IDIAP RESEARCH REPORT

Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays

Guillaume Lathoud [a,b]
Mathew Magimai.-Doss [a,b]
Jean-Marc Odobez [a,b] Hervé Bourlard [a,b]

IDIAP–RR 05-52

October 2005

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-52

Threshold Selection for Unsupervised Detection, with an Application to Microphone Arrays

Guillaume Lathoud   Mathew Magimai.-Doss   Jean-Marc Odobez
Hervé Bourlard

October 2005

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 or development 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. In brief, 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 approaches (1) use the test data itself to compensate simplifications inherent to the model, (2) permit 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 supervised "training" approach. The methodology is general and may be applied to other contexts than microphone arrays, and other performance metrics.

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 the measured value of 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, defined as follows:

$$\mathrm{FAR} \stackrel{\mathrm{def}}{=} \frac{\text{Number of false alarms}}{\text{Number of inactive samples in the ground} - \text{truth}}, \quad (1)$$

$$\mathrm{FRR} \stackrel{\mathrm{def}}{=} \frac{\text{Number of false rejections}}{\text{Number of active samples in the ground} - \text{truth}}, \quad (2)$$

where the ground-truth is an annotation of the data, a false alarm happens when a sample is “inactive” in the ground-truth and “active” in the result, and a false rejection happens when a sample is “active” in the ground-truth and “inactive” in the result.

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 (desired target value $\mathrm{FAR_T}$), see Fig. 1. The FAR must remain as constant as possible across various conditions (noisy, clean, single soure, multiple sources etc.).

Trying to obtain an *a priori* 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 experimental 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]. The present paper presents a principled way to select the threshold within a continuum of possible values, by actually *predicting* the FAR *a priori*, without training data. On *each* test data (e.g. recording), a sensible probabilistic model is fitted using the EM algorithm [4]. A threshold value is chosen based on the model, 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. The main contribution is the “model+data” posterior-based approach that (1) compensates model imperfections, using the test data itself, (2) permits to use multidimensional models in a straightforward manner.

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 measured FAR is more stable across experimental 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 on the microphone array task highlight the generalization issue. Section 4 addresses this issue without training data, by fitting a probabilistic model on *each* test recording with the EM algorithm [4]. A major contribution is a posterior-based approximation of FAR that allows to (1) select the detection threshold *a priori*, on each test recording separately, without training data, (2) compensate simplifications made in the model, using the test data itself, and (3) use more complex, multidimensional models in a straightforward

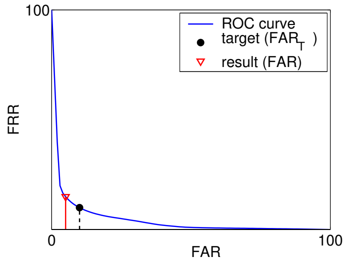


Figure 1: ROC curve. The task is to select a threshold δ_x such that the measured FAR (red triangle) is as close as possible to the desired target $\mathrm{FAR_T}$ (black dot). Ideally $\mathrm{FAR} = \mathrm{FAR_T}$. This must be achieved independently of the experimental conditions.

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.

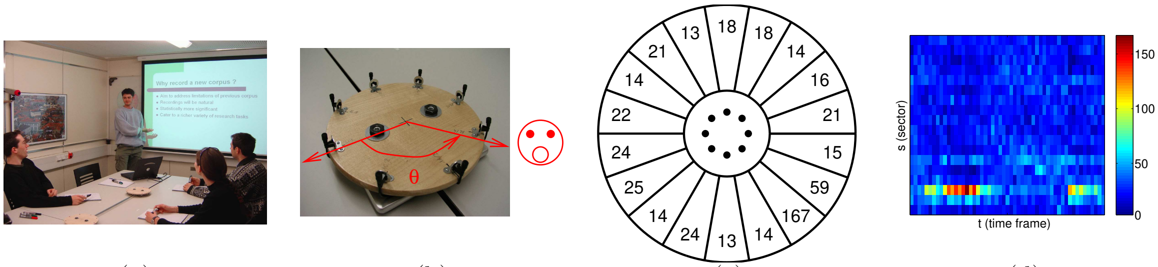
manner, as reported in Section 5. Overall, the proposed approach yields better results than the “training” approach: the variability of the FAR across experimental conditions is reduced, without using training data. Section 6 provides directions for future work, including preliminary results with another metric (FRR) that confirm the interest of the approach. Finally, Section 7 concludes. For the sake of clarity, full details, justifications and EM derivations of the probabilistic models are presented in Annexes B and C.

2 The task: detection with microphone array

A microphone array can be used to detect where and when a given person is speaking. This includes cases where multiple people speak concurrently, as often found in spontaneous multi-party speech (meetings). This section briefly summarizes a previous work that was successfully applied in both meeting rooms [6] and cars [7]. Note that the exact task addressed here is wideband acoustic source detection, of which speech is only one case. No specific measure is taken to discriminate between speech and non-speech sources. Therefore, in the rest of this paper, the terms “inactivity” and “activity” are used in order to avoid confusion with “silence” and “speech”.

Fig. 2 illustrates this approach. Both space and time are discretized, respectively into volume of spaces (e.g. radial sectors, as in Fig. 2c), and short time-frames (20 to 30 ms). For each time-frame, a discrete frequency-domain analysis called “SAM-SPARSE-MEAN” permits to estimate the “activeness” of each sector, defined as the bandwidth occupied by the acoustic sources in that sector [7]. Since speech is a wideband signal, the larger this number is, the more likely there is at least one active source in the corresponding sector. “Activeness” is the feature used for detection in the following. For details about activeness computation from the multiple waveforms, please refer to [7]. The process to transform a time-frame of samples from the multiple waveforms (Fig. 2b) to a vector of activeness values (Fig. 2c) can be summarized by the following steps:

- Process each frequency bins separately. A frequency bin is a narrowband in the FFT framework. N_{bins} is the total number of frequency bins.
- *Average* the delay-sum power within a sector (volume of space) for no additional cost [7].



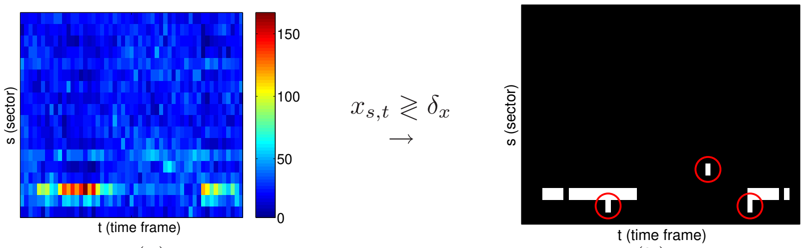
(a) Meeting room

(b) 8-mic array

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

(d) Activeness values: for all time frames $\{x_{s,t}\}$

Figure 2: Sector-based detection: multiple channels are recorded synchronously (a,b), and the activeness is estimated for each sector s (Fig. c) and each time frame t (Fig. d), as in [6, 7].



(a)
$x_{s,t} \in \{0, 1 \ldots N_{bins}\}$
Activeness value

(b)
$\hat{c}_{s,t} = 0$ or 1
final decision

Figure 3: Thresholding the activeness values $\{x_{s,t}\}$ to take the final decisions $\{\hat{c}_{s,t}\}$. Red circles mark False Alarms.

- *Sparsity assumption:* for each frequency bin, we assume that there is only one active sector, the one with maximum delay-sum power.
- Activeness $x_{s,t}$ of a given sector s at time $t \stackrel{\text{def}}{=}$ number of frequency bins where sector s is dominant at time t ($1 \leq s \leq S$, $1 \leq t \leq T$ and $0 \leq x_{s,t} \leq N_{bins}$).

Note that sparsity assumption, similar to [8], implies that, by construction, for a given time frame t, all activeness values sum to the total number of frequency bins defined by the user for the FFT (e.g. $N_{bins} = 512$): $\sum_s x_{s,t} = N_{bins}$.

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

$$\{x_{s,t}\} \quad \stackrel{\text{def}}{=} \quad \{ \ x_{s,t} \quad | \quad 1 \leq s \leq S, \quad 1 \leq t \leq T \ \} \tag{3}$$

Detection task

A final *binary* decision needs to be taken: for a given sector s and a given time frame t, is there at least one active source? A straightforward approach is to compare the activeness $x_{s,t}$ to a threshold δ_x, as illustrated in Fig. 3.

Note that the aim is to detect active sources correctly in both space and time. However, in practice the detection is never perfect and mistakes are made. See Fig. 3b for an example with False Alarms:

an "active sector" decision is taken at a time when that sector is inactive. By comparing all the final decision $\{\hat{c}_{s,t}\}$ with a ground truth $\{c_{s,t}\}$, performance metrics are derived, such as False Alarm Rate (FAR). In our case, it is the proportion of errors made on inactive (sector, time frame) pairs. For a formal definition of FAR, see Section A.

The purpose of this paper is to address the threshold selection issue: how to select δ_x so that the actual False Alarm Rate (FAR) will be equal to a desired target value $\mathrm{FAR_T}$. For example, in a practical application a user may be interested in $\mathrm{FAR_T} = 0.5\%$ of false alarms.

The goal is *not* to improve the quality of the feature $x_{s,t}$ (e.g. improve the ROC curve [2]), but rather, for a given feature $x_{s,t}$ (i.e. a given ROC curve), to be able to select a threshold δ_x (i.e. working point) that is relevant to a given task: $\mathrm{FAR}(\delta_x) = \mathrm{FAR_T}$. In the case of password verification, this would mean reaching a *user-specified* compromise between having too many false alarms (too many intruders can enter) and too many false rejections (too many users classified as intruders).

The main issue is to be able to select this threshold in an adaptive manner: for different experimental conditions (noisy/clean, indoor/outdoor, single source/multiple sources, various persons, etc.), different threshold values need to be selected to ensure that $\mathrm{FAR} = \mathrm{FAR_T}$. In other words, we need to be able to predict how well a system will perform on new, unseen data – i.e. predict its performance FAR as a function of threshold value δ_x.

3 Threshold Selection with Training Data

3.1 Data

Five recordings where made in the IDIAP meeting room, with an 8-microphone array (10 cm radius) on a table. 3 recordings (a),(b),(c) where made with 3 loudspeakers playing synthesized speech in a concurrent manner (2 or 3 loudspeakers active simultaneously). Each recording lasted 20 minutes. The loudspeaker spatial locations are known, as depicted by Fig. 4. Moreover, the use of synthesized speech permits to have an exact speech/silence segmentation for each loudspeaker. In addition, 2 recordings (d),(e) were made with humans, and annotated by a human in terms of spatial location [9] and speech/silence segmentation for each person. (d) contains a single person at several locations, while (e) contains multiple persons talking at the same time. All data used in the experiments belongs to a corpus available online [9]. The data and a detailed description are fully available at `http://mmm.idiap.ch/Lathoud/05-ICASSP`. The processing for extracting the Activeness values $\{x_{s,t}\}$ (Section 2) was done with half-overlapping time frames of 32 ms (one frame every 16 ms).

3.2 Training Approach

A classical approach is to annotate a recording (called "training data") by marking the ground-truth $c_{s,t} = 0$ or 1 for each sector s and each time frame t. Then, using this annotated data, a threshold δ_x is selected such that $\mathrm{FAR}(\delta_x) = \mathrm{FAR_T}$. Afterwards, the threshold δ_x is kept fixed and applied to any unseen data (called "test data" in the following). For training data, we used a small part (first 3 minutes) of loudspeaker recording (a). For test data, we used the remaing of (a), as well as all other recordings (b),(c),(d),(e).

The process of selecting a threshold δ_x on the training data and applying it on the test data was repeated for various desired target values $\mathrm{FAR_T}$, and the actual FAR was measured in each case. On loudspeaker data, comparison between the desired target $\mathrm{FAR_T}$ and the measured FAR is shown in Figs. 6a,b,c. The result is close to ideal, which is not surprising since the training and test conditions are very similar.

However, when applied to human data (Figs. 6d,e), the resulting curve is quite far from ideal. This is also not surprising, since the "human" condition (real speech from humans) differs a lot from the "loudspeaker" condition (synthetic speech from loudspeakers) used during training. In other words,

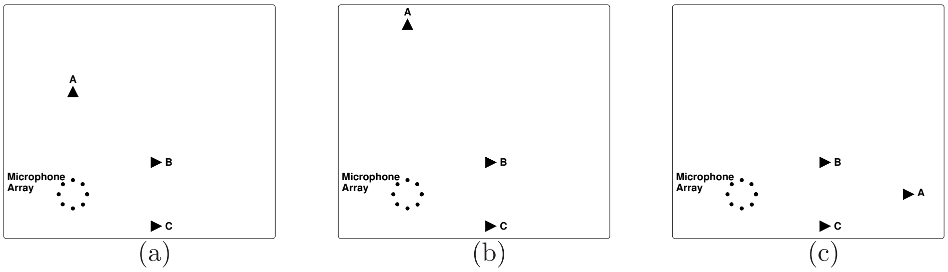


Figure 4: Setup used for the loudspeaker recordings. Each recording lasts 20 minutes, with either 2 or 3 loudspeakers playing at the same time.

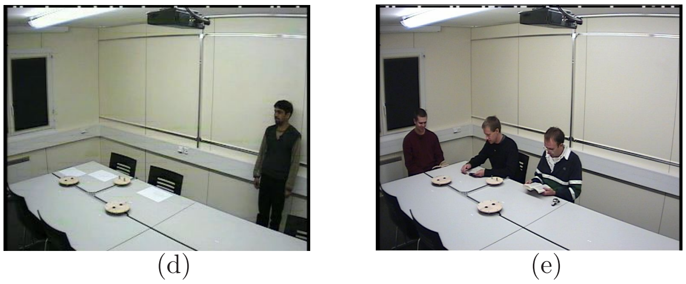


Figure 5: Recordings with humans: (d) single speaker at several locations (3 min 40 sec), (e) multiple concurrent speakers (8 min 30 sec).

a threshold δ_x selected on the training condition *does not generalize* to very different test conditions. The following sections attempt to address this issue by avoiding the use of training data.

4 Threshold Selection Without Training Data

In this section, unsupervised approaches are examined, where training data is not used. A sensible 1-dimensional probabilistic model is fitted on any unseen test data $\{x_{s,t}\}$ using the EM algorithm [4], as described in Section 4.1.

The question of selecting a threshold to take the final detection decision (as in Fig. 3) is addressed in two different manners. Section 4.2 describes an approach that only relies on the model fitted on the test data, in an unsupervised manner. Section 4.3 attempts to compensate imperfections in the model by using the same test data again, in addition to the fitted model, to select the threshold. Section 4.4 contains experimental results.

It is important to bear in mind that the threshold selection approaches described below have no impact on the ROC curve. Thus, the ROC curve is unchanged, compared to the approach with training data (Section 3). Indeed, the aim is *not* to improve the ROC curve, but rather to be able to select a working point (in terms of FAR) that will be stable across experimental conditions. This way, a robust behavior can be guaranteed, that will meet some practical specification particular to a task ($\mathrm{FAR}(\delta_x) = \mathrm{FAR_T}$) as independently as possible of the condition.

4.1 Unsupervised fit of a probabilistic model on test data

A given piece of test data $\{x_{s,t}\}$ (Fig. 7a) is extracted as described in Section 2. It is flattened into a 1-dimensional histogram, irrespective of sector in space s or time frame t (gray histogram in Fig. 7b). Next, a sensible 2-mixture model in 1-dimensional space is fitted on the histogram using

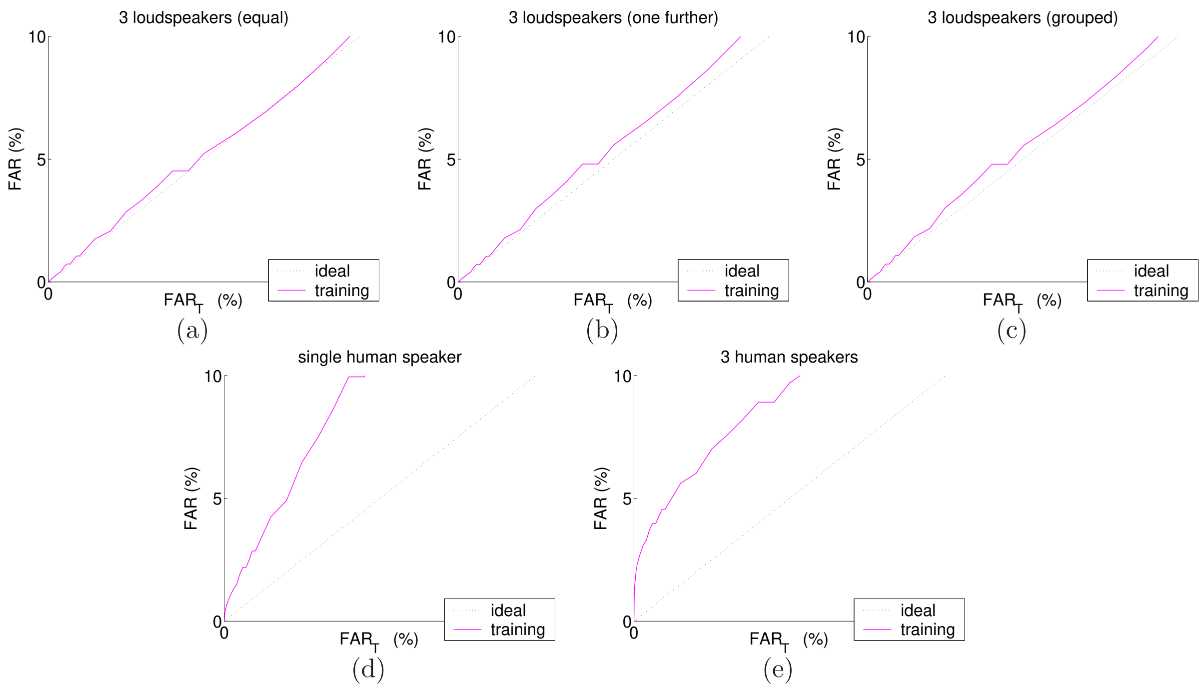
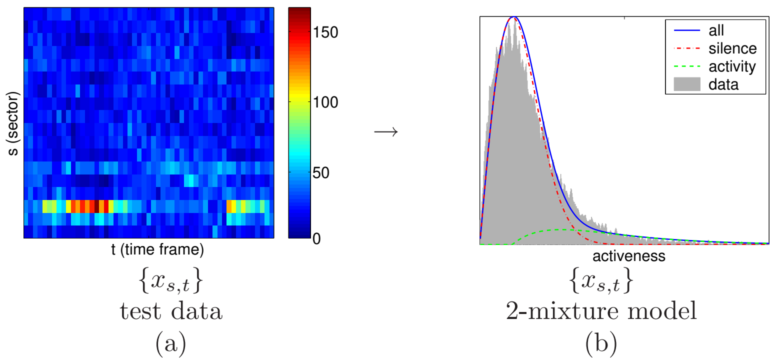


Figure 6: Threshold selection with training data, applied to loudspeaker recordings (see Figs. 4a,b,c) and human recordings (see Figs. 5d,e): comparison between target $\mathrm{FAR_T}$ and result FAR. In the "3 human speakers" case, the positive bias between FAR and $\mathrm{FAR_T}$ is due to body noises (breathing, stomps, shuffling paper) that could not be marked in the ground-truth, since their location are unkown.



$$f(x) = w_0 \cdot f_0(x) + w_1 \cdot f_1(x) \tag{4}$$

Figure 7: Unsupervised fit of a 2-mixture model. The histogram in (b) is a 1-dimensional view of all data available in (a), irrespective of space or time. w_0 and w_1 are the priors of inactivity and activity, respectively.

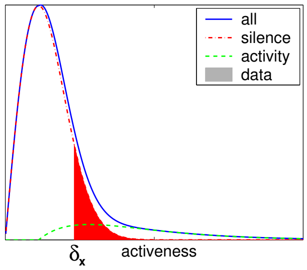


Figure 8: Estimation of the FAR for a given threshold δ_x, using the model only, through integration (red area).

the EM algorithm [4]. "2-mixture" means that one component of the model describes "inactivity", while the other component of the model describes "activity". This type of approach was introduced earlier on a different task: noise-robust automatic speech recognition, as reported in [10]. The model is described in details in Section B, along with justifications and EM derivation. In brief:

- The "inactivity" distribution $f_0(x)$ is a Rice distribution [11].
- The "activity" distribution $f_1(x)$ is a Shifted Rice distribution, which means it models values x above a moment-based statistic of f_0 (RMS value). This is a formal way to directly include in the model a practical assumption: that only values above the average "inactivity" background noise level can be distinguished from noise.

Use of such a model with the EM algorithms avoids any hand-tuning of parameters, as in [10]. The blue, green and red curves in Fig. 7b show an example of fit on the data after convergence of EM.

4.2 "model only" threshold selection

Once the model is fitted on the test data (as in Section 4.1), the next task is to select a threshold δ_x.

As illustrated in Fig. 8, one possibility is to use the model alone to estimate the FAR for a given value of the threshold δ_x:

$$\overline{\mathrm{FAR}}_1(M, \delta_x) = \int_{\delta_x}^{+\infty} f_0(x) dx \tag{5}$$

The threshold selection task then amounts to inverse the integral: select δ_x such that $\overline{\mathrm{FAR}}_1(M, \delta_x) = \mathrm{FAR_T}$.

One possible issue with this approach is over-reliance on the quality of the fit of the model. Since a model is always a simplification of reality, in some cases it may not fit well the data, as illustrated in Fig. 9. Consequently, the estimate in Eq. 5 will be very different from the actual FAR. Thus, a threshold δ_x would be selected that leads to a performance FAR very different from the desired $\mathrm{FAR_T}$. This issue is addressed in Section 4.3.

4.3 "model+data" threshold selection

This section proposes an attempt to correct a possible bad fit of the model on the test data, by using the test data again to select the threshold δ_x. Thus, the test data is used twice: first to fit the model in an unsupervised manner, as described above, second to select the threshold. Consider the definition of False Alarm Rate in Eq. 1. The rest of this section proposes to approximate numerator and denominator with their expected value, using posterior probabilities.

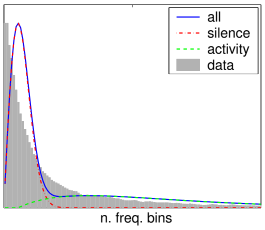


Figure 9: Example of bad fit.

Approximation of the numerator

For a given sample $x_{s,t}$, a false alarm happens when the detection decision is $\hat{c}_{s,t} = 1$ and the truth is $c_{s,t} = 0$. Since the truth $c_{s,t}$ is unknown, we estimate the probability of having a false alarm for sample $x_{s,t}$:

$$\begin{aligned} & p\left(\hat{c}_{s,t} = 1, c_{s,t} = 0\right) x_{s,t}, M, \delta_x \\ = \; & p\left(x_{s,t} > \delta_x, c_{s,t} = 0\right) x_{s,t}, M, \delta_x \\ = \; & p\left(x_{s,t} > \delta_x\right) c_{s,t} = 0, x_{s,t}, M, \delta_x \quad \cdot \quad p\left(c_{s,t} = 0\right) x_{s,t}, M, \delta_x \\ = \; & \qquad 1_{x_{s,t} > \delta_x} \qquad \cdot \qquad p_{s,t}^{(0)} \end{aligned} \tag{6}$$

where $M = \{w_0, w_1, f_0, f_1\}$ is the model, $1_{\text{proposition}}$ is the indicator function: $1_{\text{proposition}} = 1$ if proposition is true, 0 otherwise, and $p_{s,t}^{(0)}$ is the posterior probability of silence, for sample $x_{s,t}$, as derived from Bayes rule and the model M:

$$p_{s,t}^{(0)} \stackrel{\text{def}}{=} p\left(c_{s,t} = 0 \mid x_{s,t}, M\right) \tag{7}$$

$$= \frac{p\left(c_{s,t} = 0 \mid M\right) \cdot p\left(x_{s,t} \mid c_{s,t} = 0, M\right)}{p\left(c_{s,t} = 0 \mid M\right) \cdot p\left(x_{s,t} \mid c_{s,t} = 0, M\right) + p\left(c_{s,t} = 1 \mid M\right) \cdot p\left(x_{s,t} \mid c_{s,t} = 1, M\right)} \tag{8}$$

$$= \frac{w_0 \cdot f_0(x_{s,t})}{w_0 \cdot f_0(x_{s,t}) + w_1 \cdot f_1(x_{s,t})}. \tag{9}$$

Note that the last term in the last line of Eq. 6 is justified by the fact that the estimate of the posterior probability of inactivity does not depend on the threshold δ_x.

From Eq. 6, the *expected* number of false alarms is:

$$\sum_{s,t} \hat{p}\left(\hat{c}_{s,t} = 1, c_{s,t} = 0 \mid x_{s,t}, M, \delta_x\right) = \sum_{\substack{s,t \\ x_{s,t} > \delta_x}} p_{s,t}^{(0)}. \tag{10}$$

Approximation of the denominator

The *expected* number of inactive samples (i.e. $x_{s,t}$ such that $c_{s,t} = 0$) is:

$$\sum_{s,t} p_{s,t}^{(0)} \tag{11}$$

Approximation of FAR

From Eqs. 10 and 11, we propose the following FAR estimate:

$$\overline{\text{FAR}}_2(M, \{x_{s,t}\}, \delta_x) \stackrel{\text{def}}{=} \frac{\mathbf{E}\left\{\text{Number of false alarms} \mid M, \{x_{s,t}\}, \delta_x\right\}}{\mathbf{E}\left\{\text{Number of inactive samples} \mid M, \{x_{s,t}\}, \delta_x\right\}} \tag{12}$$

$$= \sum_{\substack{s,t \\ x_{s,t} > \delta_x}} p_{s,t}^{(0)} \Big/ \sum_{s,t} p_{s,t}^{(0)} \tag{13}$$

where $\mathbf{E}\{\cdot\mid\cdot\}$ is the conditional expectation.

Implementation

Determining the threshold δ_x can be done in an efficient manner, using the following steps:

- Order samples $\{x_{s,t}\}$ by decreasing value, irrespective of space or time:

$$x_1 \geq x_2 \geq \cdots \geq x_j \geq \cdots \geq x_J \tag{14}$$

  where $J = S \cdot T$ is the total number of samples.

- For each sample x_j in the order, calculate $\overline{\mathrm{FAR}}_2$ for $\delta_x = x_j$, by computing the cumulative series of the corresponding posterior probabilities $p_j^{(0)}$, and normalizing it by its last term ($\sum_j p_j^{(0)}$):

$$\overline{\mathrm{FAR}}_2\left(M, \{x_{s,t}\}, x_j\right) \quad = \quad \sum_{k=1}^{j} p_k^{(0)} \;/\; \sum_{k=1}^{J} p_k^{(0)} \tag{15}$$

- Select δ_x such that $\overline{\mathrm{FAR}}_2\left(\delta_x\right) = \mathrm{FAR_T}$ through linear interpolation.

Overall, the cost of this procedure is directly proportional to the amount of data $S \cdot T$, which itself can be reduced to a fixed, small number of samples (e.g. 100), as explained in Section B.

4.4 Experiments

Graphical results are given in Fig. 10 (dashed curves), and corresponding numerical averages are given in Tab. 2 for a practical range of small $\mathrm{FAR_T}$ values (up to 5%). The latter is the Root Mean Square (RMS) of ($\mathrm{FAR}/\mathrm{FAR_T} - 1$):

$$\sqrt{\left\langle \left(\frac{\mathrm{FAR}}{\mathrm{FAR_T}} - 1 \right)^2 \right\rangle_{\mathrm{FAR_T} < 0.5\%}} \tag{16}$$

This average metric was chosen in order to normalize results that have very different orders of magnitude (from 0.1% to 5%). Ideally it is equal to zero.

Two observations can be made:

- 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 fact that no condition-specific tuning is made in the model-based approaches, while in the "training" case, tuning was done on loudspeaker data.
- The "model+data" approach systematically improves over the "model only" approach.

Both points confirm previous expectations. It is important to bear in mind that all three approaches "training", "model only" and "model+data" have the exact same ROC curve (FRR as a function of FAR), since the decision process is the same: $x_{s,t} \gtrless \delta_x$.

Overall, although there is a major improvement over the "training" approach in terms of robustness across conditions, especially visible in Fig. 10d,e, we can see that the results are sometimes suboptimal (loudspeaker data).

The next section shows that the "model+data" approach can be applied to more complex models, thus bringing further improvement.

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. This section shows that the "model+data" approach presented in Section 4.3 can be applied to more complex multidimensional models.

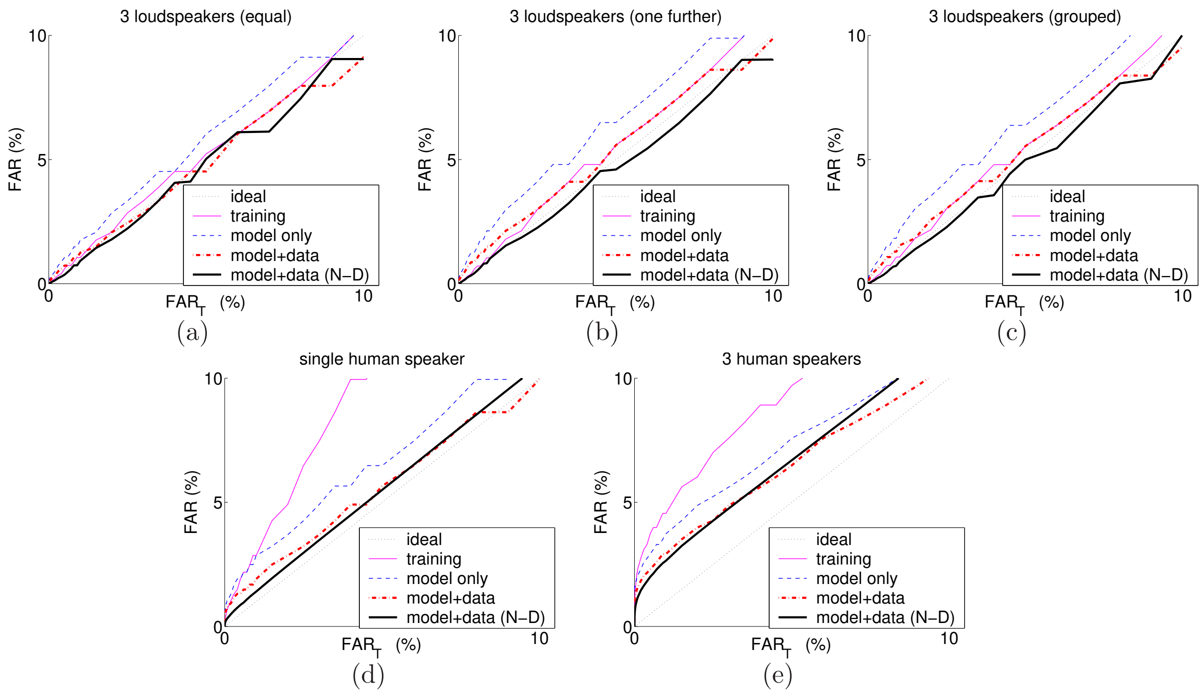


Figure 10: Threshold selection with and without training data, applied to loudspeaker recordings (see Fig. 4a,b,c) and human recordings (see Fig. 5d,e): comparison between desired target and measured False Alarm Rate. "training", "model only", "model+data" and "model+data (N-D)" correspond to Sections 3, 4.2, 4.3 and 5, respectively. In the "3 human speakers" case, the positive bias between FAR and FAR$_\mathrm{T}$ is due to body noises (breathing, stomps, shuffling paper) that could not be marked in the ground-truth, since their location are unkown.

	3 loudspeakers			1 human	3 humans
Recording	(a)	(b)	(c)	(d)	(e)
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 $(\mathrm{FAR}/\mathrm{FAR_T} - 1)$ (Eq. 16): the lower, the better. The best result for each recording is indicated in boldface.

5.1 Multidimensional model

On the microphone array detection task, as explained earlier in Section 2, we know that for a given time frame t, all numbers $x_{s,t}$ sum to a constant:

$$\sum_s x_{s,t} = N_{bins} \tag{17}$$

Overall, this motivates modelling all sectors $(x_{1,t} \cdots x_{s,t} \cdots x_{S,t})$ jointly, as described in details in Section C and briefly summarized here.

In particular, it can be expected that if all sectors are "silent", the frequency bins will be attributed to sectors somewhat randomly, so that a given sector will receive in average N_{bins}/S frequency bins. On the contrary, if at least one sector is active, it will capture more frequency bins than other sectors. Therefore, the remaining silent sectors will capture less than N_{bins}/S frequency bins in average. Intuitively, at least two probabilistic distributions should be used for "silent" sectors.

To summarize, the dependencies between the sectors are modelled through a frame state hidden variable. We assume independence between the sectors *conditioned* by the knowledge of the overall state of the time frame, as illustrated in Fig. 14. More precisely, a time frame can be "active" or "inactive", where an "active time frame" means that it contains at least one active sector. Thus, 3 probability density functions are used, one for each of 3 cases:

- Inactive frame, inactive sector: a Gamma distribution is used.
- Active frame, inactive sector: another Gamma distribution is used, with different parameters.
- Active frame, active sector: a Shifted Rice distribution is used to capture large values of Activeness, as in Section 4.

An example of fit of the 3 distributions (the two Gammas and the Shifted Rice) is depicted by Fig. 15b. The model used in experiments reported below is described in details in Section C, along with justifications and EM derivation.

5.2 Thresholding posteriors

However a threshold cannot be defined on multidimensional data.

We therefore propose to reuse the "model+data" approach presented in Section 4.3, by simply replacing the threshold on the 1-dimensional "activeness" feature:

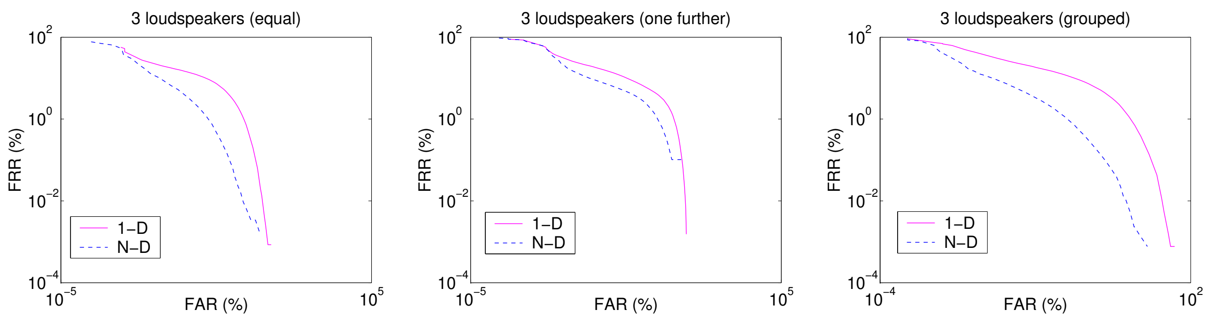
$$x_{s,t} \quad \gtrless \quad \delta_x \tag{18}$$

with a threshold on the estimate of the posterior probability of activity:

$$p\,(\,c_{s,t} = 1 \mid \{x_{1,t} \ldots x_{S,t}\}, M\,) \quad \gtrless \quad \delta_p \tag{19}$$

Thus, the exact same reasoning can be made as in Section 4.3, and the threshold on posteriors δ_p can be determined on the test data itself such that $\overline{\mathrm{FAR}}_2(M, \{x_{s,t}\}, \delta_p) = \mathrm{FAR}_\mathrm{T}$. 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.

Implementation: it is exactly similar to the 1-dimensional case (Section 4.3). Simply, the ordering of data samples is replaced with an ordering of posteriors. Thus, the cost of selecting a threshold δ_p is also directly proportional to the total number of samples $T \cdot S$.



(a) (b) (c)

Figure 11: ROC curves on the loudspeaker recordings of the 1-dimensional approaches (Eq. 18, approaches “training”, “model only”, “model+data”) and multidimensional approach (Eq. 19, approach “model+data (N-D)”).

5.3 Experiments

The results are shown in Fig. 10 and Tab. 2. 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 recording (a), results are similar to those of best one: “training”, which itself was tuned on part of (a). On recordings (b),(c),(d),(e) the multidimensional approach yields the best results of all approaches. Overall, this result is quite interesting given that the multidimensional approach does not use any training data.

However, Fig. 10 and Tab. 2 only involve the FAR prediction performance. For the sake of completeness, we also looked at the ROC curves. As explained earlier, the three 1-dimensional methods share the same ROC curve. On the contrary, the ROC curve of the multidimensional approach is different. In the case of loudspeaker recordings a systematic improvement is seen as compared to the 1-dimensional approaches (Fig. 11). ROC curves on human recordings are not reliable, as explained in Section 6.2.

6 Openings

This section provides insights about future extensions of the present work, from both theoretical and practical point of views.

6.1 Theory

In Section 4.3, it was proposed to approximate the true FAR:

$$\mathrm{FAR} \stackrel{\text{def}}{=} \frac{\text{Number of false alarms}}{\text{Number of inactive samples}} \tag{20}$$

by approximating the numerator and the denominator separately (Eq. 13):

$$\overline{\mathrm{FAR}}_2(M, \{x_{s,t}\}, \delta_x) \stackrel{\text{def}}{=} \frac{\mathbf{E}\left\{\text{Number of false alarms} \mid M, \{x_{s,t}\}, \delta_x\right\}}{\mathbf{E}\left\{\text{Number of inactive samples} \mid M, \{x_{s,t}\}, \delta_x\right\}}. \tag{21}$$

Ideally the whole ratio should be approximated at once:

$$\overline{\mathrm{FAR}}_3(M, \{x_{s,t}\}, \delta_x) \stackrel{\mathrm{def}}{=} \mathbf{E}\left\{ \mathrm{FAR} \mid M, \{x_{s,t}\}, \delta_x \right\} \tag{22}$$

$$= \mathbf{E}\left\{ \frac{\text{Number of false alarms}}{\text{Number of inactive samples}} \middle| M, \{x_{s,t}\}, \delta_x \right\}, \tag{23}$$

which is a possible direction for future work. This may apply to other metrics as well (FRR, HTER, precision, recall, etc.).

6.2 Practice: Preliminary Experiment with False Rejection Rate

Similarly to Section 4.2, a "model only" estimate of FRR can be proposed:

$$\overline{\mathrm{FRR}}_1(M, \delta_x) \stackrel{\mathrm{def}}{=} \int_0^{\delta_x} f_1(x) dx, \tag{24}$$

and similarly to Section 4.3, a "model+data" estimate can be proposed:

$$\overline{\mathrm{FRR}}_2(M, \{x_{s,t}\}, \delta_x) \stackrel{\mathrm{def}}{=} \sum_{\substack{s,t \\ x_{s,t} \leq \delta_x}} p_{s,t}^{(1)} \; / \; \sum_{s,t} p_{s,t}^{(1)} \tag{25}$$

Curves depicting FRR as a function of $\mathrm{FRR_T}$ are shown in Fig. 12. Note that the whole $[0\%, 100\%]$ interval is shown. Two observations can be made:

- On loudspeaker recordings (Figs. 12a,b,c), the "training" approach fails, while all the proposed model-based approaches provide a reasonable estimation of FRR. This is quite interesting, given that the "training" approach was tuned on part of the recording corresponding to Fig. 12a. A possible reason is that FRR is by definition linked to the distribution of "activity", which may be more variable than "inactivity" (e.g. between different spoken words), hence the "training" results are much worse than in the FAR case.
- On human recordings (Fig. 12d,e), for all approaches a large bias can be seen in the region of small $\mathrm{FRR_T}$. The reason is most likely an issue with ground-truthing: for each location, speech segments were marked as begin- and end-point, by listening to the signal. Each speech segment very often contains many short silences between words or syllables, and therefore possibly many frames where the selected "activeness" feature (Fig. 2c,d) is too low. This explains the large number of artificial false rejections. In other terms, the ground-truth of "activeness" is over-conservative for the human recordings and the FRR task. This artificially lifts up the FRR for conservative thresholds (low FAR). Thus, in the human case, it is not possible to judge or compare the FRR prediction curves.

Tab. 3 shows numerical results for loudspeaker recordings, on the $\mathrm{FRR_T} \in [0.1\%, 5\%]$ range. We can already see that:

- The "model+data" approach always performs best.
- Although not always the best, the multidimensional approach is the most robust ("maximum" column).

6.3 Extension to Multiple Classes

In this section we investigate whether the proposed threshold selection approach can be extended to a multi-class classification contect. A class Q_k can be selected from a set $\{Q_1, \cdots Q_k, \cdots Q_K\}$, from an observed data sample x and a model M, by selecting the Maximum A Posteriori (MAP):

$$\hat{k}(x, M) = \arg\max_k p(Q_k \mid x, M) \tag{26}$$

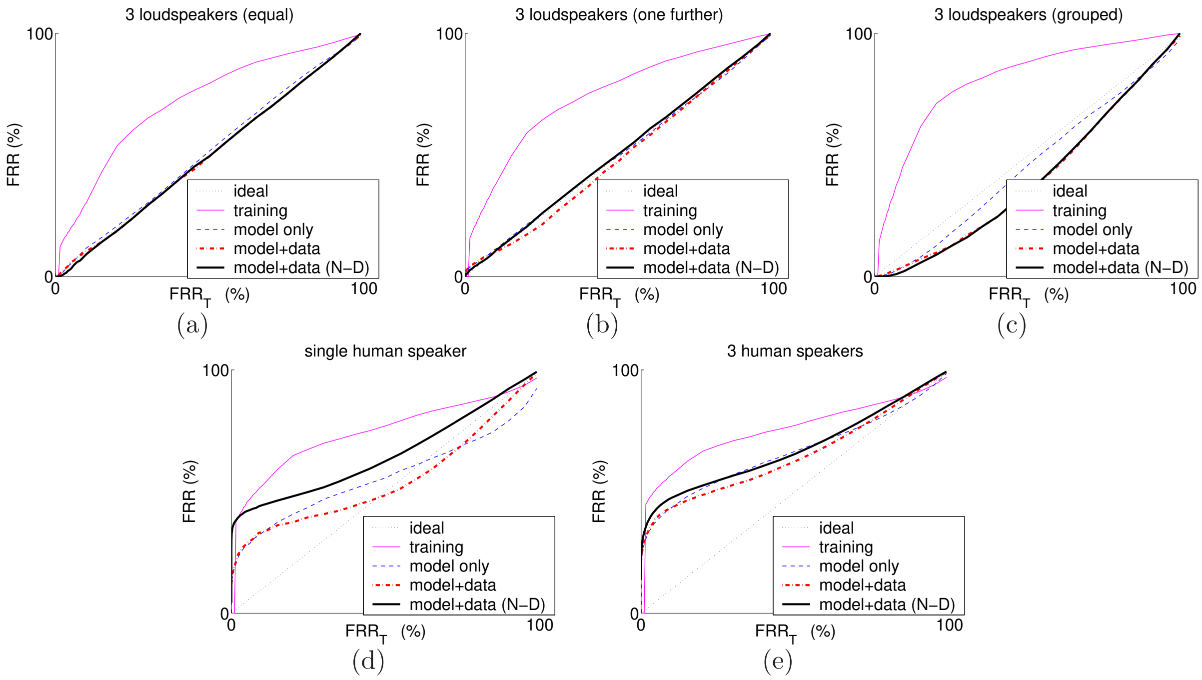


Figure 12: Threshold selection with and without training data, applied to loudspeaker recordings (see Fig. 4a,b,c) and human recordings (see Fig. 5d,e): comparison between desired target and measured False Rejection Rate. "training", "model only", "model+data" and "model+data (N-D)" correspond to Sections 3, 4.2, 4.3 and 5, respectively. Note that all $\mathrm{FRR_T}$ values from 0% to 100% are shown. Figs. d and e illustrate the ground-truthing issue with human data.

	3 loudspeakers			
Recording	(a)	(b)	(c)	Maximum
training	4.072	5.435	5.623	5.623
model only	0.400	2.356	0.846	2.356
model+data	**0.279**	2.759	**0.783**	2.759
model+data (N-D)	0.728	**0.743**	0.976	**0.976**

Table 3: RMS statistic over the interval $\mathrm{FRR_T} = [0.1\%, 5\%]$. This is the RMS of $(\mathrm{FRR}/\mathrm{FRR_T} - 1)$ (Eq. 16): the lower, the better. The best result for each recording is indicated in boldface. The rightmost column shows the maximum over all 3 recordings.

Intuitively, if all posteriors $p\left(Q_1 \mid x, M\right) \cdots p\left(Q_K \mid x, M\right)$ have comparable values, selecting the maximum is almost equivalent to a random choice. Thus, one may want to determine whether the system is *confident* in the decision $\hat{k}(x, M)$. For example, a speaker recognition system would ask the user to speak again, if the maximum posterior is below a threshold:

$$\begin{cases} \text{confident :} & p\left(Q_{\hat{k}} \mid x, M\right) > \delta_p \\ \text{not confident :} & p\left(Q_{\hat{k}} \mid x, M\right) \leq \delta_p \end{cases} \tag{27}$$

Then again, the question of selecting the threshold δ_p for a given objective criterion ($\mathrm{FAR} = \mathrm{FAR_T}$ or other) can be addressed. Indeed, Eq. 27 can be seen as a detection task, and by definition:

$$p\left(\text{correct decision} \mid x, M\right) \quad = \quad p\left(Q_{\hat{k}} \mid x, M\right) \tag{28}$$

From Eq. 28, the objective criterion can be estimated and the threshold δ_p can be selected, exactly as in Section 5.2.

7 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, but rather to rely on unsupervised fit of a model on test data. Even with an unsupervised approach, one question remains: how to select the detection threshold in an adequate manner? To that purpose, we proposed to determine the detection threshold from the unsupervised model itself. The proposed approach 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. In the proposed approach, the main novelty is a simple mechanism to compensate 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. We are currently using the proposed approach to investigate links between better voice activity detection and higher precision of speaker localization. The proposed approach could also be applied to other metrics such as False Rejection Rate (FRR), for example to detect end-points prior to automatic speech recognition, where it would be desirable to keep FRR to a small value. Finally, we showed in theory that the approach can be used in multiclass classification problems.

8 Acknowledgements

The author acknowledges 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 Bertrand Mesot for fruitful discussions and useful suggestions.

		Ground-truth	
		$c_{s,t} = 0$	$c_{s,t} = 1$
Detection	$\hat{c}_{s,t} = 0$	TN	FN
decision	$\hat{c}_{s,t} = 1$	FP	TP

Table 4: The four types of results. TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

A False Alarm Rate and False Rejection Rate

For each sector s and each time frame t:

- The ground-truth is $c_{s,t} = 0$ or 1.
- The decision taken by the system is $\hat{c}_{s,t} = 0$ or 1.

Thus, following [2], four types of cases happen, including correct classifications TP, TN and wrong classifications FP, FN, as defined in Tab. 4. The corresponding number of samples $\mathrm{N_{TP}}, \mathrm{N_{TN}}, \mathrm{N_{FP}}, \mathrm{N_{FN}}$ are counted over all sectors $s = 1 \cdots S$ and all time frames $t = 1 \cdots T$.

False Alarm Rate (FAR) is defined as follows:

$$\mathrm{FAR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{N_{FP}}}{\mathrm{N_{FP}} + \mathrm{N_{TN}}} \quad . \tag{29}$$

False Rejection Rate (FRR) is defined as follows:

$$\mathrm{FRR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{N_{FN}}}{\mathrm{N_{FN}} + \mathrm{N_{TP}}} \quad . \tag{30}$$

B 1-dimensional Model

This section decribes the 2-mixture probabilistic model $f(x) = w_0 \cdot f_0(x) + w_1 \cdot f_1(x)$ used to model the distribution of $x_{s,t}$ in Section 4, and derives the EM algorithm [4] for it.

B.1 Description

As described in details in [7] and briefly summarized in Section 2, the acoustic spectrum is divided into bins (narrowbands), and the activeness feature ($x_{s,t} \in \{0, 1, \cdots N_{bins}\}$) is the number of frequency bins where acoustic sources in sector s are dominant. In the case that there is a speech source in a sector of space, the corresponding value $x_{s,t}$ will be large because speech is wideband.

B.1.1 Inactivity: Dirac + Rice

In the case that there is no active coherent source at all in a particular frequency bin, the choice of the dominant sector is random, with equal probability $1/S$ for each sector. Hence, in the case that a sector s is completely inactive at time t, the number of bins attributed to this sector $x_{s,t}$ is a sum of realizations of such uniform random processes. It is therefore expected that $x_{s,t}$ follows a binomial distribution.

However, in real cases, even for a sector s that does not contain any active source, $x_{s,t}$ will not only result from purely random decisions, but it will also capture acoustic activity due to background noise (e.g. computer fan) and reverberations. We found visually, on some inactive sectors in the training data, that the Gamma distribution has a better fit than the binomial distribution. Gamma can therefore be used to model inactive sectors.

Moreover, the parameters of the Gamma used to model inactive sectors vary greatly between conditions, which can be roughly divided into two cases: whether at least one sector is active in a given frame or not. Therefore, ideally, two different distributions should be used. This is the issue addressed by the multidimensional model in Section 5.

In the case of a 1-dimensional model, we need to model a mixture of all inactive sectors (whether the frame is active or not) with a single distribution. We found visually, on real data, that the Rice distribution provides a better fit than the Gamma.

Finally, since some (rare) values of $x_{s,t}$ are zeroes, the inactive data is modelled by a mixture of a Dirac distribution at zero and a Rice distribution:

$$f_0(x) \stackrel{\text{def}}{=} \hat{p}\,(\,x_{s,t} \mid c_{s,t} = 0\,) \tag{31}$$

$$= w_0^D \cdot \delta_0(x) + w_0^R \cdot \mathcal{R}_{\sigma_0,V_0}(x), \tag{32}$$

where δ_0 is the Dirac distribution centered in 0 and the Rice distribution is defined as:

$$\mathcal{R}_{\sigma,V}(x) \stackrel{\text{def}}{=} \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2+|V|^2}{2\sigma^2}} I_0\left(\frac{x|V|}{\sigma^2}\right) & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \le 0 \end{cases} \tag{33}$$

where I_0 is a modified Bessel function of the first kind.

B.1.2 Activity: Shifted Rice

The distribution of $x_{s,t}$ for acoustic activity (especially speech) follows a distribution that is quite complex, varying over time and not known *a priori*. We therefore chose to use the Rice distribution for activity, because it is a flexible way to model a distribution of positive values. The shape of the Rice can vary between a pointy, Dirac-like distribution to a Gaussian-like distribution or a Rayleigh distribution.

Furthermore, it is reasonable to assume that in a small range of values around the background noise level, $x_{s,t}$ does not give any information to discriminate between activity and inactivity. Hence, similarly to [10], we only model values above the mean square value $\sqrt{2 \cdot \sigma_0^2 + V_0^2}$ [12] of the silence distribution $\mathcal{R}_{\sigma_0,V_0}$. Hence the "Shifted Rice" distribution for active sectors:

$$f_1(x) \stackrel{\text{def}}{=} \hat{p}\,(\,x_{s,t} \mid c_{s,t} = 1\,) \tag{34}$$

$$= \mathcal{R}_{\sigma_1,V_1}(x - \sqrt{2 \cdot \sigma_0^2 + V_0^2}). \tag{35}$$

Mixture of Inactivity and Activity

The likelihood $\hat{p}\,(\,x_{s,t} \mid c_{s,t}\,)$ is expressed as:

$$\hat{p}\,(\,x_{s,t} \mid c_{s,t}\,) \stackrel{\text{def}}{=} \delta(c_{s,t} - 0) \cdot f_0(x) + \delta(c_{s,t} - 1) \cdot f_1(x) \tag{36}$$

where $\delta(x) = 1_{x=0}$ is the Kronecker function, (not to be confused with the zero-centered Dirac distribution δ_0). $1_{\text{proposition}}$ is the indicator function: $1_{\text{proposition}} = 1$ if proposition is true, 0 otherwise

The priors are the weights w_0 and w_1:

$$w_0 \stackrel{\text{def}}{=} P(c_{s,t} = 0) \tag{37}$$

$$w_1 \stackrel{\text{def}}{=} P(c_{s,t} = 1) \tag{38}$$

Then the overall distribution of the data is expressed as:

$$f(x_{s,t}) \stackrel{\text{def}}{=} P(x_{s,t}) \tag{39}$$

$$= P(c_{s,t} = 0) \cdot \hat{p}\,(\,x_{s,t} \mid c_{s,t} = 0\,) + P(c_{s,t} = 1) \cdot \hat{p}\,(\,x_{s,t} \mid c_{s,t} = 1\,) \tag{40}$$

$$= w_0 \cdot f_0(x_{s,t}) + w_1 \cdot f_1(x_{s,t}). \tag{41}$$

The complete model M is defined by the list of its parameters:

$$M \stackrel{\text{def}}{=} \left(w_0, w_1, w_0^D, w_0^R, \sigma_0, V_0, \sigma_1, V_1\right). \tag{42}$$

B.2 EM Derivation

B.2.1 General

In the E-step, the posteriors are computed using Bayes rule, for example, for a given set of parameters M and a sample $x_{s,t}$:

$$p_{s,t}^{(0)}(M) \stackrel{\text{def}}{=} \hat{p}\left(c_{s,t}=0 \mid x_{s,t}, M\right) = \frac{w_0 \cdot f_0(x_{s,t})}{w_0 \cdot f_0(x_{s,t}) + w_1 \cdot f_1(x_{s,t})} \tag{43}$$

$$p_{s,t}^{(1)}(M) \stackrel{\text{def}}{=} \hat{p}\left(c_{s,t}=1 \mid x_{s,t}, M\right) = 1 - p_{s,t}^{(0)}(M) \tag{44}$$

Let us assume we have parameter values $M=\phi$, in the M-step we look for new values $\hat{\phi}$ that will increase the likelihood of the data given the model $\hat{p}\left(X \,\middle|\, M=\hat{\phi}\right)$. We will write X and C the random variables associated with the observed data $x \geq 0$ and the inactive/active state $c=0$ or 1. For random variables X and C, the s and t indices are irrelevant since all data is modelled on a single dimension, irrespective of space or time.

Let us write the KL divergence between the two distributions associated with current parameters ϕ and new parameters $\hat{\phi}$, for a given realization of X:

$$\begin{aligned} &KL\left[p\left(C \mid X, M=\phi\right), \hat{p}\left(C \,\middle|\, X, M=\hat{\phi}\right)\right] \stackrel{\text{def}}{=} \\ &\left\langle \log p\left(C \mid X, M=\phi\right)\right\rangle_{p(C \mid X, M=\phi)} - \left\langle \log \hat{p}\left(C \,\middle|\, X, M=\hat{\phi}\right)\right\rangle_{p(C \mid X, M=\phi)} \end{aligned} \tag{45}$$

where letters p and $\hat{p}$ represent the same function: they are only used to clarify where current parameters ϕ or new parameters $\hat{\phi}$ are used. In the following, ϕ and $\hat{\phi}$ are thus omitted whenever possible. The mean $\langle\cdot\rangle_.$ is calculated over all possible values of C (0 or 1). For example:

$$\left\langle \log \hat{p}\left(C \mid X\right)\right\rangle_{p(C \mid X)} = \sum_{c=0}^{1} p\left(C=c \mid X\right) \cdot \log \hat{p}\left(C=c \mid X\right) \tag{46}$$

Using Bayes rule to decompose the last term of Eq. 45, and using the fact that the KL divergence is always positive:

$$\log \hat{p}(X) \geq -\left\langle \log p\left(C \mid X\right)\right\rangle_{p(C \mid X)} + \left\langle \log \hat{p}(X, C)\right\rangle_{p(C \mid X)} \tag{47}$$

the first term in the right hand side does not depend on $\hat{\phi}$, therefore, one way to increase the likelihood $\log \hat{p}(X)$ is to find $\hat{\phi}$ that maximizes the second term $\langle \log \hat{p}\rangle_p$. The latter decomposes into:

$$\left\langle \log \hat{p}(X, C)\right\rangle_{p(C \mid X)} = \left\langle \log \hat{p}\left(X \mid C\right)\right\rangle_{p(C \mid X)} + \left\langle \log \hat{p}(C)\right\rangle_{p(C \mid X)} \tag{48}$$

To conclude, in the M-step our purpose is to find the parameter values $\hat{\phi}$ that maximizes the likelihood of the observed data:

$$\sum_{s,t} \log \hat{p}\left(X = x_{s,t}\right) \tag{49}$$

which, using Eqs. 47 and 48, can be done by maximizing $\Xi_1 + \Xi_2$, where:

$$\Xi_1 \stackrel{\text{def}}{=} \sum_{s,t} \left\langle \log \hat{p}\left(X = x_{s,t} \mid C\right)\right\rangle_{p(C \mid X = x_{s,t})} \tag{50}$$

$$\Xi_2 \stackrel{\text{def}}{=} \sum_{s,t} \left\langle \log \hat{p}(C)\right\rangle_{p(C \mid X = x_{s,t})} \tag{51}$$

B.2.2 Specific

Let us express both terms Ξ_1 and Ξ_2 as a function of the new parameters $\hat{\phi} = \left(\hat{w}_0, \hat{w}_1, \hat{w}_0^D, \hat{w}_0^R, \hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$.

$$\Xi_1 = \sum_{s,t} \sum_{c=0}^{1} p_{s,t}^{(c)} \cdot \log \hat{p}\left(X = x_{s,t} \,\middle|\, C = c, \hat{\phi}\right) \tag{52}$$

From Eqs. 32 and 35:

$$\Xi_1 = \sum_{s,t} \sum_{c=0}^{1} p_{s,t}^{(c)} \cdot \log f_c\left(x_{s,t}, \hat{\phi}\right) \tag{53}$$

$$\Xi_1 = \sum_{s,t} p_{s,t}^{(0)} \cdot \log\left(\hat{w}_0^D \cdot \delta_0(x_{s,t}) + \hat{w}_0^R \cdot \mathcal{R}_{\hat{\sigma}_0,\hat{V}_0}(x_{s,t})\right) \tag{54}$$

$$+ \sum_{s,t} p_{s,t}^{(1)} \cdot \log\left(\mathcal{R}_{\hat{\sigma}_1,\hat{V}_1}\left(x_{s,t} - \sqrt{2 \cdot \hat{\sigma}_0^2 + \hat{V}_0^2}\right)\right) \tag{55}$$

$$\Xi_1 = \sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(0)} \cdot \log \hat{w}_0^D \tag{56}$$

$$+ \sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(0)} \cdot \log \hat{w}_0^R \tag{57}$$

$$+ \sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(0)} \cdot \log \delta_0(x_{s,t}) \tag{58}$$

$$+ \sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(0)} \cdot \log \mathcal{R}_{\hat{\sigma}_0,\hat{V}_0}(x_{s,t}) \tag{59}$$

$$+ \sum_{\substack{s,t \\ x_{s,t}>\sqrt{2\cdot\hat{\sigma}_0^2+\hat{V}_0^2}}} p_{s,t}^{(1)}(\phi) \cdot \log \mathcal{R}_{\hat{\sigma}_1,\hat{V}_1}\left(x_{s,t} - \sqrt{2 \cdot \hat{\sigma}_0^2 + \hat{V}_0^2}\right) \tag{60}$$

The term in $\log \delta_0(\cdot)$ is not finite, but does not involve any parameter in $\hat{\phi}$. In the M-step, we therefore maximize the "partial likelihood", which is the sum of all other (finite) terms.

From Eqs. 37 and 38:

$$\Xi_2 = \sum_{s,t} \sum_{c=0}^{1} p_{s,t}^{(c)} \cdot \log \hat{p}(C = c) \tag{61}$$

$$\Xi_2 = \sum_{s,t} p_{s,t}^{(0)} \cdot \log \hat{w}_0 + \sum_{s,t} p_{s,t}^{(1)} \cdot \log \hat{w}_1 \tag{62}$$

Our goal is to find $\hat{\phi}$ that maximimizes $\Xi_1 + \Xi_2$. From Eqs. 56, 57, 59, 60 and 62, we can see that:

- The priors $\hat{w}_0$ and $\hat{w}_1$ only appear in Ξ_2 (Eq. 62).
- The weights $\hat{w}_0^D$ and $\hat{w}_0^R$ of the silence mixture only appear in Ξ_1, in Eqs. 56 and 57.
- The remaining parameters $\left(\hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$ are tied in a non-linear fashion through Eqs. 59 and 60. Finding their value can be done through joint, numerical optimization (e.g. simplex search, as in `fminsearch` in MATLAB) of the corresponding sum (59)+(60).

Since $\hat{w}_1 = 1 - \hat{w}_0$, we have:

$$\frac{\partial \Xi_2}{\partial \hat{w}_0} = \sum_{s,t} \left(\frac{1}{\hat{w}_0} \cdot p_{s,t}^{(0)} + \frac{1}{1-\hat{w}_0} \cdot p_{s,t}^{(1)} \right). \tag{63}$$

The new parameter $\hat{w}_0$ does not appear in Ξ_1, it is therefore the maximum of Ξ_2 with respect to w_0, which necessitates $\frac{\partial \Xi_2}{\partial w_0}(\hat{w}_0) = 0$, therefore:

$$\hat{w}_0 = \frac{\sum_{s,t} p_{s,t}^{(0)}}{\sum_{c=0}^{1} \sum_{s,t} p_{s,t}^{(c)}} = \frac{1}{T \cdot S} \sum_{s,t} p_{s,t}^{(0)} \tag{64}$$

and $\hat{w}_1 = 1 - \hat{w}_0$. This is the update of the priors of inactivity and activity.

Similarly, since $\hat{w}_0^R = 1 - \hat{w}_0^D$, from Eqs. 56 and 57 we have the maximum $\hat{w}_0^D$ at the zero $\frac{\partial \Xi_1}{\partial \hat{w}_0^D} = 0$, which yields:

$$\hat{w}_0^D = \frac{\sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(0)}}{\sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(0)} + \sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(0)}} \tag{65}$$

and $\hat{w}_0^R = 1 - \hat{w}_0^D$. This is the update of the “inactivity” mixture weights.

B.2.3 Implementation Details

EM Implementation: in practice, we observed that the possibly large amount of data $\{x_{s,t}\}$ can be conveniently reduced to a very small number of samples (e.g. 100) with approximately the same distribution. This is done by ordering the samples (from min to max) and picking 100 samples at regular intervals along the ordered list. This way the cost of each EM iteration is drastically reduced, and is independent of the amount of data (e.g. 20 minutes of recording are reduced to 100 samples).

An additional speedup can be obtained by replacing, in the M-step, the numerical optimization of $\left(\hat{\sigma}_0, \hat{V}_0, \hat{\sigma}_1, \hat{V}_1\right)$ with a moment-based update similar to the initialization described below. This way, the numerical optimization, which is itself a many-step process, is replaced with a direct, 1-step analytical update. Although this is an approximation, we observed in practice that after convergence of EM, the model parameters are almost the same as with the numerical optimization.

All results reported in the article for the 1-dimensional models use both simplifications. An example of data distribution fitted with the 1-dimensional model is depicted by Fig. 13b.

Automatic Initialization: the data $\{x_{s,t}\}$ is arbitrarily splitted between $N_{\text{low}}(\theta)$ low values and $N_{\text{high}}(\theta)$ high values, using a threshold θ. The non-zero low values are used to initialize the “inactivity” Rice distribution using the analytical moment-based approximation from [12]. This is done by first computing the mean G_a and standard deviation G_v:

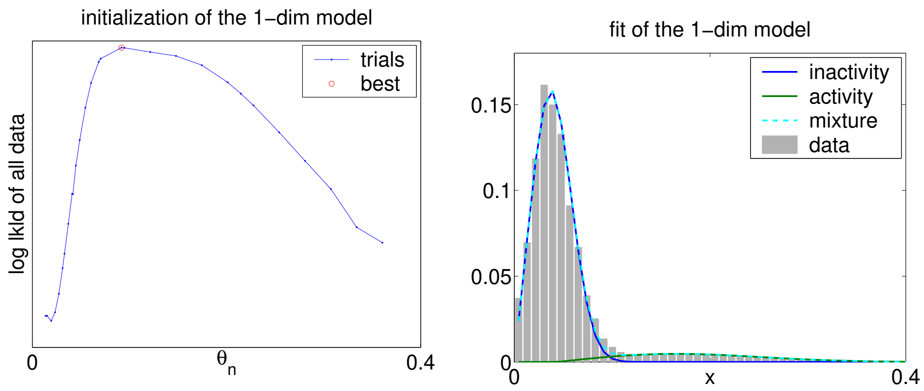
$$G_a = \frac{1}{N_{\text{low}}} \sum_{\substack{s,t \\ x_{s,t}<\theta}} x_{s,t} \tag{66}$$

$$G_v = \left(\frac{1}{N_{\text{low}}} \sum_{\substack{s,t \\ x_{s,t}<\theta}} (x_{s,t} - G_a)^2 \right)^{\frac{1}{2}} \tag{67}$$

and the parameters of the “inactivity” Rice distribution are initialized as follows:

$$V_0^{(\text{init})} \leftarrow \left[\max\left(0, G_a^2 - G_v^2\right) \right]^{\frac{1}{4}} \tag{68}$$

$$\sigma_0^{(\text{init})} \leftarrow \left[\frac{1}{2} \max\left(0, G_a - \left(V_0^{(\text{init})}\right)^2\right) \right]^{\frac{1}{2}} \tag{69}$$



(a) (b)

Figure 13: Fit of the 2-mixture model described in Section B: (a) automatic initialization, (b) final model after convergence of EM.

The "activity" Shifted Rice distribution is initialized similarly using data above

$$\max\left(\theta,\ \sqrt{2\cdot\left(\sigma_0^{(\text{init})}\right)^2+\left(V_0^{(\text{init})}\right)^2}\right). \tag{70}$$

Finally, the mixture weights w_0^D and w_0^R are initialized by counting the number of zero samples within the N_{low} low values. Priors w_0 and w_1 are initialized as follows:

$$w_0^{(\text{init})} \leftarrow \max\left(0.1, \min\left(0.9, \frac{N_{\text{low}}}{N_{\text{low}}+N_{\text{high}}}\right)\right) \tag{71}$$

$$w_1^{(\text{init})} \leftarrow 1-w_0^{(\text{init})} \tag{72}$$

where the restriction to the $[0.1, 0.9]$ interval avoids a "wrong" local maxima such as $w_0^{(init)} = 0$.

In order to have a fully automatic initialization process, a series of thresholds $\theta_1 \cdots \theta_{N_\theta}$ are derived from the data itself (e.g. $N_\theta = 30$: 15 equal-interval percentiles and 15 equal intervals between minimum and maximum). For each threshold θ_n, the moment-based initialization is done as explained above and the likelihood of the whole data is computed. The initialization yielding the maximum likelihood is selected, as depicted by Fig. 13a. This way, we avoid starting from a "wrong" local maxima (e.g. the "inactive" component capturing all data and the "active" component capturing none, or vice-versa).

C Multidimensional Model

This section describes a multidimensional model that models activeness for all sectors $(x_{1,t} \cdots x_{S,t})$ jointly, at a given time frame t. It is used in Section 5. In many places, reasoning made in details in the 1-dimensional case (Section B) is reused here in a brief form.

C.1 Description

As explained in Section 2, for a given time frame t, all values sum to a constant: $\sum_s x_{s,t} = N_{bins}$. This knowledge is enough to expect two cases:

- At a given time frame t, there is no activity. Then, as explained in B.1, the N_{bins} bins of the frequency spectrum are attributed to the various sectors in a uniformly random fashion. It

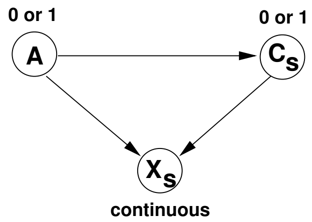


Figure 14: Graphical model for the independence assumptions used in the multidimensional model. A is the frame state (inactive or active) and C_s is the state of a given sector s (inactive or active). X_s is the observed data for sector s. On an active frame ($A = 1$) at least one sector is active ($C_s = 1$).

is therefore expected that $x_{s,t}$ will be N_{bins}/S in average. As mentioned in B.1, the Gamma distribution fits well (visual trials on real data), which is defined as:

$$\mathcal{G}_{\gamma,\beta}(x) \stackrel{\text{def}}{=} \frac{x^{\gamma-1} \cdot e^{-\frac{x}{\beta}}}{\beta^{\gamma} \cdot \Gamma(\gamma)} \tag{73}$$

where $\gamma > 0$, $\beta > 0$ and Γ is the gamma function:

$$\Gamma(\gamma) \stackrel{\text{def}}{=} \int_0^{+\infty} t^{\gamma-1} e^{-t} dt \tag{74}$$

In this case, we could expect the average $\gamma\beta = \frac{N_{bins}}{S}$.

- At a given time frame t, at least one sector contains at least one active wideband source (e.g. speech source). In such a case, the $x_{s,t}$ values corresponding to active sector(s) will be larger than the average N_{bins}/S, thus leaving less frequency bins to be randomly attributed to inactive sectors. We propose to model those values with a (Gamma + Shifted Rice) mixture, similarly to Section B.1. For the Gamma distribution, we could expect that $\gamma\beta < \frac{N_{bins}}{S}$.

Similarly to C, let us define the binary random variable "frame state" $A = 0$ or 1, which indicates whether at least one sector is active in a given time frame. A realization a_t of A is defined by:

$$a_t \stackrel{\text{def}}{=} \max_{1 \le s \le S} c_{s,t} \tag{75}$$

Next, we define the random variables $X_1 \cdots X_S$ associated with activeness values of all sectors, for a given time frame. We then need to express $p\,(\, X_1 \cdots X_S \mid A, M\,)$.

We assume that the knowledge of the frame state A is enough (the two cases mentionned above), and that further interdependences between the various values

$$X_{1:S} \stackrel{\text{def}}{=} (X_1 \cdots X_S) \tag{76}$$

need not to be modelled. This amounts to a simplifying *conditional* independence assumption:

$$p\,(\, X_{1:S} \mid A, M\,) \approx \prod_{s=1}^{S} p\,(\, X_s \mid A, M\,) \tag{77}$$

as illustrated in Fig. 14.

With the same justification, we also make the following *conditional* independence assumption. For a given sector s:

$$p\,(\, C_s \mid X_{1:S}, A\,) \approx p\,(\, C_s \mid X_s, A\,) \tag{78}$$

These two assumptions are used in the EM derivation (Section C.2).

C.1.1 Inactive sector: Dirac + Gamma

Similarly to B.1.1, we model inactivity of a sector with a (Dirac + Gamma) mixture. Two such mixtures g_{00} and g_{01} are defined, depending on the state of the frame: inactive frame $A = 0$ or active frame $A = 1$:

$$p\,(\,X_s \mid C_s = 0, A = 0, M\,) \quad \sim \quad g_{00}\,(x) \tag{79}$$
$$p\,(\,X_s \mid C_s = 0, A = 1, M\,) \quad \sim \quad g_{01}\,(x) \tag{80}$$

where:

$$g_{00}\,(x) \quad \stackrel{\text{def}}{=} \quad v_{00}^D \cdot \delta_0(x) + v_{00}^G \cdot \mathcal{G}_{\gamma_{00},\beta_{00}}\,(x) \tag{81}$$
$$g_{01}\,(x) \quad \stackrel{\text{def}}{=} \quad v_{01}^D \cdot \delta_0(x) + v_{01}^G \cdot \mathcal{G}_{\gamma_{01},\beta_{01}}\,(x) \tag{82}$$

We further constrain $\gamma_{00} > 1$ and $\gamma_{01} > 1$ so that $\mathcal{G}_{\gamma_{00},\beta_{00}}(0) = 0$ and $\mathcal{G}_{\gamma_{01},\beta_{01}}(0) = 0$. This way, zero values and strictly positive values are separately modelled by the Dirac and Gamma functions. (v_{00}^D, v_{00}^G) and (v_{01}^D, v_{01}^G) are the weights the g_{00} and g_{01} mixtures, respectively.

C.1.2 Active sector: Shifted Rice

Similarly to Section B.1.2, we model an active sector with a shifted Rice distribution, where the shift is equal to the first moment $\gamma_{01}\beta_{01}$ of the Gamma distribution $\mathcal{G}_{01}$:

$$p\,(\,X_s \mid C_s = 1, A = 1, M\,) \quad \sim \quad g_{11}\,(x) \tag{83}$$

where:

$$g_{11}\,(x) \quad \stackrel{\text{def}}{=} \quad \mathcal{R}_{\sigma_{11},V_{11}}(x - \gamma_{01}\beta_{01}). \tag{84}$$

C.1.3 Complete model

Let v_0 and v_1 denote the frame-level priors ($v_0 + v_1 = 1$):

$$v_0 \quad \stackrel{\text{def}}{=} \quad p\,(\,A = 0 \mid M\,) \tag{85}$$
$$v_1 \quad \stackrel{\text{def}}{=} \quad p\,(\,A = 1 \mid M\,)\,. \tag{86}$$

Let v_{01} and v_{11} denote the sector-level conditional priors ($v_{01} + v_{11} = 1$), given that a frame is active:

$$v_{01} \quad \stackrel{\text{def}}{=} \quad p\,(\,C_s = 0 \mid A = 1, M\,) \tag{87}$$
$$v_{11} \quad \stackrel{\text{def}}{=} \quad p\,(\,C_s = 1 \mid A = 1, M\,)\,. \tag{88}$$

The complete model is defined by the parameters:

$$M \quad \stackrel{\text{def}}{=} \quad \left(v_0, v_1, v_{01}, v_{11}, v_{00}^D, v_{00}^G, \gamma_{00}, \beta_{00}, v_{01}^D, v_{01}^G, \gamma_{01}, \beta_{01}, \sigma_{11}, V_{11}\right). \tag{89}$$

From Eqs. 77, 81, 82, 84, 85, 86, 87, 88, the complete model can be written as follows.
The priors of frame state:

$$p\,(A) \quad = \quad v_0^{\delta(A-0)} \cdot v_1^{\delta(A-1)} \tag{90}$$

where δ is the Kronecker function (not to be confused with the Dirac distribution δ_0).
The priors of sector state, for one sector s:

$$p\,(\,C_s \mid A\,) \quad = \quad \delta\,(A - 0) \cdot \delta\,(C_s - 0) + \delta\,(A - 1) \cdot v_{01}^{\delta(C_s - 0)} \cdot v_{11}^{\delta(C_s - 1)} \tag{91}$$

The likelihood of the data for *one* sector s, given the frame and sector states:

$$\begin{aligned} p\left(X_s \mid A, C_s\right) &= \delta\left(A-0\right)\delta\left(C_s-0\right)\cdot\left[v_{00}^D\cdot\delta_0\left(X_s\right)+v_{00}^G\cdot\mathcal{G}_{\gamma_{00},\beta_{00}}\left(X_s\right)\right] \\ &+ \delta\left(A-1\right)\delta\left(C_s-0\right)\cdot\left[v_{01}^D\cdot\delta_0\left(X_s\right)+v_{01}^G\cdot\mathcal{G}_{\gamma_{01},\beta_{01}}\left(X_s\right)\right] \\ &+ \delta\left(A-1\right)\delta\left(C_s-1\right)\cdot\left[\mathcal{R}_{\sigma_{11},V_{11}}\left(X_s-\gamma_{01}\beta_{01}\right)\right] \end{aligned} \tag{92}$$

C.2 EM Derivation

C.2.1 E step

In the E-step, we need to estimate posteriors of the $(A=1)$ and $(A=1, C_s=1)$ events. Posteriors of other events $(A=0)$ and $(A=1, C_s=0)$ are their respective 1-complements.

The first one is directly obtained from Bayes rule:

$$p\left(A=1 \mid X_{1:S}, M\right) = \frac{p\left(X_{1:S} \mid A=1, M\right)\cdot v_1}{p\left(X_{1:S} \mid A=0, M\right)\cdot v_0 + p\left(X_{1:S} \mid A=1, M\right)\cdot v_1} \tag{93}$$

and each of the 3 terms $p\left(X_{1:S} \mid A, M\right)$ expands as a product of individual likelihoods, given by Eq. 77. Note that in the case of a zero value $X_s=0$, an (indefinite) Dirac term will appear in all 3 terms, hence it simplifies out and only the corresponding (finite) Dirac weights remain.

The second one can be obtained from the following decomposition:

$$p\left(A, C_s \mid X_{1:S}, M\right) = p\left(C_s \mid A, X_{1:S}, M\right)\cdot p\left(A \mid X_{1:S}, M\right) \tag{94}$$

which, using the assumption made in Eq. 78, becomes:

$$p\left(A, C_s \mid X_{1:S}, M\right) = p\left(C_s \mid A, X_s, M\right)\cdot p\left(A \mid X_{1:S}, M\right) \tag{95}$$

The second term is given by Eq. 93, and the first one develops into:

$$\begin{aligned} p\left(C_s \mid X_s, A, M\right) &= \frac{p\left(X_s, C_s, A \mid M\right)}{p\left(X_s, A \mid M\right)} && (96) \\ &= \frac{p\left(X_s \mid C_s, A, M\right)\cdot p\left(C_s \mid A, M\right)\cdot p\left(A \mid M\right)}{p\left(X_s, A \mid M\right)} && (97) \\ &= \frac{p\left(X_s \mid C_s, A, M\right)\cdot p\left(C_s \mid A, M\right)}{p\left(X_s \mid A, M\right)} && (98) \end{aligned}$$

This decomposition is valid for both $(A=1, C_s=0)$ and $(A=1, C_s=1)$ events, hence:

$$\begin{aligned} p\left(C_s=1 \mid A=1, X_s, M\right) &= \frac{p\left(X_s \mid C_s=1, A=1, M\right)\cdot p\left(C_s=1 \mid A=1, M\right)}{\sum_{c=0}^{1} p\left(X_s \mid C_s=c, A=1, M\right)\cdot p\left(C_s=c \mid A=1, M\right)} && (99) \\ p\left(C_s=1 \mid A=1, X_s=x, M\right) &= \frac{g_{11}\left(x\right)\cdot v_{11}}{g_{01}\left(x\right)\cdot v_{01} + g_{11}\left(x\right)\cdot v_{11}} && (100) \end{aligned}$$

C.2.2 M step

As for the M-step, the KL divergence

$$KL\left[p\left(C_s, A \mid X_{1:S}, M=\phi\right), \hat{p}\left(C_s, A \,\middle|\, X_{1:S}, M=\hat{\phi}\right)\right] \tag{101}$$

can be written similarly to Eq. 45, where ϕ and $\hat{\phi}$ are the current parameters and new parameters, respectively. Similarly to Section B.2.1, a lower bound on the likelihood of the observed data can be found using the fact that the KL divergence is always positive:

$$\sum_t \log\hat{p}\left(X_{1:S}=\mathbf{x}_t\right) \geq \sum_t \left\langle \log\hat{p}\left(X_{1:S}=\mathbf{x}_t, C_{1:S}, A\right)\right\rangle_{p\left(C_{1:S}, A \mid X_{1:S}=\mathbf{x}_t\right)} \tag{102}$$

where $\mathbf{x}_t = (x_{1,t} \cdots x_{S,t})^{\mathrm{T}}$ is the vector of activeness values for all sectors, in the time frame t, and $C_{1:S} = (C_1 \cdots C_S)^{\mathrm{T}}$ is the corresponding vector of sector states (zeroes and ones).

From the decomposition:

$$\hat{p}\,(X_{1:S}, C_{1:S}, A) \quad = \quad \hat{p}\,(\,X_{1:S} \mid C_{1:S}, A\,) \cdot \hat{p}\,(\,C_{1:S} \mid A\,) \cdot \hat{p}\,(A)\,, \tag{103}$$

the right hand side (RHS) in Eq. 102 can be decomposed into a sum of 3 terms $\Psi_1 + \Psi_2 + \Psi_3$, where:

$$\Psi_1 \stackrel{\text{def}}{=} \sum_t \langle \log \hat{p}\,(\,X_{1:S} = \mathbf{x}_t \mid C_{1:S}, A\,)\rangle_{p(\,C_{1:S}, A \mid\, X_{1:S} = \mathbf{x}_t\,)} \tag{104}$$

$$\Psi_2 \stackrel{\text{def}}{=} \sum_t \langle \log \hat{p}\,(\,C_{1:S} \mid A\,)\rangle_{p(\,C_{1:S}, A \mid\, X_{1:S} = \mathbf{x}_t\,)} \tag{105}$$

$$\Psi_3 \stackrel{\text{def}}{=} \sum_t \langle \log \hat{p}\,(A)\rangle_{p(\,C_{1:S}, A \mid\, X_{1:S} = \mathbf{x}_t\,)} \tag{106}$$

The aim of the M-step is to find new parameter values $\hat{\phi}$ that will maximize the sum $\Psi_1 + \Psi_2 + \Psi_2$, in order to increase the overall likelihood of the observed data. From an independence assumption between the sectors, conditioned by A, similarly to Eq. 77:

$$\Psi_1 \quad = \quad \sum_{s,t} \langle \log \hat{p}\,(\,X_s = x_{s,t} \mid C_s, A\,)\rangle_{p\left(\,\underline{C_{1:S}},\; A \mid\, X_{1:S} = \mathbf{x}_t\,\right)} \tag{107}$$

which can be shown to be equal to:

$$\Psi_1 \quad = \quad \sum_{s,t} \langle \log \hat{p}\,(\,X_s = x_{s,t} \mid C_s, A\,)\rangle_{p\left(\,\underline{C_s},\; A \mid\, X_{1:S} = \mathbf{x}_t\,\right)}. \tag{108}$$

Under similar independence assumptions between the sectors, conditioned by A:

$$\Psi_2 \quad = \quad \sum_{s,t} \langle \log \hat{p}\,(\,C_s \mid A\,)\rangle_{p\left(\,\underline{C_s},\; A \mid\, X_{1:S} = \mathbf{x}_t\,\right)}. \tag{109}$$

Finally, since A does not depend on $C_{1:S}$:

$$\Psi_3 \quad = \quad \sum_t \langle \log \hat{p}\,(A)\rangle_{p(\,\underline{A} \mid\, X_{1:S} = \mathbf{x}_t\,)}. \tag{110}$$

Considering the definition of the model (Eqs. 90 to 92), it can be shown that:

- The priors v_0 and v_1 only appear in Ψ_3, under a form similar to w_0 and w_1 in Ξ_2 (Section B.2.2).
- The conditional priors v_{01} and v_{11} only appear in Ψ_2, under a form similar to w_0 and w_1 in Ξ_1 (Section B.2.2).
- The mixture weights $\left(v_{00}^D, v_{00}^G\right), \left(v_{01}^D, v_{01}^G\right)$ only appear in Ψ_1, both of them under a form similar to $\left(w_0^D, w_0^R\right)$ in Ξ_1 (Section B.2.2).
- The parameters γ_{00}, β_{00} appear only in Ψ_1, not tied to any other parameter.
- The parameters $\gamma_{01}, \beta_{01}, \sigma_{11}, V_{11}$ appear only in Ψ_1 and are tied in a non-linear fashion, similarly to the Rice and Shifted Rice in Section B.2.2.

Reasoning similar to Section B.2.2 leads to the following update equations.

For the frame-level priors:

$$\hat{v}_0 = \frac{\sum_{s,t} p_{s,t}^{(0)}}{\sum_{s,t} p_{s,t}^{(0)} + \sum_{s,t} p_{s,t}^{(1)}} \tag{111}$$

$$= \frac{1}{T \cdot S} \sum_{s,t} p_{s,t}^{(0)} \tag{112}$$

and $\hat{v}_1 = 1 - \hat{v}_0$.

For the sector-level conditional priors:

$$\hat{v}_{01} = \frac{\sum_{s,t} p_{s,t}^{(10)}}{\sum_{s,t} p_{s,t}^{(10)} + \sum_{s,t} p_{s,t}^{(11)}} \tag{113}$$

and $\hat{v}_{11} = 1 - \hat{v}_{01}$.

For the "inactive frame, inactive sector" (Dirac + Gamma) mixture weights:

$$\hat{v}_{00}^D = \frac{\sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(00)}}{\sum_{\substack{s,t \\ x_{s,t}=0}} p_{s,t}^{(00)} + \sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(00)}} \tag{114}$$

and $\hat{v}_{00}^G = 1 - \hat{v}_{00}^D$. Replacing "00" with "10" gives the update equations for $\hat{v}_{01}^D$ and $\hat{v}_{01}^G$.

Parameters γ_{00} and β_{00} are updated within the $\{\gamma_{00} > 1, \beta_{00} > 0\}$ space through numerical optimization, by maximizing the following sum (e.g. using the simplex method):

$$\sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(00)} \cdot \log \mathcal{G}_{\gamma_{00},\beta_{00}}(x_{s,t}) \tag{115}$$

Parameters $\gamma_{01}, \beta_{01}, \sigma_{11}$ and V_{11} are updated within the $\{\gamma_{01} > 1, \beta_{01} > 0, \sigma_{11} > 0, V_{11} \geq 0\}$ space, through numerical optimization, by maximizing the following sum:

$$\sum_{\substack{s,t \\ x_{s,t}>0}} p_{s,t}^{(10)} \cdot \log \mathcal{G}_{\gamma_{01},\beta_{01}}(x_{s,t}) \quad + \sum_{\substack{s,t \\ x_{s,t}>\gamma_{01}\beta_{01}}} p_{s,t}^{(11)} \cdot \log \mathcal{R}_{\sigma_{11},V_{11}}(x_{s,t} - \gamma_{01}\beta_{01}) \tag{116}$$

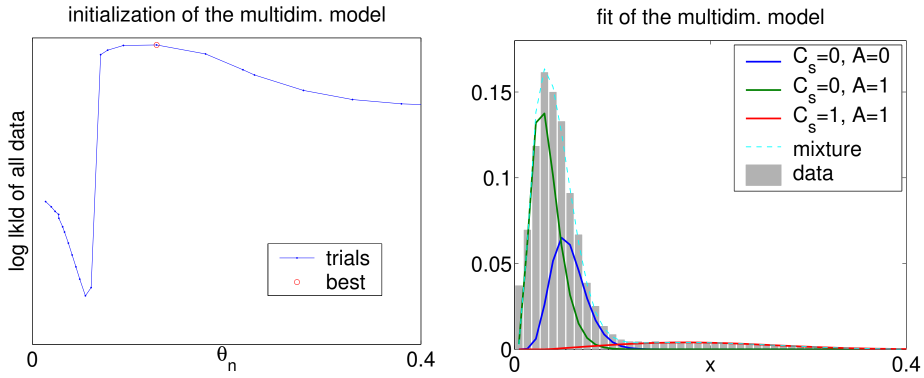
C.2.3 Implementation Details

EM Implementation: similarly to Section B.2.3, the possibly large data (e.g. 70000 frames for 20 minutes) is reduced to a fixed, small number of frames (e.g. 1000), by first ordering the frames $\{\mathbf{x}_t\}$ and then picking 1000 frames at equal intervals along the ordered list This necessitates the definition of an order between vectors $\{\mathbf{x}_t\}$. We are trying to keep approximately the same distribution of frames as in the original data, from the very inactive frames to the very active frames. Thus, data reduction is done by ordering frames $t = 1 \cdots T$ based on their maximum value $\max_s(x_{s,t})$. This way the computational cost of one EM iteration is drastically reduced, and it is independent of the size of the data.

As for the M-step, contrary to the 1-dimensional case, we found in practice that using the *exact* (likelihood increase is guaranteed), numerical optimization of the tied parameters, or using a moment-based approximation produced quite different results.

Overall the above-described data reduction and the numerical optimization are used for all multi-dimensional results reported in this paper.

Initialization: an approach similar to Section B.2.3 is used. *For the initialization only*, all data is stacked into a 1-dimensional histogram, and considered as a mixture of a Gamma with parameters $\gamma^{(\text{init})}, \beta^{(\text{init})}$ and a Shifted Rice with shift $\gamma^{(\text{init})}\beta^{(\text{init})}$. As in Section B.2.3, moment-based



(a) (b)

Figure 15: Fit of the multidimensional model described in Section C: (a) automatic initialization, (b) final distributions $\mathcal{G}_{00}, \mathcal{G}_{01}, \mathcal{R}_{11}$ and g after convergence of EM.

methods are used to initialize the Gamma, then the Shifted Rice, within an automatic, multiple initialization approach.

Finally both $\mathcal{G}_{00}$ and $\mathcal{G}_{01}$ are initialized with the same parameters $\gamma^{(\mathrm{init})}, \beta^{(\mathrm{init})}$.

Figs. 15a and 15b respectively depict an example of automatic initialization, and the final distributions after convergence of EM.

References

[1] J. Sklansky and N. 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 Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[5] W. Eadie, D. Drijard, and F. 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] S. Roweis, "Factorial Models and Refiltering for Speech Separation and Denoising," in *Proc. Eurospeech*, 2003.

[9] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proceedings of the 2004 MLMI Workshop, S. Bengio and H. Bourlard Eds, Springer Verlag*, 2005.

[10] G. Lathoud, M. Magimai.-Doss, B. Mesot, and H. Bourlard, "Unsupervised Spectral Subtraction for Noise-Robust ASR," in *Proceedings of the IEEE ASRU 2005 Workshop*, Cancun, Mexico, December 2005.

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

[12] L. Greenstein, D. Michelson, and V. Erceg, "Moment-method estimation of the ricean K-factor," *IEEE Communications Letters*, vol. 3, no. 6, June 1999.