



HAL
open science

Real-time Audio Classification based on Mixture Models

Maxime Baelde, Christophe Biernacki, Raphaël Greff

► **To cite this version:**

Maxime Baelde, Christophe Biernacki, Raphaël Greff. Real-time Audio Classification based on Mixture Models. The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017), Mar 2017, La Nouvelle-Orléans, United States. . hal-01481934

HAL Id: hal-01481934

<https://hal.science/hal-01481934v1>

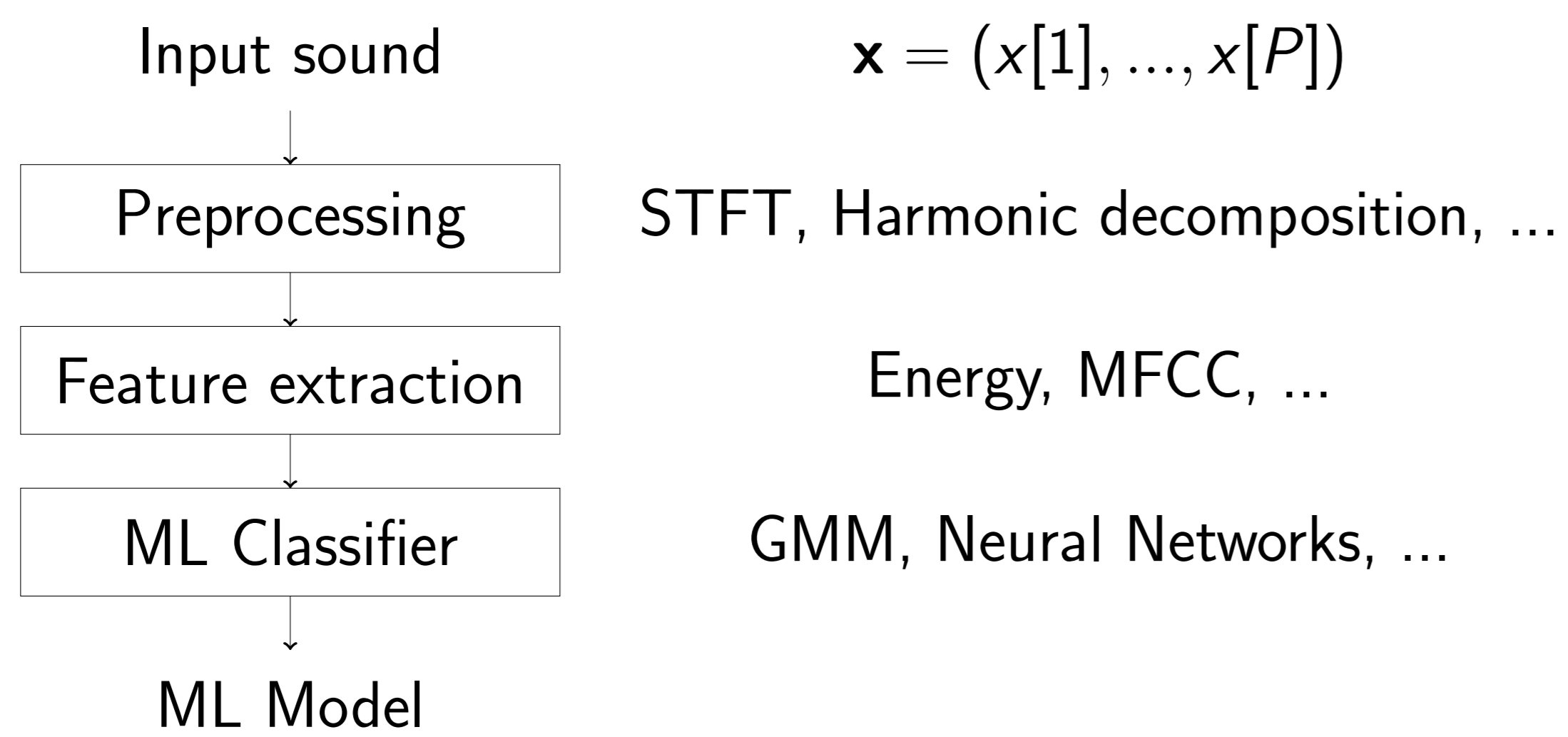
Submitted on 3 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

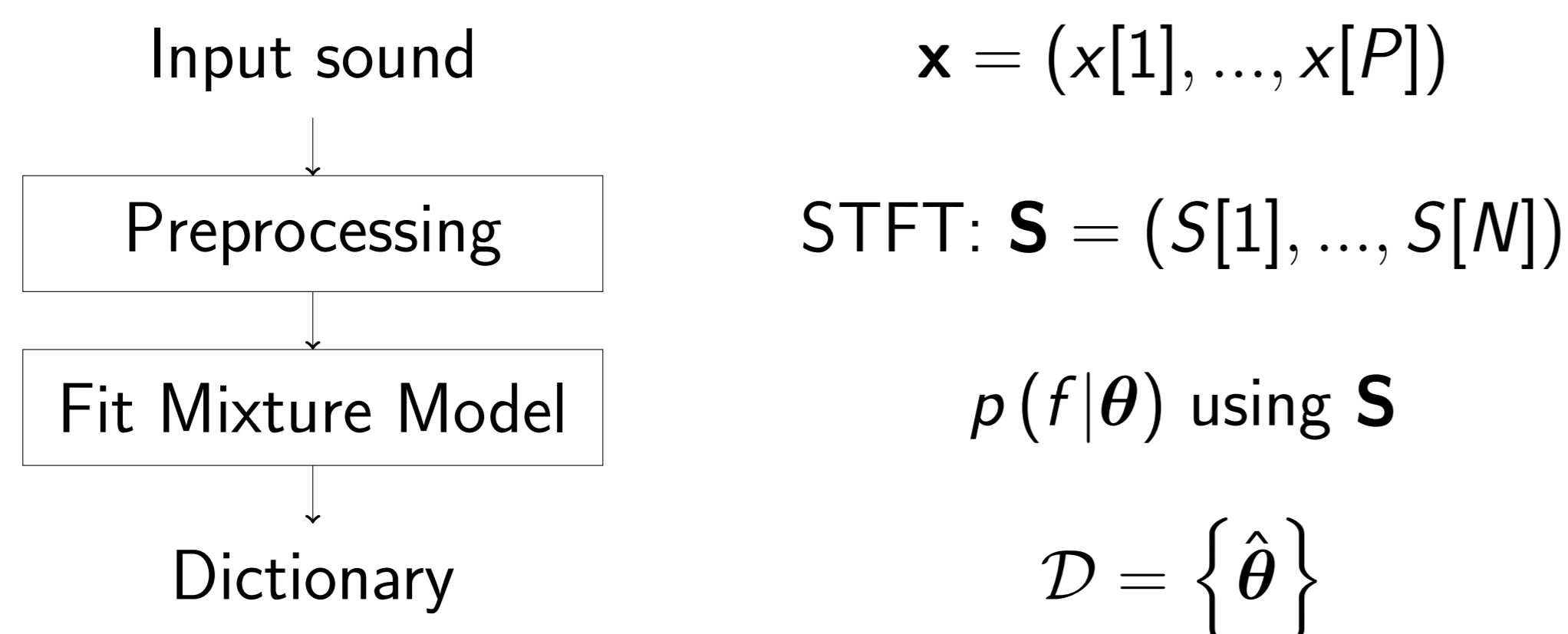
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1. INTRODUCTION

Standard Machine Learning (ML) approach for audio classification:

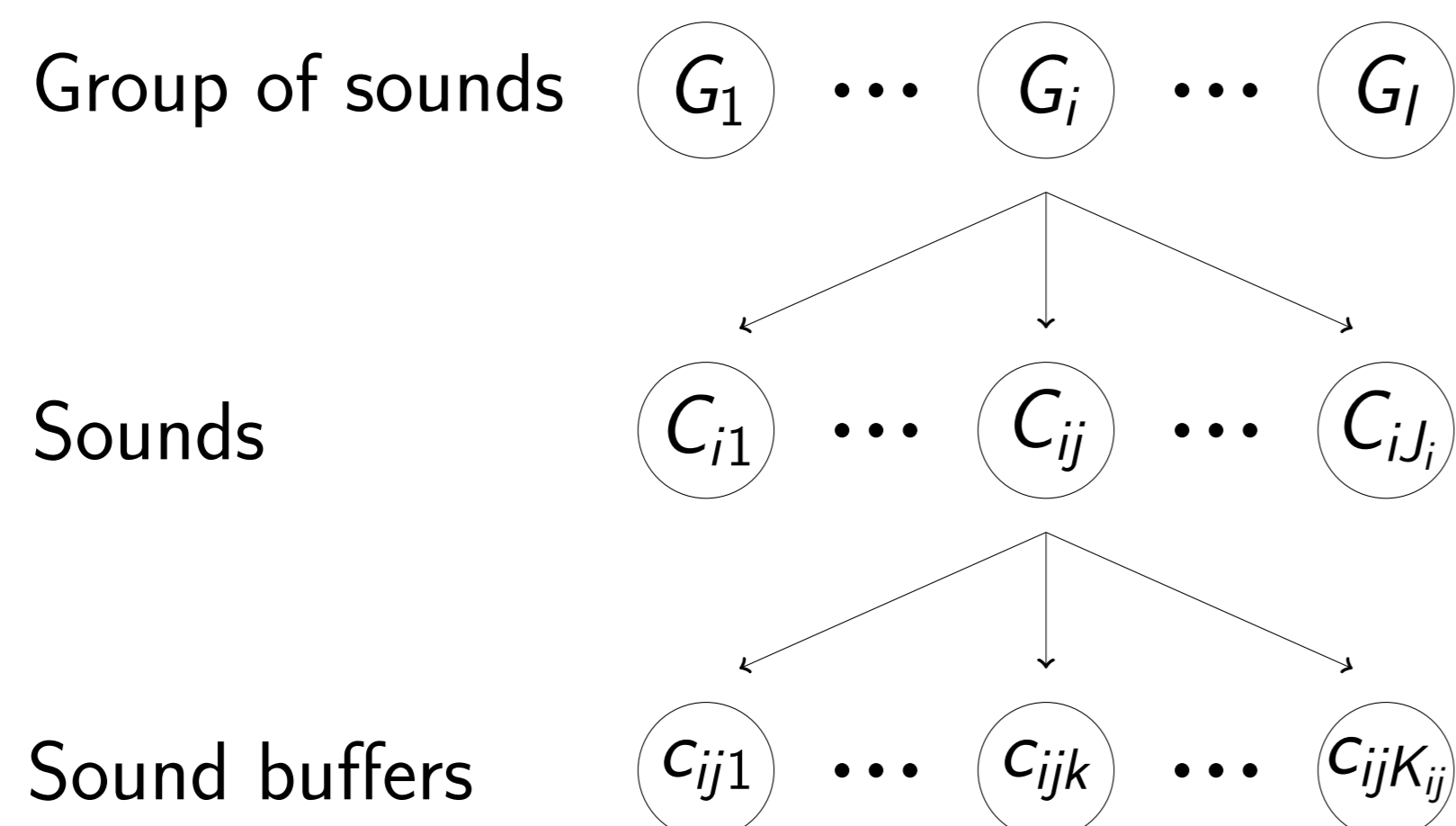


Our approach to real-time audio classification:

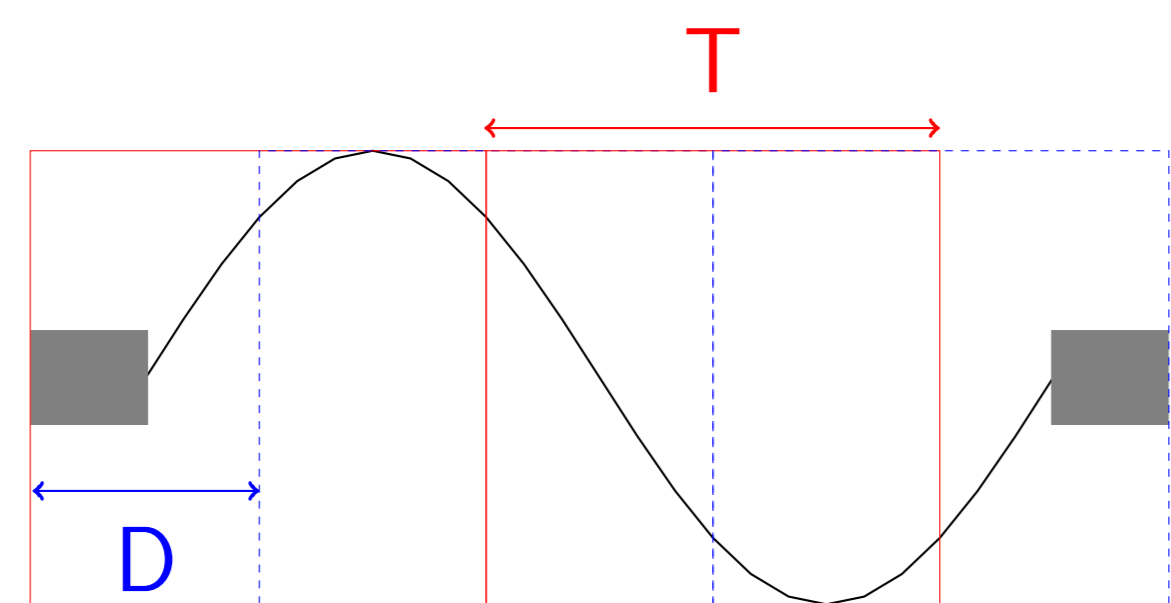


2. CREATE A DICTIONARY OF MODELS

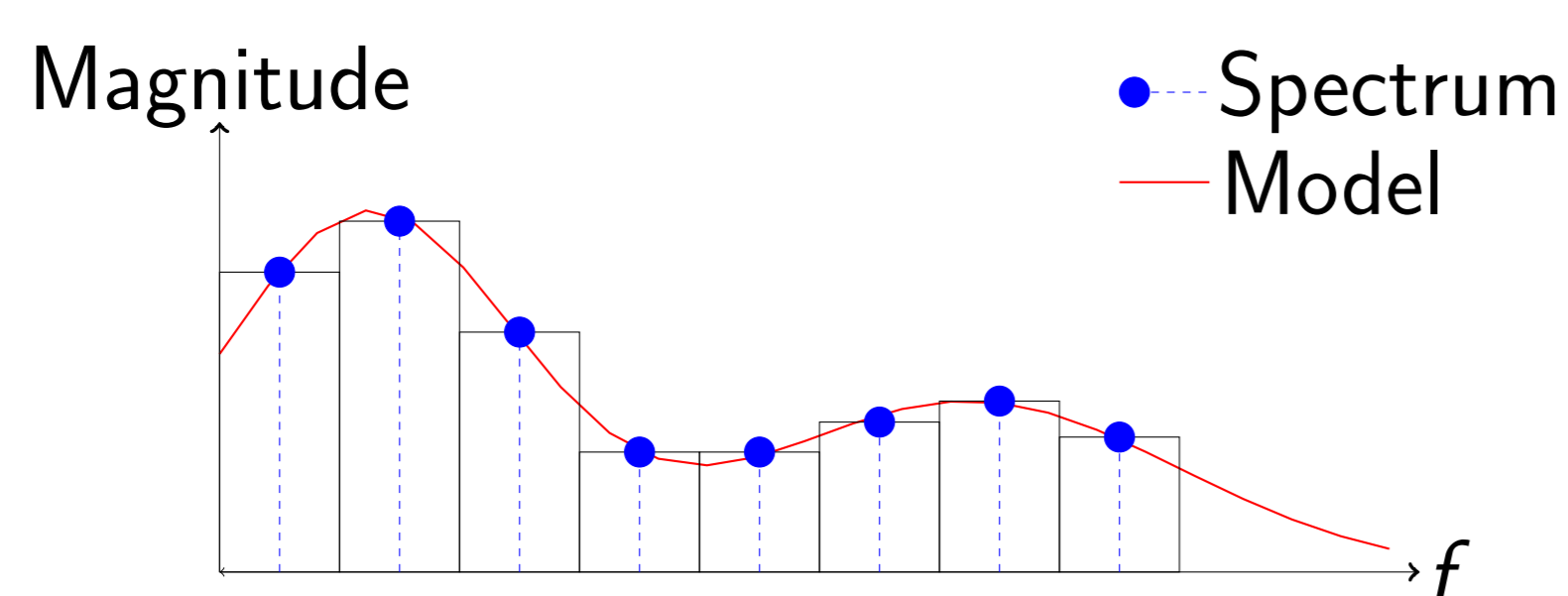
How the sounds are grouped and splitted:



Split a sound into buffers with a window size T and an overlap D :



Modeling of each buffer with a mixture model [2]:



3. SOUND MODELS

Normalized spectrum:

$$S_{ijk}[n] = N \frac{|s_{ijk}[n]|^2}{\sum_{p=1}^N |s_{ijk}[p]|^2}$$

Mixture model:

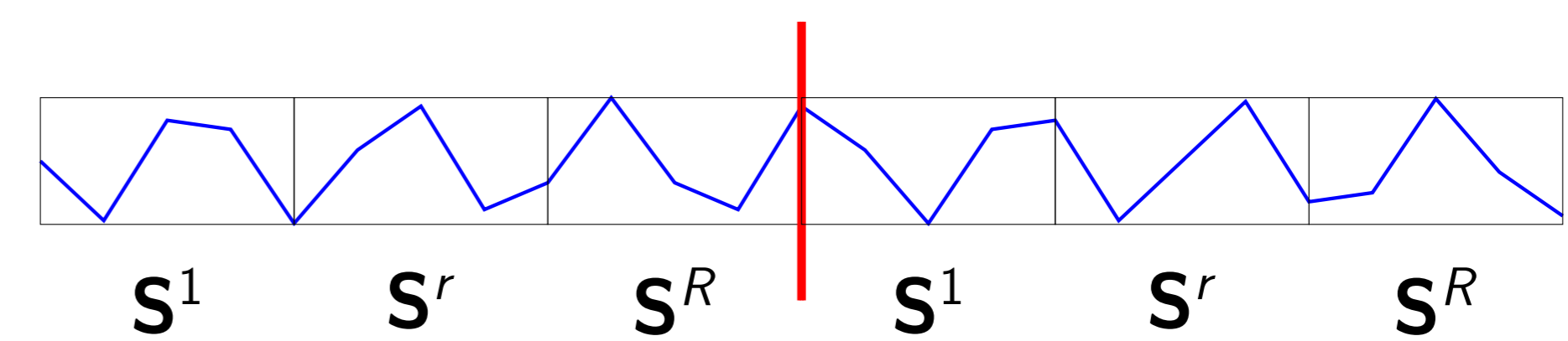
$$p(f|\theta_{ijk}) = \sum_{m=1}^{M_{ijk}} \pi_{ijk}^{(m)} \mathcal{N}\left(f \mid \mu_{ijk}^{(m)}, \left(\sigma_{ijk}^{(m)}\right)^2\right)$$

Model likelihood for binned data:

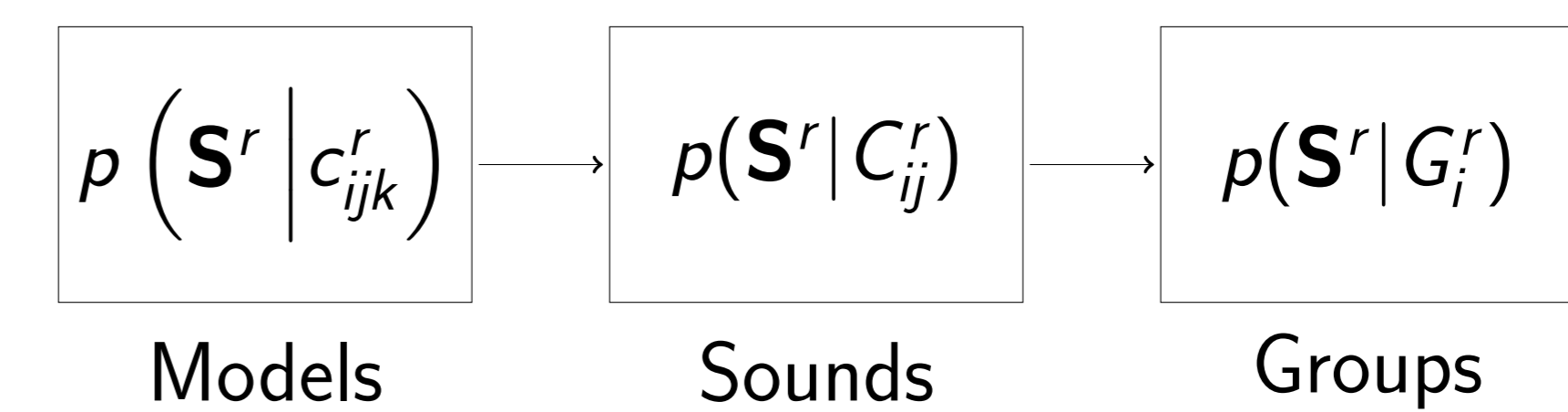
$$\mathcal{L}(\theta_{ijk}) = p(\mathbf{S}|\theta_{ijk}) = \prod_{n=1}^N \left(\int_{f[n]}^{f[n+1]} p(f|\theta_{ijk}) df \right)^{S[n]}$$

4. IDENTIFY NEW SOUNDS

Test sounds \mathbf{S} : Split with window size T and consider groups of R buffers.



Aggregate the likelihoods:



Conditional probabilities of the groups G_i^r :

$$p(G_i^r | \mathbf{S}^r) = \frac{p(\mathbf{S}^r | G_i^r) p(G_i^r)}{\sum_h p(\mathbf{S}^r | G_h^r) p(G_h^r)}$$

Aggregate the probabilities over R buffers:

$$p(G_i | \mathbf{S}) = \prod_{r=1}^R p(G_i^r | \mathbf{S}^r)$$

Final decision (for every group of R buffers):

$$\hat{G}_i = \underset{G_i}{\operatorname{argmax}} p(G_i | \mathbf{S})$$

5. RESULTS & DISCUSSION

Cross-Validation Good classification rate (%)
(Comparison with state-of-the-art methods)

Dataset	A-Volute	ESC-50	ESC-10
Our algorithm	96.5	94.0	96.0
Parametric method	73.6	45.5	73.5
Non-parametric method	46.6	53.2	76.0
Human	91.8	81.3	95.7

Parametric method: standard GMM with standard features [1]

Non-parametric method: Deep ConvNet with spectrogram features [3]

Complexity

(Example on the A-Volute database)

	$O(\text{Number of operations})$
Our algorithm	28×10^6
Parametric method	2×10^3
Non-parametric method	14×10^6

6. RESOURCES

Website available with free demonstrator of the method:



7. REFERENCES

- [1] C. Clavel, T. Ehrette, and G. Richard. "Events Detection for an Audio-Based Surveillance System". In: *2005 IEEE International Conference on Multimedia and Expo*. July 2005, pp. 1306–1309.
- [2] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.
- [3] K. J. Piczak. "Environmental sound classification with convolutional neural networks". In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2015, pp. 1–6.