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

To cite this version:

HAL Id: hal-01481934
https://hal.archives-ouvertes.fr/hal-01481934
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:
- Input sound: \( x = (x[1], \ldots, x[P]) \)
  - Preprocessing: STFT, Harmonic decomposition, ...
  - Feature extraction: Energy, MFCC, ...
  - ML Classifier: GMM, Neural Networks, ...
  - ML Model

Our approach to real-time audio classification:
- Input sound: \( x = (x[1], \ldots, x[P]) \)
  - Preprocessing: STFT: \( S = (S[1], \ldots, S[N]) \)
  - 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: \( G_1, \ldots, G_r \)
- Sounds: \( C_{ij}, \ldots, C_{ij}, \ldots, C_{ij} \)
- Sound buffers: \( r_{ij}, \ldots, r_{ij}, \ldots, r_{ij} \)

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:
\[
S[p][n] = \frac{|s[p][n]|^2}{\sum_{p=1}^{M} |s[p][n]|^2}
\]

Mixture model:
\[
p(f|\theta_{ij}) = \sum_{m=1}^{M} \pi_{ij}^{(m)} \left( f \left| \mu_{ij}^{(m)}, \sigma_{ij}^{(m)} \right) \right.
\]

Model likelihood for binned data:
\[
L(\theta_{ij}) = p(S|\theta_{ij}) = \prod_{i=1}^{N} \int_{f[i]} p(f|\theta_{ij}) df \pi_{ij}
\]

4. IDENTIFY NEW SOUNDS

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

Aggregated the likelihoods:
- Models
- Sounds
- Groups

Conditional probabilities of the groups \( G_i \):
\[
p(G_i|S) = \frac{p(S|G_i)p(G_i)}{\sum_{r=1}^{R} p(S|r_i)p(G_i)}
\]

Aggregated the probabilities over \( R \) buffers:
\[
p(G_i|S) = \prod_{r=1}^{R} p(G_i|S_r)
\]

Final decision (for every group of \( R \) buffers):
\[
\hat{G_i} = \arg\max_{G_i} p(G_i|S)
\]

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>46.6 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