A mixture model-based real-time audio sources classification method
Maxime Baelde, Christophe Biernacki, Raphaël Greff

To cite this version:
Maxime Baelde, Christophe Biernacki, Raphaël Greff. A mixture model-based real-time audio sources classification method. The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP2017, Mar 2017, New Orleans, United States. hal-01420677v2

HAL Id: hal-01420677
https://hal.archives-ouvertes.fr/hal-01420677v2
Submitted on 28 Dec 2016

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.
A MIXTURE MODEL-BASED REAL-TIME AUDIO SOURCES CLASSIFICATION METHOD

Maxime Baelde1,2, Christophe Biernacki1, Raphaël Greff2

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

ABSTRACT
Recent research on machine learning focuses on audio source identification in complex environments. They rely on extracting features from audio signals and use machine learning techniques to model the sound classes. However, such techniques are often not optimized for a real-time implementation and in multi-source conditions. We propose a new real-time audio single-source classification method based on a dictionary of sound models (that can be extended to a multi-source setting). The sound spectrums are modeled with mixture models and form a dictionary. The classification is based on a comparison with all the elements of the dictionary by computing likelihoods and the best match is used as a result. We found that this technique outperforms classic methods within a temporal horizon of 0.5s per decision (achieved 6% of errors on a database composed of 50 classes). Future works will focus on the multi-sources classification and reduce the computational load.

Index Terms— real-time, audio identification, statistical learning, mixture models, sound classification.

1. INTRODUCTION
Audio source classification is a vast and trendy topic in machine learning, and can be divided into three groups. Music Information Retrieval (MIR) aims at recognizing musical instruments [1] or musical genres [2] from musics. Automatic Speech Recognition (ASR) aims at detecting and identifying speakers in audio recordings [3]. Environmental Sound Recognition (ESR) aims at recognizing classes of sounds that are neither music nor speech [4]: for instance, airplanes or gunshots.

Typical audio source classification methods include two stages. The first one consists in extracting features from the signals using audio descriptors [5]. Common features are temporal (energy, zero crossing rate,...), spectral (Mel Frequency Cepstral Coefficient, centroid,...) or harmonic (fundamental frequency, inharmonicity,...). This step extracts relevant information from the signal and can be seen as a dimension reduction. The second one uses machine learning algorithm to model the sound classes based on the previous features. Commonly used algorithms are Gaussian Mixture Models (GMM) [6, 7], Support Vector Machines (SVM) [8, 9], Hidden Markov Models (HMM) [10, 11] or Neural Networks (NN) [12, 13]. More recently, research focuses on neural networks and uses them for features extraction and classification at the same time. These networks often involve convolutional layers and deep architecture (Deep Convolutional Neural Network, DCNN) to model fine details in signals [3, 14].

In this study, we develop an audio single-source classification system based on a dictionary of models in a probabilistic framework, using the mixture model theory [15]. Indeed, the future goal of this research is to identify several sound sources at the same time, which is merely impossible with the current techniques. The use of mixture models allows to deal with mixture of sounds and therefore to identify simultaneous audio sources. Each sound spectrum is modeled by a mixture model and all the models constitute a dictionary. The classification is performed by comparing an unknown signal with the elements of the dictionary by computing likelihoods, aggregating these probabilities and taking the Maximum A Posteriori (MAP).

The rest of the paper is organized as follows. The creation of the dictionary is presented in Section 2. The single-source identification procedure that uses the previous dictionary is detailed in Section 3. The experiments carried out to assess the performance of the method are presented in Section 4. Section 5 presents the results of the experiments. Finally, a brief discussion is presented in Section 6 and concludes the paper.

2. CREATION OF THE DICTIONARY
The classification algorithm deals with multiple classes of sounds, denoted $G_i$, $i = 1, ..., I$, and a mono-channel stream. For instance, the classes can be $G_1 = \text{airplane}$, $G_2 = \text{gunshot}$. In each group $G_i$ there are multiple sounds, labeled $C_{ij}$, $j = 1, ..., J_i$. For instance in $G_1$, $C_{ij} = \text{airplane}_1$, and in $G_2$, $C_{23} = \text{gunshot}_3$. As a typical signal processing step for real-time application, the sounds $C_{ij}$ are splitted into buffers, labeled $c_{ijk}$, $k = 1, ..., K_{ij}$ of size $T$ samples with a time shift of $D$ samples (see Fig. 1). In a signal notation, $c_{ijk} = C_{ij}[kD : kD + T]$. In real world applications, the signal does not necessarily fit correctly into the buffer. To tackle this problem, Gaussian white noise is added at the beginning and the end of each sound.
where Gaussian white noise is added at the variable. Here the random variable is the frequency set of parameters the spectrum is computed, and the pdf is parameterized by a the components, which represents the distribution of a random data \([s_{ijk}[n]]\) is a mixture of several probability density functions (pdf), the mixing coefficients \(\pi_{ijk}\) are considered since we want to keep the addi-
tivity of the spectrums. These spectrums are considered as
multiple sounds present at the same time, the energy spectrum
is used. Indeed, when two uncorrelated signals are mixed,
cause most of the information is included in the low-frequency
content of the signal. As our algorithm is designed to classify
steps until convergence:

\[L(\theta) = p(S|\theta) = \prod_{n=1}^{N} p(S[n]|\theta),\]  

where \(S = [S[1], ..., S[N]]\). Starting with an initial parameter estimate \(\theta^{(0)}\), the EM algorithm iterates the two following steps until convergence:

(E) Compute: \(Q(\theta | \theta^{(p)}) = E[\log L(\theta) | S, \theta^{(p)}] \).

(M) Maximize: \(\theta^{(p+1)} = \arg\max_{\theta} Q(\theta | \theta^{(p)}) \).

This model and the related algorithm require the number of components \(M\) in the mixture to be specified. For each buffer, the optimal number of components is chosen by minimizing the Bayesian Information Criterion (BIC) of Schwarz [18]:

\[BIC(M) = -2 \log L(\hat{\theta}) + d \log N,\]  

where \(\hat{\theta}\) denotes the Maximum Likelihood (ML) value of \(\theta\) and \(d = 3M - 1\) is the dimension of \(\theta\), with \(M\) varying from 1 to 20 in this case. The estimated parameters are gathered into \(\{\hat{\theta}_{ijk}\}\), with \(k = 1, ..., K_{ij}, j = 1, ..., J_{i}, i = 1, ..., I\), which is named the dictionary.

3. IDENTIFICATION PROCEDURE

Once the dictionary is computed (this corresponds to the learning step), it can be used to identify unknown sounds. Suppose that at a time \(t\) there is a new buffer which is transformed into a normalized energy spectrum \(S'\). The identification process

\[N(f | \mu_m, \sigma^2_m) = \frac{1}{\sqrt{2\pi \sigma^2_m}} \exp\left(-\frac{1}{2} \left(\frac{f - \mu_m}{\sigma_m}\right)^2\right).\]  

The frequency \(f\) is not known precisely: we know only how many frequencies \(S[n]\) fall into the frequency range \([f[n], f[n+1]]\). This framework is close to the one of McLaugh-
lan [16] except that \(S[n]\) is not an integer. Thus, given that the for \(i\) are independent and identically distributed (i.i.d.), the probability that one sample falls into \([f[n], f[n+1]]\) is:

\[p(S[n]|\theta) = \left(\int_{f[n]}^{f[n+1]} p(f|\theta) df\right)^{S[n]}.\]  

The Expectation-Maximization (EM) algorithm of Demp-
ster [17] is used to estimate the set of parameters \(\theta\). This algorithm finds the parameters that maximize the likelihood of the model parameterized by \(\theta\). Given that the \(S[n]\)s are i.i.d., the likelihood is:

\[\mathcal{L}(\theta) = p(S|\theta) = \prod_{n=1}^{N} p(S[n]|\theta),\]  

This is used to identify unknown sounds. The algorithm finds the parameters that maximize the likelihood of the model parameterized by \(\theta\). Given that the \(S[n]\)s are i.i.d., the likelihood is:

\[\mathcal{L}(\theta) = p(S|\theta) = \prod_{n=1}^{N} p(S[n]|\theta),\]  

where \(S = [S[1], ..., S[N]]\). Starting with an initial parameter estimate \(\theta^{(0)}\), the EM algorithm iterates the two following steps until convergence:

(E) Compute: \(Q(\theta | \theta^{(p)}) = E[\log L(\theta) | S, \theta^{(p)}] \).

(M) Maximize: \(\theta^{(p+1)} = \arg\max_{\theta} Q(\theta | \theta^{(p)}) \).

This model and the related algorithm require the number of components \(M\) in the mixture to be specified. For each buffer, the optimal number of components is chosen by minimizing the Bayesian Information Criterion (BIC) of Schwarz [18]:

\[BIC(M) = -2 \log L(\hat{\theta}) + d \log N,\]  

where \(\hat{\theta}\) denotes the Maximum Likelihood (ML) value of \(\theta\) and \(d = 3M - 1\) is the dimension of \(\theta\), with \(M\) varying from 1 to 20 in this case. The estimated parameters are gathered into \(\{\hat{\theta}_{ijk}\}\), with \(k = 1, ..., K_{ij}, j = 1, ..., J_{i}, i = 1, ..., I\), which is named the dictionary.
consists in finding the correct label $G_i^r$ by aggregating the following probabilities:

1. Compute the likelihoods of all the models of the dictionary $p(S^r | c_{ijk}^r) = p(S^r | \theta_{ijk})$, for $k = 1, ..., K_{ij}$, $j = 1, ..., J_i$ and $i = 1, ..., I$ using Eq. (2.4).

2. Compute the likelihood of the sound $C_{ij}^r$, for $j = 1, ..., J_i$ and $i = 1, ..., I$:

$$p(S^r | C_{ij}^r) = \sum_{k=1}^{K_{ij}} p(S^r | c_{ijk}^r)p(c_{ijk}^r | C_{ij}^r).$$ \hspace{1cm} (3.1)

3. Compute the likelihood the group $G_i^r$, for $i = 1, ..., I$:

$$p(S^r | G_i^r) = \sum_{j=1}^{J_i} p(S^r | C_{ij}^r)p(C_{ij}^r | G_i^r).$$ \hspace{1cm} (3.2)

4. Finally compute the conditional probability of the group $G_i^r$ for $i = 1, ..., I$:

$$p(G_i^r | S^r) = \frac{p(S^r | G_i^r)p(G_i^r)}{\sum_h p(S^r | G_h^r)p(G_h^r)}.$$ \hspace{1cm} (3.3)

The decision rule uses $R$ buffers (as in [6]) and is computed as follows. For each buffer $r$ and each class the classifier computes a probability $p(G_i^r | S^r)$. Then these probabilities are multiplied so as to aggregate the results over a larger temporal horizon:

$$p(G_i | S) = \prod_{r=1}^{R} p(G_i^r | S^r),$$ \hspace{1cm} (3.4)

where $S = [S^1, ..., S^R]$. The MAP estimate is finally taken on these probabilities: $\hat{G}_i = \arg\max_{G_i} p(G_i | S)$.

**4. EXPERIMENTS**

**4.1. Setup of the experiments**

Several experiments were carried out to assess the quality of our sound sources classification algorithm. Three databases were considered for these experiments. One database (from A-Volute) was composed of 704 video games sounds divided into 9 classes. The ESC databases [19] (50 and 10) were also considered, composed of 2000 environmental sounds from 50 classes for the former and 400 sounds from 10 classes for the latter. Each sound was resampled to 44.1kHz and the mean was subtracted to the signal because it caused instability in the fitting process. Several values were considered for the window size $T$ and the time shift $D$, mainly $T = \{512, 1024, 2048\}$ samples and $D = 512$ samples. The number of kept bins $N$ was set to $T/5$ which corresponded to approximately 8kHz in our setting. $R$ was set to 10, that corresponds approximately to 0.5s.

Each dataset were splitted into a training and a test set for cross-validation procedure. We used 80% of the set for training and the remainder for testing. The division was done in a $v$-fold manner, with $v = 5$. The recognition rate, defined by the number of buffers correctly labeled over the total number of buffers, was used to assess the performance of the system. In a cross-validation setting, a recognition rate for each fold was computed and the cross-validation recognition rate was simply the mean of the $v$ folds recognition rates.

Three probabilities were introduced in Section 3 and have to be specified:

- The probability of a buffer $c_{ijk}$ given a sound $C_{ij}$ was set to 1: $p(c_{ijk} | C_{ij}) = 1$ for $k = 1, ..., K_{ij}$, $j = 1, ..., J_i$ and $i = 1, ..., I$.
- The probability of a sound $C_{ij}$ given a class $G_i$ was uniformly distributed over the class $G_i$: $p(C_{ij} | G_i) = 1/\#\{G_i\}$ for $j = 1, ..., J_i$ and $i = 1, ..., I$, where $\#\{A\}$ means the number of elements in $A$.
- The probability of a class $G_i$ was set to the ratio between the number of sounds in $G_i$ and the total number of sounds: $p(G_i) = \#\{G_i\} / \sum_h \#\{G_h\}$ for $i = 1, ..., I$.

**4.2. Comparison with other techniques**

We compared our algorithm with other state-of-the-art techniques. The first one was a parametric method inspired from Clavel [6]. Acoustic features were extracted from the signals: energy, 8 MFCC, spectral centroid, spectral spread, plus the first and second derivatives. A Principal Component Analysis step was applied on these descriptors and only the first 13 principal components were kept. The classifier was a standard GMM.

The second one was a non-parametric method: a DCNN. A neural network is a set of neurons (computation units) stacked together in multiple layers that can perform classification or regression. The input of a layer $l$ is denoted by $h_{l-1}$, and the convolutional kernel $W_l$. The network computes an output $h_l$ by taking an activation function $g$: $h_l = g(W_l \ast h_{l-1})$, where $\ast$ is the convolution operation. Recently, the most popular activation function is the ReLU (Rectified Linear Unit), which is simply $g(x) = \max(x, 0)$ element-wise. After a convolutional layer, a max-pooling operation consisting in merging adjacent cells by taking the maximum value is applied. We used the network developed by Piczak [14]. The input was a log-mel spectrogram with its delta, considered as a 2-channel images of size $60 \times 41$. The first convolutional layer consisted in 80 filters of size $57 \times 6$, with a max-pooling of size $4 \times 3$. The second convolutional layer consisted in 80 filters of size $1 \times 3$, with a max-pooling of size $1 \times 3$. Finally, two fully connected layers composed of 5000 units each and a softmax layer ended the network.
We evaluate the complexity of the previous algorithms at the
where we carried out a listening test on the A-V olute dataset. 21
4.3. Human listening tests
The experiments were carried out for all the values of
putations needed to compute the decision for one buffer can be
5. RESULTS
5.1. Recognition rate
The experiments were carried out for all the values of \(T\). However, only the results for \(T = 2048\) are shown because they are the best, and are presented in Table 1. The results for the A-V olute’s database are the following. The parametric method achieves 73.6% and the non-parametric method 46.6%. Our algorithm outperforms the current methods by achieving a recognition rate of 96.5%. Concerning the ESC-50 database, the parametric method achieves 44%, the non-parametric method 53.2% and our algorithm 94.0%. Finally, on the ESC-10 dataset, the parametric method achieves 73.5%, the non-parametric method 76.0% and our algorithm 96.0%. It is worth noticing that a human can achieve 91.8% on the A-V olute dataset, 94.0% on the ESC-50 database and 95.7% on the ESC-10 database [19]. Our algorithm outperforms state-of-the-art methods and humans on these databases.

Table 1. The recognition rates in % of the considered methods
on the different datasets (A-V olute, ESC-50 and ESC-10).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>A-V olute</th>
<th>ESC-50</th>
<th>ESC-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric</td>
<td>73.6</td>
<td>45.5</td>
<td>73.5</td>
</tr>
<tr>
<td>Non-parametric</td>
<td>46.6</td>
<td>53.2</td>
<td>76.0</td>
</tr>
<tr>
<td>Our algorithm</td>
<td>96.5</td>
<td>94.0</td>
<td>96.0</td>
</tr>
<tr>
<td>Human</td>
<td>91.8</td>
<td>81.3</td>
<td>95.7</td>
</tr>
</tbody>
</table>

For each previous technique, the hyperparameters were adapted so as to have comparable results.

5.2. Complexity and execution time
We evaluate the complexity of the previous algorithms at the
identification step. Our algorithm is mainly concerned with
multiplication. Indeed, by computing the log-probability, computations are just additions and multiplications. All the operations needed to compute the decision for one buffer can be resumed as \(O(\#D \cdot T/5)\). The parametric method relies mainly on computing exponentials and matrix operations (inversion and determinant), so the complexity is roughly \(O(d^3)\), where \(d\) is the dimension of the feature vector. The neural network uses convolutions which are the more expensive operations when evaluating the decision. The complexity is
thus \(O\left(\sum_{i=1}^{L} n_{i-1}(s_i) s_i n_l(m_l^1 m_l^2)\right)\), with \(L\) the number of layers, \(n_{i-1}\) the number of input channels for layer \(l\), \(s_i\) the dimensions of the input, \(n_l\) the number of filters of the layer \(l\) and \(m_l^1\) the size of the output.

Considering the settings of the experiments, the numbers of operations needed for one buffer for our algorithm are \(O(28\cdot 10^6)\) (A-V olute database), \(O(63 \cdot 10^6)\) (ESC-10) and \(O(120 \cdot 10^6)\) (ESC-50), for the parametric method \(O(2 \cdot 10^3)\) and for the neural network \(O(14 \cdot 10^6)\).

We also report the time needed to compute the decision. The machine used a Intel Core i7-5820K CPU @3.30GHz. Using a dictionary composed of approximately 70,000 models, it takes 185ms to the system to infer the decision for one buffer. By using an NVidia GeForce GTX 960 to compute the multiplications, this computational time can be reduced to 35ms.

6. DISCUSSION AND CONCLUSION
The aim of this algorithm is to perform real-time audio multi-source identification. Up to now, this study examined only the single-source case. As we can see in the results, our technique outperforms standard methods (GMM), more recent algorithms (DCNN) and even humans on both benchmark [19] and industrial databases. Despite these good recognition rates, the main advantage of our algorithm is that it can theoretically handle multi-source conditions. The learning algorithm of classic techniques would have to learn every combination of sound classes, which is practically impossible because of the combinatorial and the amount of data required for training. Some research on neural networks begins to study polyphonic identification, as in [20].

The energy spectrum is chosen so as to extend this
the multi-source setting. Indeed, we make the assumption that the additivity of two energy spectrums is preserved (uncorrelated sources). Mixture models were employed because they allow a flexible modeling of any signal. Moreover, in a multi-source setting, the signal model is a mixture of mixture, which is well defined in the mixture model framework.

The real-time constraint requires that the signals are not
known by advances. Besides, we want the system to have a
low latency. This is why such values of \(T\) and \(D\) were chosen. However, it is not reach yet. Indeed, the larger the dictionary the longer it takes to compute the decision. This is why future work will try to reduce the computation time, for instance by organizing the dictionary in a binary tree.
7. REFERENCES


