A Fast One-Pass-Training Feature Selection Technique for GMM-based Acoustic Event Detection with Audio-Visual Data

Taras Butko and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{taras.butko, climent.nadeu}@upc.edu

Abstract

Acoustic event detection becomes a difficult task, even for a small number of events, in scenarios where events are produced rather spontaneously and often overlap in time. In this work, we aim to improve the detection rate by means of feature selection. Using a one-against-all detection approach, a new fast one-pass-training algorithm, and an associated highly-precise metric are developed. Choosing a different subset of multimodal features for each acoustic event class, the results obtained from audiovisual data collected in the UPC multimodal room show an improvement in average detection rate with respect to using the whole set of features.

Index Terms: acoustic event detection, feature selection, hill-climbing approach, hidden Markov models, one-against-all strategy, GMMs

1. Introduction

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in the signals that are captured by one or several microphones. In AED, like in similar application areas, we face the problem of the large number and variety of features proposed in the literature [1]. In fact, there are many features which exploit acoustic content such as subband energies computed in short-time windows, time evolution parameters, modulation spectrum, level of harmonicity, etc. which are used [2][3]. Although in speech recognition the MFCC features (or alternative features which have a lot in common with them) became the de-facto standard for front-ends in many applications, the situation in AED is not so clear yet. Very often authors do not present strong or clear arguments in favor of a particular feature set they propose, and the final decision about feature subset selection is mainly based on their prior knowledge. For instance, for music detection and segmentation, features based on the harmonicity of the waveform are preferable [4], while for classification of generic sounds the features which model the spectral envelope are widely used [2]. Indeed, the problem is even more acute when other types of data, like video data, are used besides audio.

Recently, we used for AED a combination of standard ASR features together with a set of "perceptual" features [3]. Posteriorly, in order to enhance the detection of particular sounds, new features coming not only from the audio modality but also from video were proposed [5]. In that work, video features improved the detection of all acoustic events (AE) while the features coming from an acoustic localization system improved accuracy only for some of them. These results meant an additional motivation for us to perform feature selection in order to find the best feature set for each particular class of interest.

Actually, feature selection plays a central role in the tasks of classification and data mining, since redundant and irrelevant features often degrade the performance of classification algorithms [6]. It is worth mentioning that the system that got the highest accuracy in the last AED international evaluation campaign (CLEAR'07) [7] used an Adaboost feature selection algorithm to improve the baseline detection rate [8]. Unlike in that work, where the authors compare different feature sets of the same size, the main objective in our current feature selection work is to filter out features from the initial feature set which are redundant and even harm the detection accuracy.

In this paper, we propose a fast feature selection technique that avoids retraining of acoustic models during the evaluation of the candidate feature set. The conventional hill-climbing approach is used as a searching strategy to conduct the feature selection. We have also developed a new metric to evaluate a candidate feature set that overcomes the problem of insufficient number of AE instances in the database. A GMM-based AED system is employed which is composed of binary detectors and uses a one-against-all strategy [5][9] for training and testing.

2. AED with one-against-all detection approach

In our current work we consider 12 classes of AEs which naturally occur in meeting-room environments, like in [4], [5], [7] and [8]: "Door knock", "Door open/slam", "Steps", "Chair moving", "Spoon/cup jingle", "Paper work", "Key jingle", "Keyboard typing", "Phone ring", "Applause", "Cough", and "Speech".

Since we employ a one-against-all detection strategy for AED, only two models are used for each AE, which will herewith be called "Class" and "non-Class". The first model is trained using the signals coming from one class of interest, while the second model is trained using the rest of signals. In total, 12 binary detectors working in parallel are needed to perform detection of all AEs [5].

Gaussian Mixture Models (GMMs) are used, with continuous densities, 16 components, and diagonal covariance matrices. We consider the AE "Class" as a minor in time with respect to the "non-Class". This assumption implies an asymmetry between the two classes of AEs. An example of audio waveform with ground truth labels is depicted in Fig. 1.

Given a sequence of observed acoustic vectors $X = \{x_1, x_2, \ldots, x_T\}$, the likelihood that corresponds to the sequence of AEs $W = w_1, w_2, \ldots, w_M$ is:

$$P(X|W) = P(\{x_1, x_2, \ldots, x_T\}|w_1, w_2, \ldots, w_M) \quad (1)$$

Denoting by *i(t)={class, non-class}* the model associated to frame t, with observation vector x_t, and assuming that each

frame is independent of every other frame, the logarithm of the probability $P(X|W)$ can be expressed as a sum:

$$Q(X|W) = \log(P(X|W)) = \sum_{t=1}^{T} Q(x_t|w_{i(t)}) \quad (2)$$

where T is the total number of frames in the sequence X, and $Q(x_t|w_{i(t)})$ is the local log-likelihood of the frame t given the AE model from which observation vector x_t came.

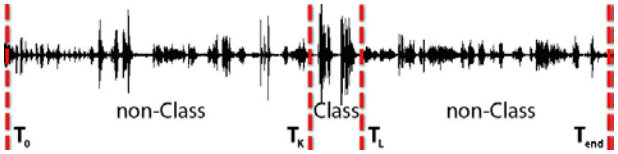

Fig.1. *Audio waveform with ground truth labels.*

For illustration, we consider the waveform in Fig. 1. The problem of detecting the AE in the interval $[T_K...T_L]$ with observation sequence $X = [x_K, ..., x_L]$ can be formulated as a hypothesis test, being H_0 "Class" in the interval $[T_K...T_L]$ is not detected, and H_1 "Class" in this interval is detected. The necessary condition for detecting "Class" (hypothesis H_1) in that interval is the log-likelihood ratio (LLR) of "Class" and "non-Class" models exceeds the given non-negative threshold value P:

$$Q(X|w_{class}) - Q(X|w_{non-class}) > P \quad (3)$$

Although $P=0$ is the natural choice, $P>0$ is used to avoid false alarms in the case when the LLR is too small.

In Fig. 2 we display the LLRs for all AEs in the labeled development database. Squares correspond to the "Class" AE instances ("Chair moving" in our case) while crosses correspond to "non-Class" instances. Negative values of the LLR are substituted by 0. As we can see from that plot, most of the "Class" instances have higher LLR values than the "non-Class" ones. We consider the parameter P as a threshold (the horizontal line in Fig. 2), and we selected $P=270$ for illustrative purposes. The ith AE instance is detected as "Class" if its LLR ΔL_i is above P, otherwise it is detected as "non-Class". Thus, all "Class" instances (squares in Fig. 2) below the P line are misses, and all "non-Class" instances (crosses) are false alarms. All AE instances that are around the horizontal line P are very sensitive to the selection of the value P.

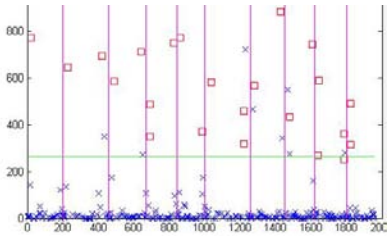

Fig.2. *The LLRs corresponding to "Chair moving" class.*

3. Feature subset selection

A search of the optimal feature set requires a state space, an initial state, a termination condition, and a search engine together with an evaluation (objective) function [6]. The state space includes all possible combinations of features, and the search is terminated after finding a feature set with the highest value of the evaluation (objective) function.

A natural objective function is the empirical average loss, defined in [10] as the total count of classification errors in the development database. However, it is not adequate to use this objective function for feature selection in our case since it is not sensitive enough to small changes of detection accuracy caused by additional features. For illustration, we present in Table 1 the relation between the number of features and the number of errors for the "Chair moving" AE. Notice that the metric is not useful to decide about selecting the features 5th, 6th and 7th since the number of errors does not change when those three features are included in the feature vector.

Table 1. *Relation between the number of features and the number of errors for the "Chair moving" AE.*

Number of features	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**...**	**32**
Number of errors	*25*	*9*	*4*	*2*	*2*	*2*	*2*	…	*8*

3.1. Developing of a new, more precise, metric

We want a *soft detection accuracy* metric (*SDA*) that counts not just correct and incorrect detections of AEs, but uses the degree of correctness or confidence of the decisions given by ΔL_i. In fact, in terms of the *one-against-all* detection strategy discussed in the previous section, a high LLR ΔL_i with respect to the threshold P of the ith AE instance corresponds to a high confidence of the decision that the given signal segment belongs to the "Class" AE. If the LLR has a value close to P, the confidence of the decision made by the AED system is very low.

The *SDA* metric is obtained by averaging the scores Ω_i obtained for all the segments that have "Class" ground truth label, where the score Ω_i lies in the range [-1 +1] and it is computed from the LLR ΔL_i with the expression:

$$\Omega_i = f(\Delta L_i, P) = \begin{cases} -\dfrac{P}{\Delta L_i} + 1, & \text{if } \Delta L_i > P \\ \dfrac{\Delta L_i}{P} - 1, & \text{otherwise} \end{cases} \quad (6)$$

In our experiments, to compute the metric *SDA(**ΔL**)*, where ***ΔL*** is a vector formed by the LLR values ΔL_i of all the "Class" instances, P is selected in such a way that the numbers of misses and false alarms are equal (equal error rate).

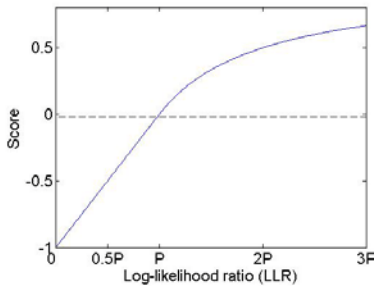

Fig. 3. *The normalization function f(.) that is used in SDA metric computation.*

3.2. Conventional feature selection approach

The conventional hill-climbing feature selection approach is widely used when the feature set is relatively small. In this approach, the initially chosen feature vector is iteratively enlarged by adding a single feature that maximally increases the performance score. The process is stopped when no more performance improvement can be achieved by adding new features. The main disadvantage of that approach is its large

computational load. Indeed, during each iteration, the GMMs have to be retrained with the EM algorithm. In the following, we propose a new way of evaluating the feature set that does not require retraining of GMM models at each iteration.

3.3. One-pass-training feature selection approach

According to (2) the log-likelihood of the AE instance from the interval $[T_K \ldots T_L]$ given the observation vector X and AE model w is represented as:

$$Q(X|w) = \sum_{t=T_K}^{T_L} Q(x_t|w) \tag{7}$$

According to (7) the log-likelihood of any AE instance is estimated as the accumulation of the local log-likelihoods of the observation vector given the AE model. The LLR for the ith AE instance is obtained as:

$$\begin{aligned} &\Delta L_i(X) = Q(X|w_{class}) - Q(X|w_{non-class}) = \\ &(\sum_{t=T_K}^{T_L} Q(x_t|w_{class}) - \sum_{t=T_K}^{T_L} Q(x_t|w_{non-class}) = \sum_{t=T_K}^{T_L} \Delta L_i(x_t) \end{aligned} \tag{8}$$

where $\Delta L_i(x_t)$ is the local LLR corresponding to the frame t. Taking into account that the observation vector is modeled via GMM, the local log-likelihood is computed as:

$$\begin{aligned} &Q(x_t|w) = \log(\sum_{i=1}^{M} \alpha_i N(\overline{x_i}, \Sigma_i)) \\ &N(\overline{x_i}, \Sigma_i) = \frac{1}{2\pi^{N/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x_t-\overline{x_i})^T \Sigma_i^{-1} (x_t-\overline{x_i})} \end{aligned} \tag{9}$$

where M is the total number of mixture components and N is the total number of features. $\overline{x_i}$ and Σ_i are the mean vector and the covariance matrix of the ith mixture component, respectively.

The contribution of each feature to the local log-likelihood is not easy to estimate due to the sum of the logarithms in (9). Therefore we approximate the local log-likelihood estimation by the log-likelihood obtained from the single dominant component in the mixture:

$$\begin{aligned} &Q(x_t|w) \approx \tilde{Q}(x_t|w) = \log(\alpha_k N(\overline{x}_k, \Sigma_k)) \\ &k = \arg\max_i (\alpha_i N(\overline{x}_i, \Sigma_i)) \end{aligned} \tag{10}$$

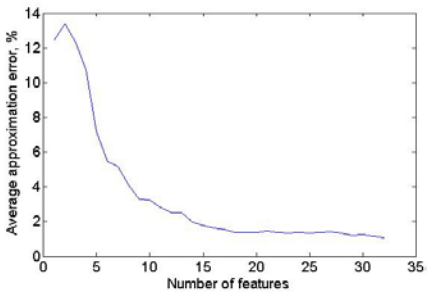

Fig. 4. *The average relative difference between LLRs estimated from (9) and (10).*

The average error of the approximation (10) is depicted in Fig. 4. In the experiment the LLRs from all "Class" AE instances ΔL_i are obtained using (9) and (10) and the average relative difference between these 2 values (in %) is depicted along the vertical axis as a function of the number of features using the development part of the database described in Section 4. Notice in Fig. 4 that the relative difference between these two values is very small (less than 5%) provided that more than 5 features are used.

Taking into account the approximation of the log-likelihood estimation (10) and assuming diagonal covariance matrix Σ_k, we further obtain:

$$\begin{aligned} &\tilde{Q}(x_t|w) = \log(\alpha_k \frac{1}{2\pi^{N/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x_t-\overline{x}_k)^T \Sigma_k^{-1} (x_t-\overline{x}_k)}) \\ &= \log\frac{1}{2\pi^{N/2}} + \log\alpha_k + \sum_{i=1}^{N} \log\frac{1}{\sigma_i} e^{-\frac{(x_i-\overline{x}_i)^2}{2\sigma_i^2}} \end{aligned} \tag{11}$$

where σ_i^2 are the diagonal elements of the matrix Σ_k.

Thus the local log-likelihood can be decomposed into the sum of the following terms: a constant term, the logarithm of the mixture weight, and the sum of the components that can be considered as the contribution coming from each feature. Thus the local LLR $\Delta L_i(x_t)$ is computed as:

$$\Delta L_i(x_t) = \log(\tilde{Q}(x_t|w_{class})) - \log(\tilde{Q}(x_t|w_{non-class}) = \sum_{f=0}^{N} \delta_f(x_t) \tag{12}$$

where $\delta_0(x_t)$ is the log difference between the weights from the dominant mixtures corresponding to "Class" and "non-Class" GMMs, and $\delta_f(f>0)$ is the contribution, in terms of LLR, from the fth feature. Finally, the expression (8) can be rewritten as:

$$\Delta L_i(X) = \sum_{t=T_K}^{T_L} \Delta L_i(x_t) = \sum_{t=T_K}^{T_L} \sum_{f=0}^{N} \delta_f(x_t) \tag{13}$$

The modified feature selection approach consists of the following steps:

- Perform an initial training of the GMM models using all available N features.
- Compute the local LLRs $\delta_f(x_t)$ for each frame t of the input signal and for each feature f from the set of N features.
- Perform the hill-climbing search to find a subset of features that maximize $SDA(\Delta \boldsymbol{L})$. The vector $\Delta \boldsymbol{L}$ is obtained with the expression (13) and the values $\delta_f(x_t)$ are taken from the previous step.

Note, the proposed approach doesn't require the EM training of GMM models at each iteration. This means a large save in terms of computational load when the number of features is not small.

4. Experiments

In order to assess the performance of the proposed feature selection approach, the subset of isolated AEs from a recently recorded multimodal database [5] was used. The video signals were recorded with 5 calibrated cameras at pixel resolution 768x576 and 25 fps. Audio signals were collected from 6 T-shaped 4-microphone clusters, and sampled at 44.1 kHz. All sensors were synchronized. In the recorded scenes, 4 subjects performed several times the 12 AEs employed in this work, adding up to around 100 instances for every AE, and 2 hours. The recorded dataset has been divided into three parts for training (to create GMM models), developing (to perform feature selection) and testing (to present the evaluation results).

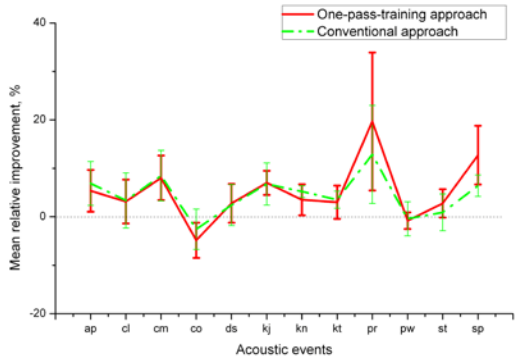

Fig. 5. *Comparison between the conventional and the one-pass-training features selection techniques.*

In our work, we started with an initial feature set that consists of 16 frequency-filtered (FF) logfilter -bank energies with their first time derivatives (in total, 32 features). We also added features coming from the acoustic source localization system (1 additional feature for all classes) and video signals (1 additional feature for 5 classes) [5].

Table 2. *The number of selected features for each AE using the one-pass-training approach.*

AEs	Number of selected features				
	FF-based		Loc	Video	**Total**
	Static	Dynamic			
Door knock	9	10	1	---	**20**
Door open/slam	10	13	1	1	**25**
Steps	10	11	1	1	**23**
Chair moving	9	11	1	1	**22**
Cup clink	7	8	1	---	**16**
Paper work	11	9	1	1	**22**
Key jingle	9	9	0	---	**18**
Keyboard typing	8	8	0	1	**17**
Phone ring	7	6	0	---	**13**
Applause	8	9	1	---	**18**
Cough	8	10	1	---	**19**
Speech	7	10	0	---	**17**

Fig. 5 summarizes the mean relative improvement obtained by the system based on selected features with respect to the system that uses the whole feature vector. The dashed line corresponds to the results obtained with the baseline conventional approach, and the solid curve corresponds to the one-pass-training approach explained in Section 3. The standard deviation is plotted with vertical lines. It has been calculated from 8 scores, which were obtained by using different combinations of partitions of the database for training, development and testing. The results in Fig. 5 correspond to the testing part of the database (unseen data that is not used during the feature selection process). According to them, the detection rate increases for all classes, except "cough" and "paper work", by using any of the two feature selection techniques. Observe that the new feature selection technique does not work much differently from the conventional feature selection approach. The average of the mean relative improvement across the AEs (horizontal axis) equals to 4.5% for the conventional technique and 5.0% for the one-pass-training approach achieving the feature compression ratio of 1.67 in later case.

In Table 2, the number of selected features for different categories of features and different classes of AEs is displayed.

We decompose the 32 FF features into 16 static parameters and 16 dynamic parameters. The next two columns correspond to the number of selected features coming from localization and video, respectively. According to that table, both static and dynamic FF-based features contribute to final accuracy. The video features are an important additional source of information for detection of the five AEs for which video features are extracted. The acoustic localization feature was selected for eight AEs, but not for the other four.

5. Conclusions

In this work, by using a fast one-pass-training feature selection approach we have selected the subset of multimodal features that shows the best detection rate for each class of AEs, observing an improvement in average accuracy with respect to using the whole set of features.

A new, more precise, metric is proposed to perform feature selection, which overcomes the problem of insufficient number of AE instances in the database. By accumulating the contribution of each feature to the LLR, the developed one-pass-training algorithm requires much less computational time for doing feature selection than the conventional approach.

Future work lines aim at considering a larger set of initially chosen features, and comparing the proposed feature selection approach with other dimensionality reduction techniques such as LDA or PCA. We are also motivated to compare our method with existing filter methods in terms of speed and efficiency.

6. Acknowledgements

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government.

7. References

[1] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project", CUIDADO Project Report, 2003.

[2] D. Li, I.K. Sethi, N. Dimitrova and T. McGee, "Classification of general audio data for content-based retrieval", Pattern Recognition Letters 22, pp. 533–544, 2001

[3] A. Temko, C. Nadeu, "Acoustic Event Detection in Meeting-Room Environments", Pattern Recognition Letters, v. 30/14, pp 1281-1288, Elsevier, 2009

[4] S. H. Srinivasan, M. Kankanhalli, "Harmonicity and dynamics-based features for audio", IEEE Proc. ICASSP, pp. 321-324, 2004

[5] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion", in Proc. Interspeech, 2009.

[6] R. Kohavi, G. John, "Wrappers for feature subset selection", Artificial Intelligence, Spec. Issue on Relevance, vol. 97, pp. 273-324, 1997

[7] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification systems", in Multimodal Technologies for Perception of Humans, LNCS, vol. 4122, Springer, 2007.

[8] X. Zhou, X. Zhuang, M. Lui, H. Tang, M. Hasgeawa-Johnson, T. Huang, "HMM-Based Acoustic Event Detection with AdaBoost Feature Selection", in Multimodal Technologies for Perception of Humans, LNCS, vol. 4625, Springer, 2008.

[9] R. Rifkin, A. Klautau, "In defense of One-Vs-All Classification", Journal of Machine learning Research, vol. 5, pp.101-141, 2004.

[10] R. O. Duda, P. E. Hart, D.G. Stork, "Pattern classification", New York, John Wiley-Interscience, 2001.