Real-time Audio Classification based on Mixture Models

Maxime Baelde, Christophe Biernacki, Raphaël Greff

Real-time Audio Classification based on Mixture Models

Maxime Baelde[1,2], Christophe Biernacki[1], Raphaël Greff[2]

[1]University Lille 1, CNRS, Inria   [2]A-Volute

1. INTRODUCTION

Standard Machine Learning (ML) approach for audio classification:

Input sound — $\mathbf{x} = (x[1], \ldots, x[P])$
↓
Preprocessing — STFT, Harmonic decomposition, ...
↓
Feature extraction — Energy, MFCC, ...
↓
ML Classifier — GMM, Neural Networks, ...
↓
ML Model

Our approach to real-time audio classification:

Input sound — $\mathbf{x} = (x[1], \ldots, x[P])$
↓
Preprocessing — STFT: $\mathbf{S} = (S[1], \ldots, S[N])$
↓
Fit Mixture Model — $p(f|\boldsymbol{\theta})$ using $\mathbf{S}$
↓
Dictionary — $\mathcal{D} = \left\{\hat{\boldsymbol{\theta}}\right\}$

2. CREATE A DICTIONARY OF MODELS

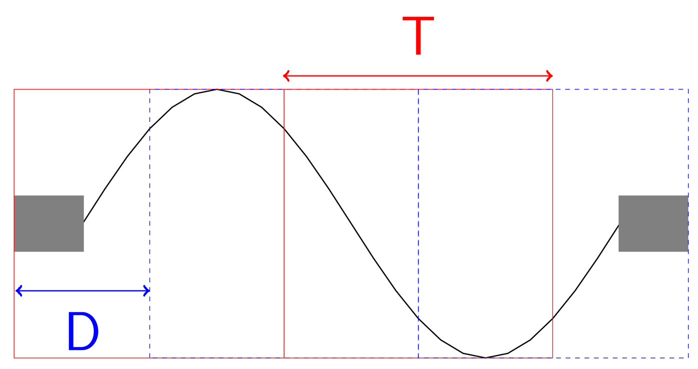
How the sounds are grouped and splitted:

Group of sounds: $G_1 \cdots G_i \cdots G_I$

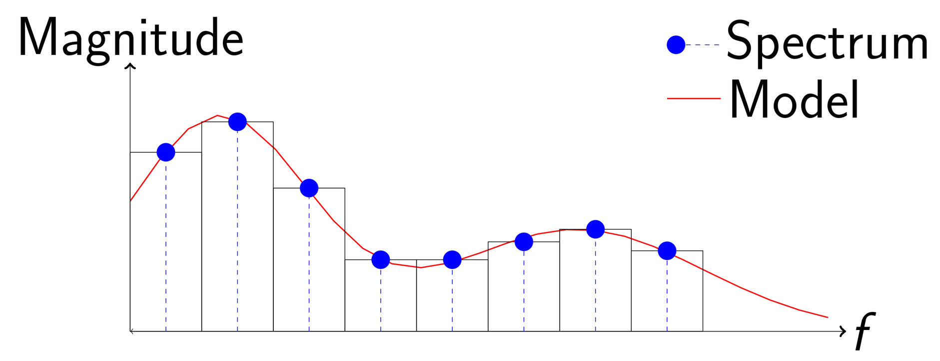
Sounds: $C_{i1} \cdots C_{ij} \cdots C_{iJ_i}$

Sound buffers: $c_{ij1} \cdots c_{ijk} \cdots c_{ijK_{ij}}$

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:

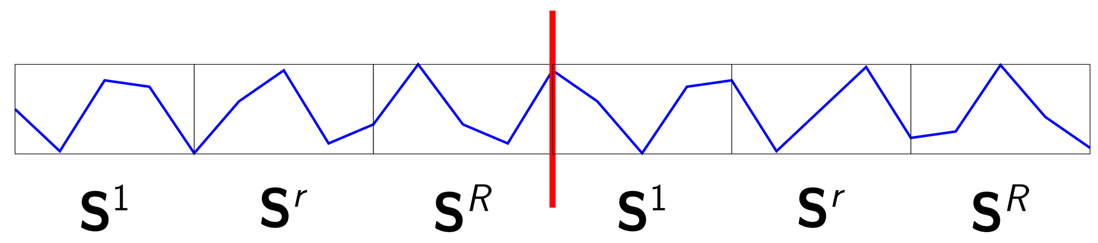
$$p(f|\boldsymbol{\theta}_{ijk}) = \sum_{m=1}^{M_{ijk}} \pi_{ijk}^{(m)} \mathcal{N}\left(f \,\middle|\, \mu_{ijk}^{(m)}, \left(\sigma_{ijk}^{(m)}\right)^2\right).$$

Model likelihood for binned data:

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

4. IDENTIFY NEW SOUNDS

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



Aggregate the likelihoods:

$p\left(\mathbf{S}^r \middle| c_{ijk}^r\right)$ → $p(\mathbf{S}^r|C_{ij}^r)$ → $p(\mathbf{S}^r|G_i^r)$

Models — Sounds — Groups

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):

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

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-paramatic method: Deep ConvNet with spectrogram features [3]

Complexity
(Example on the A-Volute database)

	O(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.