (at least for low frequencies).

5.2. Performance measures

The performance of the respective algorithms are tested according to their (a) hit percentage and (b) localisation accuracy. The hit percentage $\mathcal{Z}$ is defined as the percentage of time frames in which the algorithm estimates a source position in the *vicinity* of the true source, with the vicinity threshold being set to $\Upsilon_{\mathrm{hit}} = \pm 10^\circ$, since this seems to be a realistic value for a minimum physical separation between two (human) speakers, if they are clustered around an array in a real scenario. For the two-source case we define two kinds of hit percentages,

1. $\mathcal{Z}_1$: the percentage of frames where the algorithm localises *at least* one source within its vicinity (i.e. to within $\Upsilon_{\mathrm{hit}} = \pm 10^\circ$), and
2. $\mathcal{Z}_2$: the percentage of frames where the algorithm localises both sources within their respective vicinities.

In all our evaluation samples the speakers were always active, with no significant speech pauses. This obviates the need for a separate voice activity detector for the performance evaluations.

The localisation accuracy is measured by the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\mathrm{E}\{(\theta - \widehat{\theta})^2\}}, \tag{14}$$

where the expectation is replaced by a temporal average in practice. For this averaging we only consider the $\widehat{\theta}$ in frames where the sources have been localised to within the hit threshold.

The results for each combination of background noise, SNR and position are averaged over all the speakers for the *single* speaker case. This corresponds to averaging over 30 speaker signals for each setting of background noise and SNR. Further, for every combination of background noise type and SNR, each speaker in the set was simulated at each of the three azimuth positions, and against every other speaker in a corresponding 'interference' position. The results corresponding to the same azimuthal *difference* in the speaker location are averaged over all such speaker combinations, e.g., results for $\{\theta_1 = 60^\circ, \theta_2 = 90^\circ\}$ and $\{\theta_1 = 90^\circ, \theta_2 = 120^\circ\}$ are averaged over all the speaker combinations in these positions. This corresponds to an average over $10 \cdot 9 = 90$ combinations for an azimuthal separation of 60° and over $2 \cdot 10 \cdot 9 = 180$ combinations for an azimuthal separation of 30°. For reasons of space, only the results for the sources separated by $\Delta\theta_q = 30^\circ$ are presented. This is also the most challenging scenario, given the close proximity of the sources.

Table 1: *Parameters for the evaluations,* $f_{\mathrm{s}} = 8\,\mathrm{kHz}$

DFT length K (ms)	Frame shift (ms)	Window type/ length (ms)	$\mathcal{I}$	$\mathcal{I}_{\mathrm{SRP}}$	Υ_θ ($^\circ$)
128	16	von Hann/128	5	5	10

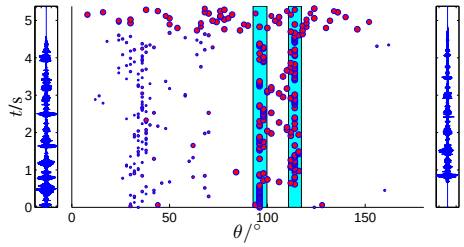
Table 2: *Localisation of concurrent speakers, azimuthal separation* $= 30^\circ$.

Noise type	SNR (dB)	$\mathcal{Z}_1$ (%)			$\mathcal{Z}_2$ (%)			RMSE ($^\circ$)		
		SRP	MoG_r	MoG	SRP	MoG_r	MoG	SRP	MoG_r	MoG
Clean	N/A	100	100	100	82.54	91.74	97.96	2.14	2.47	2.35
White, diffuse	0	97.09	99.34	100	67.73	93.25	99.69	2.86	3.33	2.52
	10	99.62	99.98	100	71.82	92.30	99.74	2.32	2.62	2.22
	20	100	99.88	100	77.16	90.87	98.78	2.02	2.36	2.25
Babble, diffuse	0	99.82	99.95	100	79.93	91.34	99.14	1.88	2.62	2.40
	10	100	99.98	100	80.81	91.01	98.86	1.83	2.53	2.48
	20	100	99.98	100	81.90	90.88	98.85	1.99	2.41	2.50

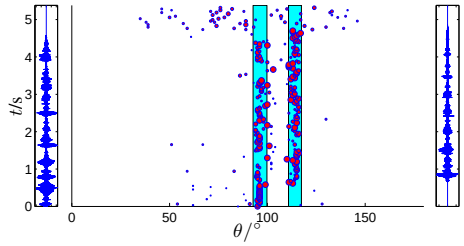
6. Results and Discussion

In Figure 2 we illustrate the performance of the algorithm on a sample setup consisting of two simultaneously active sources at $\theta_1 = 90^\circ$ and $\theta_2 = 120^\circ$ respectively. In the M-SRP plots, each marker corresponds to a maximum of the spectrally averaged cost function of (13). The *size* of the marker is proportional to the value of the cost function at that position, with the absolute maximum being normalised to unity. Thus, markers of smaller size indicate less intense peaks in (13), and may correspond *either* to a weaker source or an erroneous estimate. In the MoG_r plots, the markers indicate the localised sources in that time frame, with the size of the markers being proportional to the *a posteriori* probability of the localised source (maximum value, again, being normalized to unity). In the plots of the complete MoG approach (i.e., including the tracking), the markers correspond to the localised sources in the time-averaged model and their size is proportional to their token (normalised to 1). Thus, while the MoG_r plots indicate the reliability of estimation of a source, the MoG plots indicate the introduction of a grace period during speech pause and when the speaker fades, and the reduction of ghost clusters. The benefit of the source number estimation and model re-computation is visible by the low number of ghost clusters in both the MoG approaches.

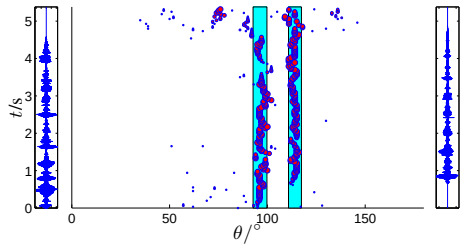
In Table 2 we see that the M-SRP and MoG approaches localise the sources with comparable accuracy. This is partly to be expected as the MoG algorithm utilises the SRP function to obtain the individual source estimates. With respect to the hit percentage for single source localisation, the performance of M-SRP and MoG_r are very similar at high SNRs. In low SNR conditions we have a flattening of the spectrally averaged SRP function along with the introduction of spurious maxima, which leads to false estimates of the source positions using this approach. On the other hand, as the MoG approach performs a localisation in each frequency bin, enough information is available (due to the correct localisation in bins with high SNR) to localise the sources. Consequently MoG_r performs better than M-SRP under low SNR conditions. The introduction of the tracking mechanism further improves the performance of the MoG-based approach, as this framework can preserve the source location estimate over frames in low-SNR segments or during speech pauses.



(a) M-SRP, two closely spaced sources



(b) MoG_r for the same situation



(c) MoG with tracking

Figure 2: *Performance of the M-SRP,* MoG_r*, and MoG algorithms for the case of two concurrent, closely spaced sources.*

7. Conclusions

When the number of sources to be localised is not known, or when it is time-variant, traditional model-order estimation algorithms are difficult to formulate. This problem is compounded for speech signals given their sparsity, non-stationarity, disjointness of spectra, and the real-time requirements. This contribution has presented an information-theory inspired algorithm for source number estimation. This was integrated into a previously proposed localisation and tracking approach and the combination was rigorously evaluated. The framework outperforms the generic SRP in the case of concurrent sources, more so when the spatial separation between the sources to be localised decreases. The source number estimation and model re-computation significantly reduces the number of ghost-clusters without compromising on the localisation accuracy. With the increase in computational power and the amount of portable computation power available (e.g. in smartphones), such localisation methods are becoming increasingly relevant for modern hearing aids, to assist, e.g, in the beamforming and source separation stages.

8. References

[1] M. Wax and T. Kailath, "Determining the number of signals by information theoretic criteria," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9, 1984, pp. 232–235.

[2] F. Cong, A. Nandi, Z. He, A. Cichocki, and T. Ristaniemi, "Fast and effective model order selection method to determine the number of sources in a linear transformation model," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2012.

[3] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, Sep. 2008.

[4] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian mixture model," *Digital Signal Processing*, vol. 21, no. 1, pp. 66 – 76, 2011.

[5] Q. Nguyen Dinh and C.-H. Lee, "Model-based clustering of DOA data using von Mises mixture model for sound source localization," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 13, no. 1, pp. 59–66, 2013.

[6] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice Hall, 1975.

[7] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Ltd., 2006.

[8] S. Rickard and Ö. Yilmaz, "On the approximate W-Disjoint orthogonality of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002.

[9] Ö. Yilmaz, A. Jourjine, and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, Jul. 2004.

[10] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin: Springer-Verlag, 2001, pp. 157–180.

[11] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. New York, USA: John Wiley & Sons, Ltd., 2008, pp. 135–170.

[12] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, III, "Performance of real-time source-location estimators for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 593–606, Jul. 2005.

[13] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," U.C. Berkeley, Tech. Rep. TR-97-021, 1998.

[14] J. Blauert and N. Xiang, *Acoustics for engineers: Troy lectures*. Springer Verlag, 2008.