Real-time Audio Classification based on Mixture Models
Maxime Baelde, Christophe Biernacki, Raphaël Greff

To cite this version:
1. INTRODUCTION

Standard Machine Learning (ML) approach for audio classification:
\[ x = (x[1], ..., x[P]) \]

Preprocessing
- STFT, Harmonic decomposition, ...

Feature extraction
- Energy, MFCC, ...

ML Classifier
- GMM, Neural Networks, ...

ML Model

Our approach to real-time audio classification:
\[ x = (x[1], ..., x[P]) \]

Preprocessing
- STFT: \( S = (S[1], ..., S[N]) \)

Fit Mixture Model
- \( p(f|\theta) \) using \( S \)

\[ D = \{ \theta \} \]

2. CREATE A DICTIONARY OF MODELS

How the sounds are grouped and splitted:
- Group of sounds: \( G_1, ..., G_r \)
- Sounds: \( C_{ij}, ..., C_{ij} \)
- Sound buffers: \( S_{ijk}, ..., S_{ijk} \)

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

Modeling of each buffer with a mixture model [2]:
- Magnitude
- Spectrum
- Model

3. SOUND MODELS

Normalized spectrum:
\[ S_{ijk}[n] = N \frac{|s_{ijk}[n]|^2}{\sum_{\beta=1}^{M_{ij}} |s_{ijk}[\beta]|^2} \]

Mixture model:
\[ p(f|\theta_{ijk}) = \sum_{\beta=1}^{M_{ij}} \pi_{ijk}[\beta] \left( f_{ijk}[\beta], \left( \sum_{\beta=1}^{M_{ij}} \pi_{ijk}[\beta] \right) \right)^2. \]

Model likelihood for binned data:
\[ L(\theta_{ijk}) = \sum_{n=1}^{N} \left( \int_{f[n]} \int_{f[n-1]} \int_{f[n]} p(f|\theta_{ijk}) df \right) \pi_{ijk}[n]. \]

4. IDENTIFY NEW SOUNDS

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

Aggregate the likelihoods:
\[ p(S|G) = \sum_{r=1}^{R} p(S|G_r) p(G_r). \]

Conditional probabilities of the groups \( G_r \):
\[ p(G_r|S) = \frac{p(S|G_r) p(G_r)}{\sum_r p(S|G_r) p(G_r)}. \]

Aggregate the probabilities over \( R \) buffers:
\[ p(G_r|S) = \prod_{r=1}^{R} p(G_r|S). \]

Final decision (for every group of \( R \) buffers):
\[ \hat{G}_r = \text{argmax}_C p(G_r|S). \]

5. RESULTS & DISCUSSION

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>A-Volute ESC-50 ESC-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our algorithm</td>
<td>96.5 94.0 96.0</td>
</tr>
<tr>
<td>Parametric method</td>
<td>73.6 45.5 73.5</td>
</tr>
<tr>
<td>Non-parametric method</td>
<td>46.8 53.2 76.0</td>
</tr>
<tr>
<td>Human</td>
<td>91.8 81.3 95.7</td>
</tr>
</tbody>
</table>

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: \( 28 \times 10^6 \)
- Parametric method: \( 2 \times 10^3 \)
- Non-parametric method: \( 14 \times 10^6 \)

6. RESOURCES

Website available with free demonstrator of the method:

7. REFERENCES

