Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm

Christian Zieger, Maurizio Omologo
Fondazione Bruno Kessler-irst, via Sommarive 18, 38050 Povo, Trento, Italy
zieger@fbk.eu, omologo@fbk.eu

Abstract

This work proposes a system for acoustic event classification using signals acquired by a Distributed Microphone Network (DMN). The system is based on the combination of Gaussian Mixture Models (GMM) and Support Vector Machines (SVM). The acoustic event list includes both speech and non-speech events typical of seminars and meetings. The robustness of the system was investigated by considering two scenarios characterized by different types of trained models and testing conditions. Experimental results were obtained by using real-world data collected at two sites. The results in terms of classification error rate show that in each scenario the proposed system outperforms any single classifier based system.

Index Terms: audio classification, distributed microphone network, GMM, SVM.

1. Introduction

Acoustic scene analysis aims to describe any acoustic event that can be observed in a given environment, in terms of space, time, and type, by means of one or more microphones signals. This paper focuses on acoustic event classification which deals with the interpretation of the nature of sounds.

In the literature some papers on acoustic event classification can be found: [1] focused on detecting a single event (laughter); in [2, 3] speech and music classification was explored; in [4] acoustic events for medical telesurvey were considered; in [5] audio events were detected to automatically extract highlights from baseball, golf and soccer matches; in [6] detection of animal sounds was investigated, in [7, 8, 9] sounds were classified for surveillance purposes.

This work deals with the classification of speech and non-speech events, where the considered non-speech events are typical sounds that may occur during seminars or meetings. This problem was first addressed under the CHIL project (http://chil.server.de/servlet/is/101/) as described in [10], which compared three different systems under the CLEAR ’06 evaluation on acoustic event detection and classification organized by NIST. In this paper, which focuses only on the problem of classification, we extend the previous work from using a single microphone to exploiting the acoustic diversity observed by a set of microphones placed far from each other. A new algorithm is proposed which combines SVM and GMM and aims to classify different possible events, among which speech. The main goal is to develop a system that is robust to the possible mismatch between training and testing conditions. The mismatch can be caused by many factors, such as different types of microphones, different positions of the acoustic sources, and different background noises. Moreover, the acoustic properties of the same event, as for instance in the case of the sound produced by a “falling object”, can change from one room to another according to the different room size, the type of furniture, the floor, etc. To this purpose, in this work two different sites were used for real experiments dealing with the possible mismatch between training and testing conditions. The proposed system combines two different levels of acoustic information fusion, namely microphone level and identification level. We will show that the system outperforms each single classifier based system in terms of classification error rate.

This research is being conducted under the European Project DICIT (see http://dicit.fbk.eu for more details) which studies the use of the acoustic event classification to increase the performance of distant-talking speech recognition or speaker recognition systems in noisy and cocktail-party contexts. For this reason, we are also interested in measuring the system performance in terms of misses and false alarms of speech, as addressed in the experimental section of this work.

The paper is organized as follows: section 2 describes the proposed system, section 3 gives a description of the addressed task, section 4 presents the experimental results and the last section gives some conclusions.

2. System Description

In order to increase system robustness against all possible locations of the acoustic source, a Distributed Microphone Network is here adopted, which comprises at least one microphone per wall to ensure a good spatial coverage. Moreover the proposed system is based on the combination of two different classifiers, GMM and SVM, in order to exploit the complementary characteristics of each classifier. In fact, GMM usually performs very well under matched conditions, but the performance decreases under mismatched conditions. On the other hand, SVM is characterized by a high discriminative power, as it is able to separate non-linear separable classes thanks to a non-linear mapping of the feature space.

The block diagram of the proposed system is depicted in Figure 1. An acoustic event is acquired through a single microphone and the front-end codes the signals into feature vector sequences. In the scoring process, a GMM/SVM combined classifier computes a score vector by matching the feature vector sequences to the given L models, which are trained on events related to the specific scenario. Finally, the event classification is performed based on the score vectors derived from all of the microphones.

In the following of this section, the front-end, the GMM and SVM based classifiers, the utilized score fusion method and the decision making technique are described.
2.1. Front-end

The front-end codes the acoustic event sampled at 44.1 kHz into a 38 dimensional feature vector sequence composed of 12 Mel Frequency Cepstral Coefficients (MFCCs) with first and second derivatives and the first and second derivatives of the log energy. The analysis window is 20 ms and the analysis step is 10 ms. Pre-emphasis is applied to each input signal by using a first order FIR filter.

2.2. GMM

The GMM is a parametric statistical model and is completely defined when the weights $w_k$, the mean vectors $\mu_k$, the covariance matrices $\Sigma_k$ and the number of mixtures $M$ are defined. In the following a general GMM will be denoted as $G = (w_k, \mu_k, \Sigma_k)$.

The likelihood function computed on the feature vector $x$ with the model $G$ is

$$ p(x|G) = \sum_{k=1}^{M} w_k p_k(x) $$

where $p_k(x)$ is the k-th multidimensional Gaussian probability density function with mean vectors $\mu_k$ and covariance matrices $\Sigma_k$.

The score of a feature vector sequence $X = \{x_1, \ldots, x_T\}$ for the l-th event model is defined as:

$$ S_{GMM_l}(X) = \sum_{t=1}^{T} \log p(x_t|G_l) $$

2.3. SVM

SVM is a two-class classifier technique. It is based on hyperplane separators chosen in order to maximize the distance between the hyperplane and the closest training vectors (maximum margin), called support vectors. SVM computes the following discriminative function on the test feature vector $x$ in the following way:

$$ f(x) = \sum_{j=1}^{L} a_j t_j K(x, x_{sv_j}) + \rho $$

where $K(\cdot, \cdot)$ is the Kernel function, $t_j$ are the ideal outputs (1 or -1 depending upon which class the support vector belongs to). The support vectors $x_{sv_j}$, the support vector number $L$, the coefficients $a_j$ and the constant $\rho$ are obtained through an optimization process during the training phase.

The extension to the multi-class classification is obtained through the “one-against-one” approach in which $L(L-1)/2$ classifiers are trained with data of two different classes and in the classification process a voting strategy is adopted [11].

The score of a feature vector sequence $X$ for the event model $l$ is the ratio between the number of feature vectors classified as event $l$ and total number of feature vectors

$$ S_{SV M_l}(X) = \frac{T_l}{T} $$

In order to reduce the computational complexity in the training and testing processes, each event sample of the training data was quantized using the K-means clustering algorithm considering a number of centroids $Nc$. Clearly, this operation reduces the performance of the classification, but at the same time it provides a smaller number of support vectors, this way reducing the computational load.

2.4. GMM/SVM Score Fusion

The classifier score fusion combines GMM and SVM scores through a weighted sum. The scoring procedure for the feature vector produced at i-th microphone is illustrated in Figure 2.

Score normalization is applied to both GMM and SVM scores. Given the score vector $S_{Ci} = [S_{GMM_1} S_{GMM_2} \cdots S_{GMM_L}]$ at i-th microphone for one of the two classifiers, each component of the score vector is normalized such that it is in the range between 0 and 1 in the following way:

$$ \hat{S}_{Ci} = \frac{S_{Ci} - \min(S_{Ci})}{\max(S_{Ci}) - \min(S_{Ci})} $$

For the weights computation, a Matcher Weighting (MW) is applied. Each score is weighted by a factor inverse proportional to the Classification Error Rate (CER), so that the weights for more accurate classifiers are higher than those of less accurate classifiers:

$$ w_{g,s} = \frac{1}{CER_g + 1/CER_s} $$

where $CER_g$ and $CER_s$ are the CER for the GMM and SVM respectively and $w_g$, $w_s$ are the corresponding weights.

2.5. Classification

Once an event has been classified (on the basis of the highest score) for each microphone, the resulting scores are combined using two methods: Major Voting (MV) and Maximum Likelihood (ML). As for MV, the classified event is identified as the one that gets the majority of votes (for cases with equal vote numbers the event is randomly selected among the best ones). As for ML, the classified event corresponds to the event with the highest likelihood obtained among all the given microphones.
3. Task Description

The data used for training and testing the given system are extracted from the isolated acoustic event database and the meeting recordings collected under the CHIL project, at Universitat Politecnica de Catalunya (UPC) and at FBK-irst site.

Concerning the list of possible events, a list of 12 possible non-speech events was defined which includes: coughing, laughing, applauding, door opening/slamming, door knocking, keyboard typing, spoon/cup jingling, key jingling, paper rustling, phone ringing, chair moving, footsteps. Speech also represents a possible event to classify.

Our aim is to implement a system that is robust under all kinds of conditions. For this reason two scenarios were set-up according to different types of restrictions in order to test the robustness of the system:

- **SD**: a Site Dependent scenario, which assumes the prior knowledge of the site type. During training, one model for each site is obtained with data acquired by all the microphones of that site. In the scoring process the model corresponding to the related site is selected.

- **SI**: a Site Independent scenario, which does not make any assumptions about the site type. During training, a single model is obtained by using the data acquired by all the microphones installed in both sites.

The first scenario could be typical of applications in which the system always operate in the same room. The second one is more challenging, since it is free of any site constraints and the portability from one room to another is highly desirable. The system performance in the **SI** scenario should demonstrate the system robustness against all possible mismatches from one room to another.

### 3.1. Acoustic Event Database

As shown in the room and microphone layouts reported in Figures 3 and 4, the addressed environments differ each other in terms of size and sensor distribution in space.

The **UPC** database of isolated acoustic events was recorded using 84 microphones, namely, Mark III (array of 64 microphones), three T-shape clusters (4 mics per cluster), 4 tabletop directional and 4 omni-directional microphones. Approximately 60 repetitions per event class were recorded. The events could occur in different positions in the room.

The **FBK**-irst database of isolated acoustic events was recorded with 32 microphones: 7 T-shaped arrays, composed by 4 microphones each one, and 4 table microphones. Approximately 50 repetitions per event class were recorded. The events could occur in different positions in the room.

The recorded meetings were held in the same rooms: six meetings at **UPC** and five meetings at **FBK**. For each meeting 4 participants were free to move in the room and talk. The background noise of the rooms was due to different types of noise sources, as PC fans and air-conditioning system. For a more detailed description of the two databases, one can refer to [10].

The experiments described in the next section refer to the use of a subset of microphones in each room, selected in order to ensure a good spatial coverage. The limited number of microphones also allowed a reduced computational load. As for **UPC** room, five microphones were selected, one placed on each wall and one on the table. As for **FBK**, eight microphones were selected, one from each of the seven T-shaped arrays and one on the table. The test set consists of 585 events: 223 events derive from the **FBK** database and 362 from the **UPC** one. Note that 131 events are labeled as speech.

4. Experimental Results

The following results were obtained running the proposed system on the acoustic event database described above. The system was characterized by having 512 mixtures in each GMM and 400 centroids in each SVM component. Radial Basis Function (RBF) was used as the Kernel function. The parameter which determines the RBF width was set to 0.5.

The experimental results referred to the **SD** and **SI** scenarios are reported in tables 1 and 2, respectively. Observing the results, in general the **GMM** / **SVM** fusion outperforms each individual classifier performance in both scenarios. In particular, as far as the minimum CER is concerned, one can notice that for a fixed classification technique there is always a single microphone system performing better than the system combining all microphones through **MV** or **ML**, but the microphone combination yields always better results than the average case.

The performance difference between **ML** and **MV** method depends on the scenario. **MV** outperforms **ML** for **SD**, while **ML** outperforms **MV** for **SI**. One possible explanation is that in the **SI** scenario, the universal model is trained with data from both sites, hence more heterogeneous. Thus, the **ML** strategy may be more reliable.

Overall, the system demonstrated robustness even under the more challenging **SI** scenario, showing comparable results as in **SD**. The performance loss from **SD** to **SI**, in particular for the **UPC** case, is due to the fact that **FBK** and **UPC** training materials are unbalanced (in the **FBK** case there are more microphones than in the **UPC** case). Moreover some events of the same type, as mentioned in the introduction, have heterogeneous acoustic
properties, which can cause a larger variance of the feature statistical distribution.

Finally, Tables 3 and 4 report experimental results in terms of speech misses and false alarms. Let us notice that in general the GMM based classifier performs better than the SVM based one in the FBK site, while the SVM is better than GMM based one in the UPC site. Thanks to the combination of the two classifiers the results are now good given the material of both sites.

### 5. Conclusions

This paper addressed the problem of acoustic event classification proposing a new technique that exploits a DMN and the combination of GMM and SVM. The proposed system was tested under two scenarios characterized by different trained models and testing conditions. Based on the use of real-world data, the experimental work showed that the system performs always better than each individual based classifier one. It is better on average compared with each single microphone based system, also in the more challenging Site Independent task.

![Table 1](image1)

<table>
<thead>
<tr>
<th>SITE</th>
<th>GMM</th>
<th>SVM</th>
<th>FUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>2.24</td>
<td>2.24</td>
<td>1.34</td>
</tr>
<tr>
<td>worse</td>
<td>3.13</td>
<td>7.62</td>
<td>2.69</td>
</tr>
<tr>
<td>average</td>
<td>2.40</td>
<td>5.43</td>
<td>1.95</td>
</tr>
<tr>
<td>MV</td>
<td>2.24</td>
<td>4.03</td>
<td>1.79</td>
</tr>
<tr>
<td>ML</td>
<td>2.24</td>
<td>5.38</td>
<td>2.69</td>
</tr>
</tbody>
</table>

![Table 2](image2)

<table>
<thead>
<tr>
<th>SITE</th>
<th>GMM</th>
<th>SVM</th>
<th>FUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>2.69</td>
<td>2.24</td>
<td>1.34</td>
</tr>
<tr>
<td>worse</td>
<td>4.03</td>
<td>8.07</td>
<td>3.13</td>
</tr>
<tr>
<td>average</td>
<td>3.19</td>
<td>5.43</td>
<td>2.24</td>
</tr>
<tr>
<td>MV</td>
<td>3.58</td>
<td>6.27</td>
<td>2.24</td>
</tr>
<tr>
<td>ML</td>
<td>3.13</td>
<td>6.27</td>
<td>1.79</td>
</tr>
</tbody>
</table>

![Table 3](image3)

<table>
<thead>
<tr>
<th>SITE</th>
<th>GMM</th>
<th>SVM</th>
<th>FUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>0</td>
<td>0</td>
<td>0/0</td>
</tr>
<tr>
<td>worse</td>
<td>0.07</td>
<td>8.95</td>
<td>0/0.33</td>
</tr>
<tr>
<td>average</td>
<td>0.03</td>
<td>5.07</td>
<td>0/0.06</td>
</tr>
<tr>
<td>MV</td>
<td>0</td>
<td>0</td>
<td>0/0</td>
</tr>
<tr>
<td>ML</td>
<td>0</td>
<td>0</td>
<td>0/0</td>
</tr>
</tbody>
</table>

![Table 4](image4)

<table>
<thead>
<tr>
<th>SITE</th>
<th>GMM</th>
<th>SVM</th>
<th>FUSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>0/0</td>
<td>0</td>
<td>0/0</td>
</tr>
<tr>
<td>worse</td>
<td>0.07</td>
<td>8.95</td>
<td>0/0.33</td>
</tr>
<tr>
<td>average</td>
<td>0.03</td>
<td>5.07</td>
<td>0/0.06</td>
</tr>
<tr>
<td>MV</td>
<td>0</td>
<td>0</td>
<td>0/0</td>
</tr>
<tr>
<td>ML</td>
<td>0</td>
<td>0</td>
<td>0/0</td>
</tr>
</tbody>
</table>

6. **References**


