THRESHOLD SELECTION FOR UNSUPERVISED DETECTION, WITH AN APPLICATION TO MICROPHONE ARRAYS

Guillaume Lathoud, Mathew Magimai.-Doss and Hervé Bourlard

IDIAP Research Institute, CH-1920 Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{lathoud,mathew,bourlard}@idiap.ch

ABSTRACT

Detection is usually done by comparing some criterion to a threshold. It is often desirable to keep a performance metric such as False Alarm Rate constant across conditions. Using training data to select the threshold may lead to suboptimal results on test data recorded in different conditions. This paper investigates unsupervised approaches, where no training data is used. A probabilistic model is fitted on the test data using the EM algorithm, and the threshold value is selected based on the model. The proposed approach (1) does not use training data, (2) uses the test data itself to compensate for simplifications inherent to the model, (3) permits the use of more complex models in a straightforward manner. On a microphone array speech detection task, the proposed unsupervised approach achieves similar or better results than the "training" approach. The methodology is general and may be applied to other contexts than microphone arrays, and other performance metrics than FAR.

1. INTRODUCTION

This paper deals with the detection task. For example, in the case of speech source detection, each data sample needs to be classified as either "active" or "inactive". Usually some criterion ("activeness" in Fig. 2c) is compared to a threshold. Various possible values of the threshold correspond to various (FAR, FRR) "working points" on the Receiver Operating Characteristic (ROC) curve (Fig. 1). FAR is False Alarm Rate and FRR is False Rejection Rate. This paper investigates *automatic threshold selection*: the main focus is *not* to improve the global characteristic of the detector (ROC curve), but rather to be able to select *a priori* a user-specified working point (target value $\mathrm{FAR_T}$), see Fig. 1. The FAR must remain as constant as possible across various conditions (noisy, clean etc.).

Trying to obtain *a priori* a fixed, given $\mathrm{FAR_T}$ could be useful for intrusion detection, as in password verification, where the number of false alarms needs to be stable across users and noise conditions, in order to make the system usable for regular users as well as efficient enough to detect unwanted intruders. With "training" approaches, a threshold value is usually selected on training data, on which the true classification (ground-truth) is known. The threshold is then kept fixed and applied on new, unseen test data. If training and test data represent very different conditions (e.g. noisy and clean), a fixed threshold leads to suboptimal results. Although variations exist, such as time-varying threshold learning approaches [1] and validation approaches [2], all are intrinsically limited by the overall variety of the "training" data: this is the "generalization" issue.

Alternatively, unsupervised approaches allow for *condition-dependent threshold selection*, on the test data itself, as in a heuristical study on Electro-Encephalogram classification [3]. In the present paper, the threshold value is selected in a principled way, by *predicting* the FAR *a priori*, without training data. On *each* test data (recording), a probabilistic model is fitted using EM [4]. From the fitted model, a threshold value is chosen, such that an estimate of the FAR will be close to a user-specified target value $\mathrm{FAR_T}$. These approaches realize composite hypothesis testing [5], where the result can be sensitive to the quality of the parameter estimation. This paper proposes a "model+data" posterior-based approach that compensates model imperfections, using the test data itself, and permits to use multidimensional models in a straightforward manner.

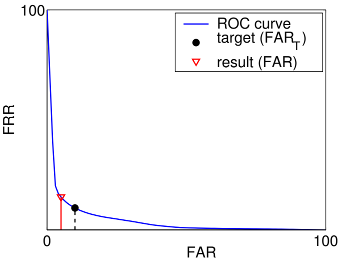


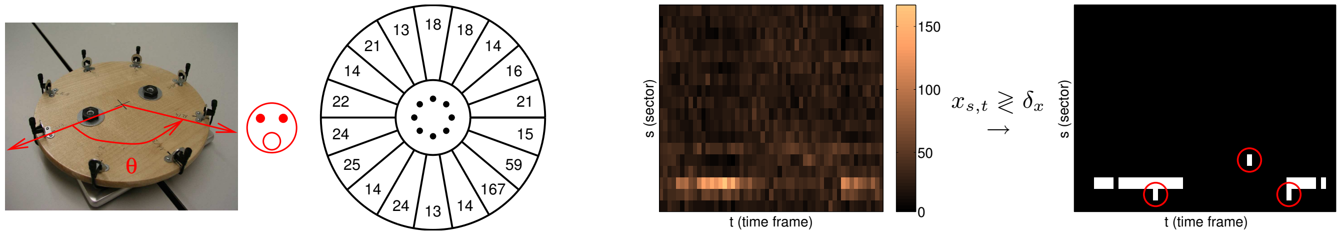
Fig. 1. ROC curve. The task is to select a threshold δ_x such that the obtained FAR (red triangle) is as close as possible to the target $\mathrm{FAR_T}$ (black dot). Ideally $\mathrm{FAR} = \mathrm{FAR_T}$.

Approach:	training	model only	model+data	
Dimensionality	$D = 1$	$D = 1$	$D = 1$	$D > 1$
Probabilistic model	none	EM fitting on "test" data. (no ground-truth)		

Table 1. Threshold selection approaches used in this article.

Results are reported on a microphone array detection task, where speakers in a meeting room must be correctly detected and located. Both space and time are discretized, and for each (sector of space, time frame) pair an "activeness" value is estimated, as in [6, 7]. Compared to the "training" approach, unsupervised model-based approaches (see Tab. 1) "generalize" better. The obtained FAR is more stable across conditions, without using training data. The proposed approach is generic, and could be applied to other tasks than microphone array detection, and other metrics than FAR. A preliminary experiment on FRR confirms its superiority over "training".

The rest of this paper is organized as follows. Section 2 describes the microphone array speech detection task. Section 3 describes the "training" approach, and experiments highlight the generalization issue. Section 4 presents the proposed unsupervised model-based approaches, along with experimental results. Application to multidimensional models is presented in Section 5. Section 6 concludes. The main focus being the threshold selection task, the probabilistic models briefly summarized here (see [8] for full details).



(a) 8-mic. array on a table

(b) Activeness values $(x_{1,t} \cdots x_{S,t})$ for one time frame t

(c) Activeness values $\{x_{s,t}\}$ for all time frames (Eq. 1)

(d) Final decisions $\{\hat{c}_{s,t}\}$ (0 or 1). A red circle is a false alarm.

Fig. 2. Sector-based detection (20-degree sectors): multichannel waveforms (a) are transformed into "activeness" values (b,c), as in [6, 7], which are thresholded to obtain the final decision (d). A false alarm happens when ground-truth $c_{s,t} = 0$ and final decision $\hat{c}_{s,t} = 1$.

2. THE TASK: DETECTION WITH MICROPHONE ARRAY

A microphone array (Fig. 2a) can be used to detect jointly where and when a given person is speaking, as already reported in meeting rooms [6] and cars [7], and briefly summarized here. It can detect multiple people talking concurrently, as often happens in spontaneous multi-party speech [9], as in meetings.

Both space and time are discretized, respectively into volumes of spaces (e.g. 20-degree radial sectors around the array, Fig. 2b), and short time-frames (20 to 30 ms). For each time-frame, a discrete frequency-domain analysis called "SAM-SPARSE-MEAN" [7] estimates the "activeness" of a sector as the bandwidth occupied by the acoustic sources in that sector. Since speech is a wideband signal, the larger "activeness" is, the more likely there is at least one active source in the corresponding sector. A time-frame t of samples from multiple microphones (Fig. 2a) is transformed into a vector $(x_{1,t} \cdots x_{S,t})$ of activeness values (Fig. 2b) as follows [6, 7]:

- Process each FFT frequency bin separately.
- *Average* the delay-sum power [10] within a sector of space.
- *Sparsity assumption:* for each frequency bin, only one active sector, the one with maximum delay-sum power.
- Activeness $x_{s,t}$ of a given sector s = number of frequency bins where sector s is dominant, at time t ($1 \leq s \leq S$ and $1 \leq t \leq T$).

Repeating this process over time yields a spatio-temporal pattern of "activeness" (Fig. 2c). The set of all values $x_{s,t}$ is written:

$$\{x_{s,t}\} \overset{\text{def}}{=} \{ \, x_{s,t} \quad | \quad 1 \leq s \leq S, \quad 1 \leq t \leq T \, \} \tag{1}$$

Detection task: One final *binary* decision $\hat{c}_{s,t} = 0$ or 1 is taken for each activeness value $x_{s,t}$ by comparing it to a threshold δ_x (Figs. 2c,d). Errors are made such as False Alarms (circled in Fig. 2d). Performance metrics such as FAR [2] are derived by comparing all final decisions $\{\hat{c}_{s,t}\}$ with a ground truth $\{c_{s,t}\}$:

$$\text{FAR} \overset{\text{def}}{=} \frac{\text{Number of false alarms}}{\text{Number of silent samples}} \tag{2}$$

$$= \frac{\text{card}\,\{\,(s,t) \quad | \quad c_{s,t} = 0 \text{ and } \hat{c}_{s,t} = 1\,\}}{\text{card}\,\{\,(s,t) \quad | \quad c_{s,t} = 0\,\}} \tag{3}$$

where card $\{\cdot\}$ is the cardinal of a set. The purpose here is to select δ_x so that the actual $\text{FAR}(\delta_x) = \text{FAR}_\text{T}$ (e.g. 0.5%). The main focus is *not* to improve the ROC curve, but rather be able to select a user-specified working point on the ROC curve (Fig. 1). For various conditions (noisy/clean, different people, etc.), the ROC curve may change. Thus, adaptive approaches are desirable, where for various conditions different threshold values δ_x are selected, ensuring that $\text{FAR}(\delta_x, \text{condition}) = \text{FAR}_\text{T}$.

Data: Five real 16kHz audio sequences were taken from a meeting room audio-visual corpus available online [11], recorded with a horizontal circular 8-mic array (10 cm radius) set on a table (Fig. 2a). Complete data and description can be found at: `http://mmm.idiap.ch/Lathoud/05-ICASSP`
Seq. #1 to #3 were recorded with either 2 or 3 simultaneously active loudspeakers, at various locations. Seq. #4 has a single human speaker at various locations. Seq. #5 has multiple concurrent human speakers. Total duration exceeds 1 hour. Activeness values $\{x_{s,t}\}$ are extracted as explained above. Time frames are 32 ms long, half-overlapping (one frame every 16 ms).

3. THRESHOLD SELECTION WITH TRAINING DATA

A classical approach is to use "training" data where the ground-truth $\{c_{s,t}\}$ is known, and select a threshold δ_x such that $\text{FAR}(\delta_x) = \text{FAR}_\text{T}$. The threshold δ_x is then kept fixed and applied to any unseen "test" data. For "training" we used the first 3 minutes of Seq. #1, for "test" the remaining part of Seq. #1, and Seqs. #2 to #5.

The training/testing process was repeated for various target values FAR_T. In Fig. 4, FAR curves compare target FAR_T and obtained FAR. The FAR curve is close to ideal ($Y = X$) on loudspeaker data, but quite far from ideal on human data. Both can be explained by the big difference between the "human" condition (real speech from humans) and the "loudspeaker" condition used during training (synthetic speech from loudspeakers). The threshold δ_x selected on the training condition *does not generalize* to the test condition. The next section addresses this issue without training data.

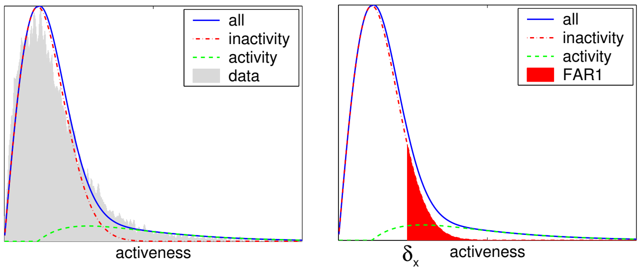
4. THRESHOLD SELECTION WITHOUT TRAINING DATA

This section proposes unsupervised approaches, where training data is not used. A probabilistic model is fitted on unseen test data using the EM algorithm [4]. The threshold value δ_x is derived from the model, such that an estimate $\overline{\text{FAR}}(\delta_x)$ is equal to FAR_T. Experiments show that the corresponding working point is closer to the target (Fig. 1), than with the "training" approach. Full details on model-based approaches are available in [8].

4.1. Unsupervised fit of a probabilistic model on test data

For a given recording, the data set $\{x_{s,t}\}$ is collected into 1 dimension, irrespective of sector in space s or time frame t (gray histogram in Fig. 3a). As shown in [8], it can be fitted with a sensible probabilistic model with 2 components f_0 ("inactivity") and f_1 ("activity"). Each component is assumed to follow a Rice distribution [12], which describes the probability density of the envelope of the sum of a sinusoidal wave and a zero mean narrow-band Gaussian noise.

No manual tuning is needed and the EM cost is very small [8], similarly to [13]. The three curves in Fig. 3a show an example of fit.



(a)
2-mixture model fitted
on test data $\{x_{s,t}\}$
$f(x) = w_0\ f_0(x) + w_1\ f_1(x)$

(b)
model-only estimation of FAR
$\overline{\mathrm{FAR}}_1(M, \delta_x)$

Fig. 3. Unsupervised fit of a 2-mixture model $M = \{w_0, w_1, f_0, f_1\}$. The histogram in (a) is a 1-dimensional view of all data $\{x_{s,t}\}$, irrespective of sector in space s or time t. w_0 and w_1 are the priors of inactivity and activity, respectively.

4.2. "model only" threshold selection

Once the model is fitted on the test data, the threshold value δ_x can be used using the model M alone (Fig. 3b), such that $\overline{\mathrm{FAR}}_1(M, \delta_x) = \mathrm{FAR_T}$, where:

$$\overline{\mathrm{FAR}}_1(M, \delta_x) \overset{\mathrm{def}}{=} \int_{\delta_x}^{+\infty} f_0(x)dx \qquad (4)$$

Since a model is always a simplification of reality, in some cases it may not fit well the data, and the $\overline{\mathrm{FAR}}_1$ estimate will be very different from the actual FAR. The selected threshold δ_x would then lead to a FAR performance very different from the desired $\mathrm{FAR_T}$.

4.3. "model+data" threshold selection

We propose to correct a possible bad fit of the model by using the test data itself. Consider the definition of FAR in Eq. 3. Numerator and denominator can be approximated with their respective conditional expectations, using posterior probabilities.

Approximation of the numerator: For a given sample $x_{s,t}$, it can be shown [8] that the probability of having a false alarm is:

$$p\,(\,c_{s,t} = 0,\ \hat{c}_{s,t} = 1 \mid x_{s,t}, M, \delta_x\,) \quad = \quad p_{s,t}^{(0)} \cdot 1_{x_{s,t} > \delta_x} \qquad (5)$$

where $1_{\text{proposition}}$ is the indicator function: $1_{\text{proposition}} = 1$ if proposition is true, 0 otherwise, and $p_{s,t}^{(0)}$ is the posterior probability of inactivity, for sample $x_{s,t}$, as derived from Bayes rule: $p_{s,t}^{(0)} = w_0 f_0(x_{s,t}) \ /\ [w_0 f_0(x_{s,t}) + w_1 f_1(x_{s,t})]$. From Eq. 5, the *expected* number of false alarms is:

$$\sum_{s,t} p\,(\,c_{s,t} = 0,\ \hat{c}_{s,t} = 1 \mid x_{s,t}, M, \delta_x\,) \ = \sum_{\substack{s,t \\ x_{s,t} > \delta_x}} p_{s,t}^{(0)} \qquad (6)$$

Approximation of the denominator: the *expected* number of silent samples (i.e. $x_{s,t}$ such that $c_{s,t} = 0$) is $\sum_{s,t} p_{s,t}^{(0)}$.

Approximation of FAR:

$$\overline{\mathrm{FAR}}_2(M, \{x_{s,t}\}, \delta_x) \overset{\mathrm{def}}{=} \sum_{\substack{s,t \\ x_{s,t} > \delta_x}} p_{s,t}^{(0)} \ / \sum_{s,t} p_{s,t}^{(0)} \qquad (7)$$

Implementation: Determining δ_x can be done in an efficient manner, by first ordering samples $\{x_{s,t}\}$ by decreasing value, irrespective of s or t, and second computing cumulative sums of posteriors $p_{s,t}^{(0)}$. The computational cost can be drastically decreased [8] by reducing the data to a fixed, small number of samples (e.g. 100).

4.4. Experiments

Fig. 4 shows the resulting FAR curves. Tab. 2 shows The Root Mean Square (RMS) of $(\mathrm{FAR}/\mathrm{FAR_T}\ - 1)$ for a practical range of small $\mathrm{FAR_T}$ values (up to 5%). This RMS metric was chosen in order to normalize results that have very different orders of magnitude (from 0.1% to 5%). Ideally the RMS value is equal to zero.

Compared to the "training" result, both model-based approaches yield a degradation on loudspeaker data and an improvement on human data. This can be explained by the absence of condition-specific tuning in the model-based approaches, in contrary to "training". Note that the "model+data" approach systematically improves over the "model only" approach.

Overall, there is a major improvement over the "training" approach in terms of robustness across conditions, especially visible in Fig. 4, Seq. #4. This is a positive result since the model-based approaches have the exact same ROC curve as the "training" approach. The next section shows that the "model+data" approach can be applied to more complex models, thus bringing further improvement.

	loudspeakers			humans	
Seq.	#1	#2	#3	#4	#5
training	**0.109**	0.142	0.154	1.898	3.929
model only	0.576	1.022	0.977	1.780	3.119
model+data	0.217	0.494	0.443	1.121	2.344
model+data (N-D)	0.117	**0.078**	**0.121**	**0.452**	**1.846**

Table 2. RMS statistic over the interval $\mathrm{FAR_T} = [0.1\%, 5\%]$. This is the RMS of $(\mathrm{FAR}/\mathrm{FAR_T}\ - 1)$: the lower, the better. The best result for each recording is indicated in boldface.

5. APPLICATION TO MULTIDIMENSIONAL MODELS

All previous approaches (training and model-based) were in 1-dimensional space: each detection decision $\hat{c}_{s,t}$ was taken based on one sample $x_{s,t}$ only. However, the "model+data" approach (Section 4.3) can be applied to more complex multidimensional models. Prior knowledge that for a given time frame t, all activeness values sum to a constant $\left(\sum_s x_{s,t} = \text{constant}\right)$ leads to joint modelling of all sectors $(x_{1,t} \cdots x_{S,t})$, as described in details in [8].

Thresholding posteriors: The "model+data" approach presented in Section 4.3, is modified by replacing the threshold on the 1-dimensional "activeness" feature $x_{s,t} \gtrless \delta_x$ with a threshold on the posterior probability of activity:

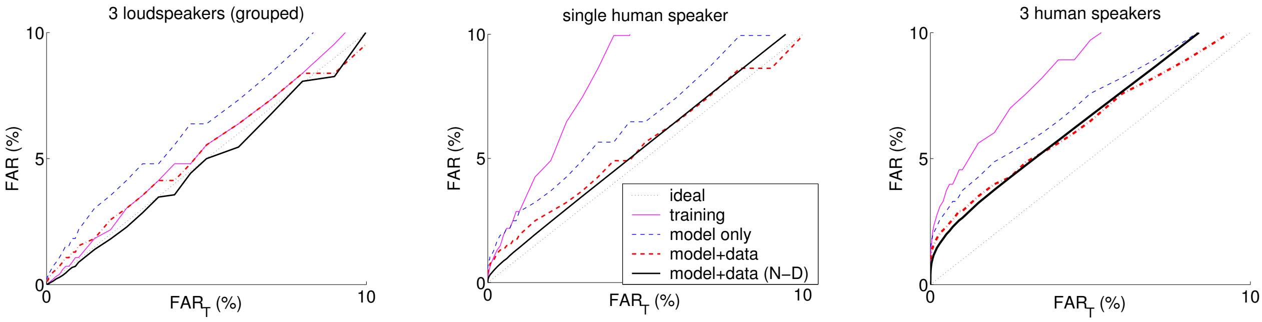
$$p\,(\,c_{s,t} = 1 \mid \{x_{1,t} \ldots x_{S,t}\}, M\,) \quad \gtrless \quad \delta_p \qquad (8)$$

Similar to Section 4.3, the threshold on posteriors δ_p can be determined on the test data itself such that $\overline{\mathrm{FAR}}_2(M, \{x_{s,t}\}, \delta_p) = \mathrm{FAR_T}$.

With a model in multidimensional space, the goal is to capture relations between several data samples $(x_{1,t} \cdots x_{s,t} \cdots x_{S,t})$. Thus, it is hoped that the model will fit the data better, which in turn will yield an estimate $\overline{\mathrm{FAR}}_2$ closer to the actual FAR.

Experiments: Fig. 4 and Tab. 2 show the results ("N-D"). In all recordings, for larger values $\mathrm{FAR_T} > 5\%$, the results are similar to those of the 1-dimensional "model+data" approach. For lower values $\mathrm{FAR_T} < 5\%$, in all recordings a systematic improvement is seen over the 1-dimensional "model+data" approach. On Seq. #1, results are similar to those of the best one: "training", which itself was tuned on part of Seq. #1. On all other recordings the multidimensional approach yields the best results of all approaches.

Overall, this result is quite interesting given that no training data is used. Note that the multidimensional approach also improves the ROC curve, as reported in [8].



Seq. #3 (Seqs. #1, #2 very similar) Seq. #4 Seq. #5

Fig. 4. FAR curves: comparison between target FAR$_\mathrm{T}$ & obtained FAR. Ideally FAR=FAR$_\mathrm{T}$. "training", "model only", "model+data" and "model+data (N-D)" correspond to Sections 3, 4.2, 4.3 and 5, respectively. In Seq. #5, the constant, positive bias is due to body noises (breathing, stomps, shuffling paper) marked as "inactivity" in the ground-truth, since their locations are unknown.

All the approaches presented here can also be applied to the FRR metric. Results [8] confirm the superiority of model-based approaches over the "training" approach (Fig. 5).

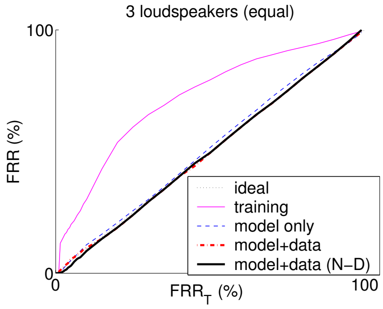


Fig. 5. Seq. #1 results with FRR metric (whole $[0\%, 100\%]$ range). Note that "training" was tuned on Seq. #1.

6. CONCLUSION

The purpose of this paper was to achieve detection so that a user-specified working point is reached, in terms of FAR. It was shown that using training data leads to the generalization issue: the detection threshold selected on training conditions may not be adequate on different test conditions. An alternative is *not* to use any training data, through unsupervised fit of a model on test data. However, the question is then: how to select the detection threshold in an adequate manner? An unsupervised model-based approach was proposed, that is robust across conditions and permits to predict the FAR as accurately or better than the "training" approach, on the microphone array task considered here. The main contribution of the paper is a method to compensate for the possible mismatch between an unsupervised model and the test data, by estimating conditional expectations over the test data itself. In particular, it allows use of complex multidimensional models in a straightforward manner. The proposed approach is generic, thus it could be applied to other tasks than microphone array sector-based detection. It can also be applied to other metrics such as FRR, for example to detect end-points prior to automatic speech recognition.

7. ACKNOWLEDGEMENTS

The authors acknowledge the support of the European Union through the AMI and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The authors would like to thank Dr Jean-Marc Odobez and Bertrand Mesot for fruitful discussions and useful suggestions.

8. REFERENCES

[1] J. Sklansky and N.J. Bershad, "The dynamics of time-varying threshold learning," *Information and Control*, vol. 15, pp. 455–486, December 1969.

[2] S. Bengio, J. Mariéthoz, and M. Keller, "The expected performance curve," in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

[3] T. Sugi, M. Nakamura, A. Ikeda, and H. Shibasaki, "Adaptive EEG spike detection: determination of threshold values based on conditional probability," *Frontiers Med. Biol. Engng*, vol. 11, no. 4, pp. 261–277, 2002.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Stat. Society, Series B*, vol. 39, pp. 1–38, 1977.

[5] W.T. Eadie, D. Drijard, and F.E. James, *Statistical Methods in Experimental Physics*, North Holland, 1971.

[6] G. Lathoud and M. Magimai.-Doss, "A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers," in *Proc. of ICASSP'05*, 2005.

[7] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-Based Detection for Hands-Free Speech Enhancement in Cars," *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Multimicrophone Speech Processing*, 2006.

[8] G. Lathoud and M. Magimai.-Doss, "Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays," IDIAP-RR-05-52, 2005.

[9] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation," in *Proc. of Eurospeech*, 2001.

[10] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[11] G. Lathoud, J.M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proc. of the MLMI'04 Workshop*, 2005.

[12] S.O. Rice, "Mathematical analysis of random noise," in *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954, pp. 133–254.

[13] G. Lathoud, M. Magimai.-Doss, B. Mesot, and Hervé Bourlard, "Unsupervised Spectral Subtraction for Noise-Robust ASR," in *Proc. of the IEEE ASRU Workshop*, 2005.