# AN MFCC-GMM APPROACH FOR EVENT DETECTION AND CLASSIFICATION

L. Vuegen[a,c,d] B. Van Den Broeck[b,c,d] P. Karsmakers[b,c,d] J. F. Gemmeke[a] B. Vanrumste[b,c,d] H. Van hamme[a]
[a]*ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium*
[b]*ESAT-SISTA, KU Leuven, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium*
[c]*iMinds, Future Health Department, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium*
[d]*MOBILAB, TM Kempen, Kleinhoefstraat 4, 2440, Geel, Belgium*

## ABSTRACT

This abstract explores Gaussian Mixture Models (GMM) estimated from Mel Frequency Cepstral Coefficients (MFCCs) for acoustic event detection and classification. To limit the impact of silence, a shared background model is used. An average F-score of 48% for the office life subtask is obtained. However, the analysis reveals that the proposed method has difficulties to cope with the large intra-class variations (e.g. time durations, dynamic range, characteristic sounds) in the provided dataset.

***Index Terms*—** Acoustic Event Detection, Mel-Frequency Cepstral Coefficients, Gaussian Mixture Models.

## 1. INTRODUCTION

The use of Gaussian Mixture Models is a well-known approach in the domain of speech- and speaker recognition applications. Research shows that this technique can achieve competitive results especially in the conjunction with auditorily motivated features (e.g. MFCCs) [1]. Therefore, this work will examine the use of an MFCC-GMM baseline acoustic event detector and classifier on the publicly available database from the IEEE-AASP challenge. The remainder of this abstract is organized as follows: feature extraction and training phase will be briefly discussed in section 2. Section 3 handles the used event detector and classifier. The executed experiments and obtained results from the subtasks office life and office synthetic are given in section 4. Finally, the conclusions are discussed in section 5.

## 2. FEATURE EXTRACTION AND TRAINING

Figure 1 is a flowchart of the feature extraction and training phase which has been used during this work. This process starts by iteratively loading the waveform (.wav) files from each event. Next, the corresponding MFCC features, including the first and second derivatives, are computed.

Labeling the extracted features into the actual event features and background features happens in two stages. First, the provided annotation files from the 2 different annotators are used to locate the event features. The earliest onset mark of both annotators is used as onset and the latest offset mark of both is used as offset in order to reduce the probability of labeling event features as belonging to the background. Next, a threshold on the first MFCC-coefficient, further denoted as $C_0$, is applied to remove the within-event silence (e.g. silence between 2 phone rings) from the remaining event data. Frames with a $C_0$ below a threshold can be assumed as low-energetic and are therefore added to the background features.

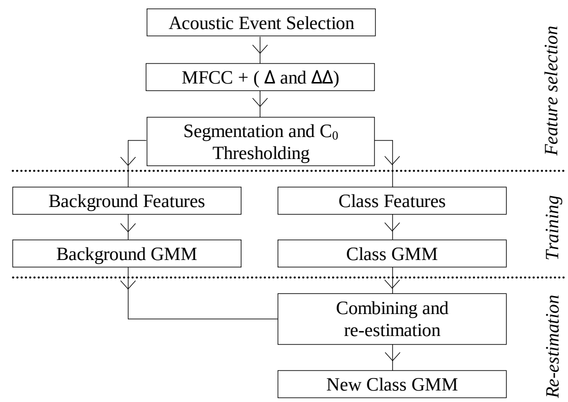


Figure 1: Flowchart of feature extraction and training phase.

The training process starts by estimating a shared background GMM and all the class GMMs (16 in total) on basis of the background data and class event data respectively. Training is achieved by applying the Expectation-Maximization (EM) algorithm as explained in [1].

Finally, the class GMMs will be combined and re-estimated in the presence of the background model where the latter is not re-estimated. This is preferred over relying on each of the class-GMMs to model the silence frames independently. This way, the shared background GMM will produce the same score for each of the class assumptions and hence the impact of silence frames on the model likelihoods will be minimized. EM reestimation of the weight of a class Gaussian in a mixture is achieved by replacing the standard expression for its posterior with:

$$p(i|x_t,\lambda) = \frac{\lambda_i N(x_t|\mu_i,\Sigma_i)}{\sum_{j=1}^{M}\lambda_j \mathrm{N}(x_t|\mu_j,\Sigma_j) + \lambda_{prior}\sum_{k=1}^{N}\lambda_{back,j} N(x_t|,\mu_{back,j},\Sigma_{back,j})}, \quad (1)$$

Compared to the standard EM-algorithm, (1) is expanded with an additional term in the denominator, i.e. the contribution of the background GMM to the data likelihood. The weighting prior $\lambda_{prior}$ defines the amount of probability mass that is assigned to the background model in the maximization step and is re-estimated in (2). .

$$\lambda_{prior} = 1 - \left(\sum_{j=1}^{M}\lambda_j\right), \quad (2)$$

$\lambda_{prior}$ is initialized with a class-independent constant.

In the experimental setup will the influence of the proposed technique examined by combining the shared background GMM with the class GMMs. This by a) applying the adapted EM-formula with an initial $\lambda_{prior}$ of 0.2 and b) a linear combination of the background and class GMMs with a ratio 1/5 respectively.

## 3. EVENT DETECTOR AND CLASSIFIER

The event detector and classifier for the subtask office life starts by extracting the MFCC features (including 1st and 2nd derivative) from the acoustic event script (see Figure 2). A posteriorgram is computed by comparing these features with both the estimated background model and all the class GMMs. Next, the posteriors from each class are moving averaged filtered with a window size depending on the minimum class duration observed in the training dataset. This smoothens the class posteriors and takes the minimum occurring time duration of each class more or less into account.

Detecting events in the office life subtask is based on $C_0$ thresholding. It can be assumed that an event has occurred when the value of $C_0$ was above a predefined threshold during a certain period of time. The values of $C_0$ and minimum time duration are mentioned further in the next section.

Finally, classification of the detected events is done by determining which GMM model produces the highest averaged a-posteriori score. In case the detected event is classified as background, it will be neglected and therefore removed as a detected event.

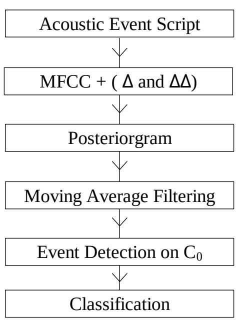


Figure 2: Flowchart of the event detector and classifier.

## 4. EXPERIMENTS AND RESULTS

In order to determine the performance of the proposed method, the following parameters are examined in function of the averaged event based F-score: a) resampling to a lower sample frequency i.e. 16kHz, b) influence of the first and second derivative, c) the number of Gaussians and d) re-estimation or just linear combination both with a $\lambda_{prior}$ of 0.2.

Reproducible research shows that window sizes equal to a third of the minimum time duration (with a maximum of 1sec.) seen in the training set for each event class provides the best results. The value of $C_{0,\min}$ on the other hand is estimated by applying an exponential filter (2) on the first MFCC-coefficient $C_0$. This technique allows changing the threshold in function of the noise in the loaded script.

$$C_{0,\min}(n) = (1-\alpha)C_{0,\min}(n-1) + \alpha C_0(n) + margin, \tag{2}$$

The parameter α in (2) is a constant between 0 and 1 and determines the rate of adaption. In our setup 2 different adaptation parameters, denoted as $\alpha_{short}$ and $\alpha_{long}$ with as value 1/50 and 1/500 respectively, were used. The following constraint (3) determines which adaptation parameter must be used.

$$\begin{cases} C_0(n) \le C_{0,\min}(n-1) \rightarrow \alpha = \alpha_{short} \\ C_0(n) > C_{0,\min}(n-1) \rightarrow \alpha = \alpha_{long} \end{cases}, \tag{3}$$

In case when an event occurs it can be assumed that $C_0(n)$ is larger than $C_{0,min}(n-1)$. This implies that the slow adaptation parameter will be used and therefore $C_{0,min}(n)$ remains almost the same as $C_{0,min}(n-1)$. On the other hand, when no event occurs, $C_0(n)$ will be smaller than $C_{0,min}(n-1)$. This allows the use of a fast adaptation parameter such that $C_{0,min}(n)$ follows the noise floor. The used margin in (2) is simply the standard deviation of the estimated $C_{0,min}$ over the entire script.

Figure 3 shows the obtained averaged F-scores with previous parameter settings and the following observations are made:

- Applying a down sampling to 16kHz has almost no effect on the averaged F-score. A possible explanation is that the higher frequency bands doesn't contain sufficient characteristic information of the occurring event.
- The usage of the first and second derivative has a small positive effect on the F-score.
- Applying the proposed re-estimation algorithm does not increase the F-score sufficiently.

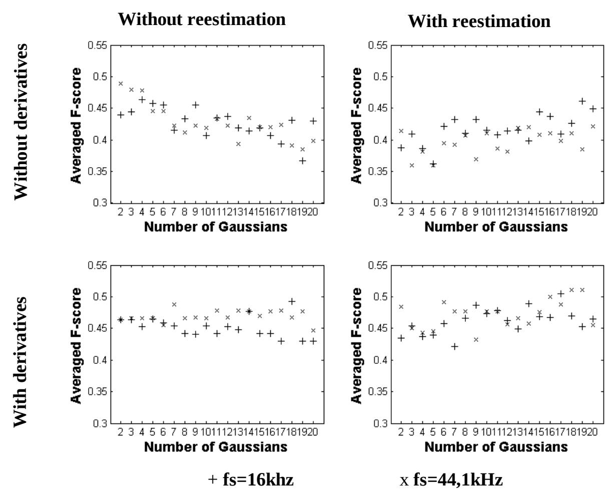


Figure 3: Averaged results of the event based F-score

Table 1 on the other hand gives the associated evaluation metrics corresponding to the highest achieved averaged F-score (in Figure 3). i.e. 51.13% and 49.24% for with and without reestimation respectively. These F-scores occurs both when the derivatives are included and 18 Gaussians were used. The corresponding sample frequency is 44.1kHz for with reestimation and 16kHz for without and as one can see, no major differences occurs between both methods.

Table 1: Results on the Office Live Dataset for various metrics.

| Metric | Evaluation Method | | | | | |
|---|---|---|---|---|---|---|
| | Event Based | | Class-Wise Event Based | | Frame Based | |
| Re-estimation | Y | N | Y | N | Y | N |
| R | 43.22 | 41.82 | 43.04 | 44.01 | 45.90 | 43.75 |
| P | 63.49 | 60.67 | 42.69 | 40.60 | 73.39 | 70.53 |
| **F-score** | **51.13** | **49.24** | **39.77** | **39.31** | **56.28** | **53.89** |
| AEER | 1.06 | 1.12 | 0.94 | 0.89 | 0.88 | 0.93 |
| Offset R | 35.68 | 35.23 | 35.26 | 37.20 | - | - |
| Offset P | 52.65 | 51.66 | 36.88 | 36.43 | - | - |
| Offset F-score | 42.28 | 41.66 | 33.58 | 34.37 | - | - |
| Offset AEER | 1.29 | 1.32 | 1.18 | 1.10 | - | - |

Table 2 and 3 gives the corresponding F-scores of the office synthetic task. The same parameters were used as in Table 1 and as one can see, the achieved results drop considerably, even for the easiest combination, i.e. a SNR of 6dB and the lowest degree of overlapping. One of the reasons of a lower score is that the proposed detection algorithm expects only 1 event when an event is detected during a certain time span. Second, research shows that GMMs are extremely sensitive to distortion by noise. Even the smallest amount of noise can cause an enormous drop of performance which is shown in [4].

Table 2: F-scores on the Office Synthetic Dataset (with re-estimation).

| | | Density | SNR -6 | SNR 0 | SNR 6 |
|---|---|---|---|---|---|
| Evaluation method | Event based | low | 15.38 | 0.00 | 0.00 |
| | | medium | 0.00 | 0.00 | 0.00 |
| | | high | 0.00 | 0.00 | 4.71 |
| | Class-wise event Based | low | 16.67 | 0.00 | 0.00 |
| | | medium | 0.00 | 0.00 | 0.00 |
| | | high | 0.00 | 0.00 | 3.81 |
| | Frame Based | low | 11.59 | 0.38 | 0.00 |
| | | medium | 0.71 | 0.00 | 0.03 |
| | | high | 0.79 | 0.85 | 7.35 |

Table 3: F-scores on the Office Synthetic Dataset (without re-estimation).

| | | Density | SNR -6 | SNR 0 | SNR 6 |
|---|---|---|---|---|---|
| Evaluation method | Event based | low | 15.38 | 0.00 | 0.00 |
| | | medium | 0.00 | 0.00 | 0.00 |
| | | high | 0.00 | 0.00 | 0.00 |
| | Class-wise event Based | low | 16.67 | 0.00 | 0.00 |
| | | medium | 0.00 | 0.00 | 0.00 |
| | | high | 0.00 | 0.00 | 0.00 |
| | Frame Based | low | 11.48 | 0.38 | 0,00 |
| | | medium | 0.69 | 0.00 | 0.17 |
| | | high | 0.79 | 1.89 | 3.26 |

## 5. DISCUSSION

The overall performance of the proposed method were not so promising as hoped, especially for office synthetic subtask. The most obvious explanation is that Gaussian Mixture Models have difficulties to cope with a) the large variation in the characteristic sounds of some classes (e.g. phone and alert) and b) the relative low amount of training examples in the dataset. This results in a harder classification problem and therefore reducing the accuracy of the classifier. Besides, the large variation in time duration and energy increases the difficulty of the detection task and therefore also decreasing overall the performance of the system.

## 6. ACKNOWLEDGMENT

This work was performed in the context of following projects: ALADIN (IWT-SBO project contract 100049), IWT doctoral scholarships (contract 111433 and 121565) and FallRisk. The iMinds FallRisk project is cofounded by iMinds (Interdisciplinary Institute for Technology), a research institute founded by the Flemisch Government. Companies and organizations involved in the project are COMmeto, Televic Healthcare, TP Vision, Verhaert and Wit-Gele Kruis Limburg, with project support of IWT.

## 7. REFERENCES

[1] D. A. Reynolds and R.C. Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, in *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83, 1995.

[2] A. Mesaros, T. Heittola, A. Eronen, And T. Virtanen, “Acoustic Event Detection In Real Life Recordings”, in *the 2010 European Signal Processing Conference*, 2010.

[3] S. Ntalampiras, “A Novel Holistic Modelling Approach for Generalized Sound Recognition”, in *IEEE Signal Processing Letters*, Vol. 20, No. 2, February 2013.

[4] L. Vuegen, P. Karsmakers And B. Vanrumste, “Comparitive Study of Sound Recognition Algorithms”, in *K.H. Kempen proceedings of Master Theses*, June 2012.