Fusion of Audio and Video Modalities for Detection of Acoustic Events

Taras Butko[1,2], *Andrey Temko*[1,2], *Climent Nadeu*[1,2], *and Cristian Canton*[1]

[1]Department of Signal Theory and Communications
[2]TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
{butko, temko, climent, ccanton}@gps.tsc.upc.edu

Abstract

Detection of acoustic events (AED) that take place in a meeting-room environment becomes a difficult task when signals show a large proportion of temporal overlap of sounds, like in seminar-type data, where the acoustic events often occur simultaneously with speech. Whenever the event that produces the sound is related to a given position or movement, video signals may be a useful additional source of information for AED. In this work, we aim at improving the AED accuracy by using two complementary audio-based AED systems, built with SVM and HMM classifiers, and also a video-based AED system, which employs the output of a 3D video tracking algorithm to improve detection of steps. Fuzzy integral is used to fuse the outputs of the three classification systems in two stages. Experimental results using the CLEAR'07 evaluation data show that the detection rate increases by fusing the two audio information sources, and it is further improved by including video information.

Index Terms: acoustic event detection, fuzzy integral, multimodality, support vector machines, hidden Markov models, video 3D tracking

1. Introduction

The detection of the acoustic events (AEs) that are naturally produced in a meeting room may help to describe the human and social activity that takes place in it. Additionally, the robustness of automatic speech/speaker recognition systems may be increased by a previous detection of the non-speech sounds lying in the captured signals. Recently, several papers have reported works on acoustic events detection (AED) for different meeting-room environments and databases, e.g. [1] [2] [3]. The CLEAR'07 international evaluations in seminar conditions have shown that AED is a challenging problem. In fact, 5 out of 6 submitted systems showed accuracies below 25%, and the best system got 33.6% accuracy (see [2] [3] for results, databases and metrics). The single main factor that accounts for those low detection scores is the high degree of overlap between sounds, especially between the targeted acoustic events and speech.

The overlap problem may be faced by developing more efficient algorithms either at the signal level, at the model level or at the decision level. Another approach is to use an additional modality that is less sensitive to the overlap phenomena present in the audio signal. In this work we aim at using two different audio-based detectors and including video information using a fusion approach. Actually, the above mentioned seminar databases include both video and audio information from several cameras and microphones hanged on the walls of the rooms.

Actually, the information about movements and positions of people in a meeting room may be correlated with AEs that take place in it. For instance, the sources of events such as "door slam" or "door knock" are associated to given positions in the room; other events such as "steps" and "chair moving" are accompanied with changes of position of participants in the meeting room. Motivated by the fact that the "steps" sound class accounted for almost 35% of all AEs in the CLEAR'07 evaluation database, in this work we use video 3D tracking information aiming to improve the representation and detection of that particular class and, consequently, to improve the AED accuracy of the whole set of 12 targeted AEs.

In our work, two diverse audio-based AED systems, built respectively with SVM and HMM, and a VIDEO-based "steps" detection system, are fused by means of Fuzzy Integral (FI) [4] [5], a fusion technique which is able to take into account the interdependences among information sources. The reported experiments are carried out with CLEAR'07 evaluation data which consist of several interactive seminars.

The rest of this paper is organized as follows: Section 2 describes video and audio-based systems of AED. Fuzzy integral is described in Section 3. Section 4 presents experimental results and discussions, and Section 5 concludes the work.

2. Acoustic Event Detection systems

In this work, detection of AEs is carried out with one VIDEO-based and two audio-based systems. The use of the three AED systems is motivated by the fact that each system performs detection in a different manner. The video-based system uses information about position of people in the room. The HMM-based AED system segments the acoustic signal in events by using a frame-level representation of the signal and computing the state sequence with highest likelihood. The SVM-based system does it by classifying segments resulting from consecutive sliding windows. The difference in the nature of the considered detection systems makes the fusion promising for obtaining a superior performance.

2.1. Video-based detection system for the class "steps"

2.1.1. Video 3D tracking algorithm

Person tracking is carried out by using multiple synchronized and calibrated cameras as described in [6]. Redundancy among camera views allows generating a 3D discrete reconstruction of the space being these data the input of the tracking algorithm. A particle filtering (PF) [7] approach is followed to estimate the location of each person inside the room at a given time *t*. In order to keep an affordable computational load, a single PF is assigned to every tracked person and an interaction model among filters is defined.

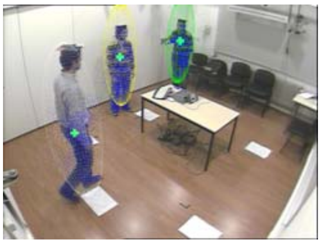


Figure 1. *Position of people in the meeting room provided by the video 3D tracking algorithm*

The combination of the estimated 3D location together with geometric descriptors allows discarding spurious objects such as furniture and a simple classification of the person's pose as standing or sitting. The performance of this algorithm over a large annotated database [6] showed the effectiveness of this approach.

2.1.2. Feature extraction and "steps" detection

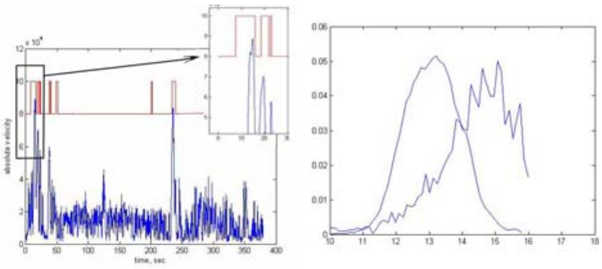
The output of the 3D tracking algorithm is the set of coordinates of all the people in the room, which are given every 40ms. From those coordinates, we have to generate features that carry information correlated with "steps". We assume that information about movements of people is relevant for "steps" detection. The movements of people in the meeting room can be characterized by a velocity measure. In a 2D plane, the velocity can be calculated in the following way:

$$v = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} \qquad (1)$$

where *dx/dt* and *dy/dt* are the values of velocity along *x* and *y* axes, respectively. Those values are calculated using a smoothed derivative non-casual filter *h* applied to the vector of positions of each person in the room. We tried several shapes of the impulse response of the derivative filter; best results were obtained using a linear non-casual filter with the impulse response *h(n) = [-m ... -2 -1 0 1 2 ... m]* (zero corresponds to the current value and *L=2*m+1* is the length of the filter).

Usually more than one person is present in the room, and each person has its own movement and velocity. The maximum velocity among the participants in the seminar is used as a current feature value for "steps"/ "non-steps" detection.

Figure 2 (a) plots the maximum value of velocity among participants for a 6min seminar along with the corresponding ground truth labels. From it we can observe that there is a high degree of correspondence between peaks of velocity and true "steps".



(a) (b)

Figure 2. *Values of the velocity during one development seminar ((a) bottom) and reference "steps" labels ((a), top), and the histograms of log-velocities for "non-steps" ((b), left hump) and "steps" ((b), right hump)*

The normalized histograms of the logarithm of velocity for "steps" and "non-steps" obtained from development seminars are depicted in Figure 2 (b), from which it can be seen that "steps" are more likely to appear with higher values of velocity.

The jerky nature of the "steps" hump results from a more than 10 times scarcer representation of "steps" with respect to "non-steps" in the development database. These two curves are approximated by two Gaussians via EM algorithm. During detection on testing data the final decision for "steps"/ "non-steps" classes is made using the Bayesian rule:

$$P(w_j \mid x) = P(x \mid w_j)P(w_j)\text{, } j=\{1,2\} \qquad (2)$$

where $P(w_1)$ and $P(w_2)$ are prior probabilities for the class "steps" and the meta-class "non-steps" respectively, which are computed using the prior distribution of these two classes in development data and $P(x|w_j)$ are likelihoods given by the Gaussian models.

To have a better detection of "steps" the length *L* of the derivative filter *h(n)* and several types of windows applied on *h(n)* were investigated and the best detection on development data is achieved with a Hamming window of 2 seconds.

2.2. SVM-based AED system

The SVM-based AED system used in the present work is the one that was also used for the AED evaluations in CLEAR 2007 [3] with slight modifications. The sound signal from a single MarkIII array microphone is down-sampled to 16 kHz, and framed (frame length/shift is 30/10ms, a Hamming window is used). For each frame, a set of spectral parameters has been extracted. It consists of the concatenation of two types of parameters: 1) 16 Frequency-Filtered (FF) log filter-bank energies, along with the first and the second time derivatives; and 2) a set of the following parameters: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth. In total, a vector of 60 components is built to represent each frame. The mean and the standard deviation parameters have been computed over all frames in a 0.5sec window with a 100ms shift, thus forming one vector of 120 elements.

SVM classifiers have been trained using 1-vs-1 scheme on the isolated AEs, from two databases of isolated AEs, along with segments from the development data seminars, that include both isolated AEs and AEs overlapped with speech. The MAX WINS (pair-wise majority voting) [8] scheme was used to extend the SVM to the task of classifying several classes. After the voting is done, the class with the highest number of winning two-class decisions (votes) is chosen.

2.3. HMM-based AED system

In this AED detection task, Hidden Markov Models (HMM) are used like in continuous speech recognition, aiming at maximizing the posterior probability of the acoustic event sequence $W=(w_1,w_2, ...,w_M)$, given the observations $O =(o_1, o_2, ..., o_T)$:

$$W_{max}\text{= argmax } P(W|O) = \text{argmax}(P(O|W)P(W)) \qquad (3)$$

where we assume that *P*(*W*) is the same for all event sequences. The HTK toolkit is used [9].

Firstly, the input signal from a single MarkIII-array microphone is down-sampled to 16 kHz, and 13 FF coefficients with their first time derivatives are extracted, using a Hamming window of size 20ms with shift 10ms.

There is one HMM for each acoustic event class, with five emitting states and fully connected state transitions. A similar HMM is used for silence. The observation distributions of the states are Gaussian mixtures with continuous densities, and consist of 9 components with diagonal covariance matrices. The "speech" class is modelled with 15 components as its observation distribution is more complex. Actually, the chosen HMM topology showed the best results on the development data. In the detection phase, which is indeed carried out with the Viterbi algorithm, after temporal segmentation of the signal according to the optimum path, the log-likelihood of each hypothesized AE is computed, since it will be used during the posterior fusion.

3. Fusion of information sources

3.1. Fuzzy integral and fuzzy measure

We are searching for a suitable fusion operator to combine a finite set of information sources $Z=\{1,...,z\}$. Let $D=\{D_1,D_2,...,D_z\}$ be a set of trained classification systems and $\Omega=\{c_1,c_2,...,c_N\}$ be a set of class labels. Each classification system takes as input a data point $x\in\Re^n$ and assigns it to a class label from Ω. Alternatively, each classifier output can be formed as an N-dimensional vector that represents the degree of support of a classification system to each of N classes. We suppose these classifier outputs are commensurable, i.e. defined on the same measurement scale (most often they are posterior probability-like). Let's denote h_i, $i=1,..,z$, the output scores of z classification systems for the class c_n. Assuming the sequence h_i, $i=1,..,z$, is ordered in such a way that $h_1\le...\le h_z$, the Choquet fuzzy integral can be computed as

$$M_{FI}(\mu,h)=\sum_{i=1}^{z}\left[\mu(i,...,z)-\mu(i+1,...,z)\right]\,h_i \qquad (4)$$

where $\mu(z+1)=\mu(\text{ø})=0$. $\mu(S)$ can be viewed as a weight related to a subset S of the set Z of information sources. It is called *fuzzy measure* and for $S,T\subseteq Z$ has to meet the following conditions:

$\mu(\text{ø})=0,\ \mu(Z)=1$, Boundary

$S\subseteq T\Rightarrow\mu(S)\le\mu(T)$, Monotonicity

3.2. Synchronization and normalization of system outputs

In order to fuse 3 information sources (SVM-, HMM-, and VIDEO-based systems), their outputs must be synchronized in time. In our case, the SVM system provides voting scores every 100ms, the VIDEO system every 40ms, and the HMM system gives segments of variable length which represent the best path throw the recognition network The outputs of the 3 systems were reduced to a common time step of 100ms. For that purpose the output of the VIDEO-based system was averaged on each interval of 100ms, while for the HMM system each segment was broken into 100ms-long pieces.

On the other hand, to make the outputs of information sources commensurable we have to normalize them to be in the range [0 1] and their sum equal to 1.

As it was said in Section 2.2, when the SVM classification system is used alone, after voting, the class with the highest number of winning two-class decisions (votes) is chosen. In case of a subsequent fusion with other classification systems numbers of votes obtained by non-winning classes were used to get vector of scores for the classes. For the HMM system, each hypothesis of an AE given by the optimal Viterbi segmentation of the seminar is posteriorly decoded by the trained HMM models of winning and each non-winning acoustic event class in order to obtain the corresponding log-likelihood values which form vector of scores. In the case of VIDEO-based AED system we obtain scores for the two classes "steps" and "non-steps" as the distance between the feature vector and the decision boundary. To make the scores of VIDEO-based and HMM-based systems positive *min-max* normalization is used.

The *soft-max* function is applied to the vector of scores of each detection system. This function is defined as:

$$q_i\big|_{normalized}=\exp(k*q_i)/\sum\nolimits_i\exp(k*q_i) \qquad (5)$$

where the coefficient k controls the distance between the components of the vector $[q_1,\ q_2,\ ...,q_N]$. For instance, in extreme case when $k=0$, the elements of the vector after *soft-max* normalization would have the same value $1/N$, and when $k\to\infty$ the elements tend to become binary. The normalization coefficients are different for each AED system, and they are obtained using the development data.

3.3. One-stage and two-stage fuzzy integral approaches

In our case, not all information sources give scores for all classes. Unlike SVM and HMM-based systems, which provide information about 15 classes, the VIDEO-based system scores are given only for the class "steps" and the meta-class "non-steps". Fusion of information sources using fuzzy integral can be done either by transforming (extending) the score for "non-steps" from the VIDEO-based system to 14 classes which do not include "steps" or, vice-versa, transforming (restricting) the scores for 14 classes provided by the SVM and HMM-based systems to one score for the meta-class "non-steps". In the former case, the fusion is done at one stage with all the classes. In the latter, a two-stage approach is implemented, where on the first stage the 3 detection systems are used to do "steps" / "non-steps" classification and on the second stage the subsequent classification of the "non-steps" output of the first stage is done with both SVM and HMM-based systems. The one-stage and two-stage approaches are schematically shown in Figure 3.

For one-stage fusion (Figure 3 (a)) the score V of "non-steps" of the VIDEO-based system was equally distributed among the remaining 14 classes assigning to each of them a score V before applying *soft-max* normalization. At the first stage of the two-stage approach (Figure 3 (b)), all the classes not labelled as "steps" form the "non-steps" meta-class. The final score of "non-steps" is chosen as the maximum value of the scores of all the classes that form that meta-class.

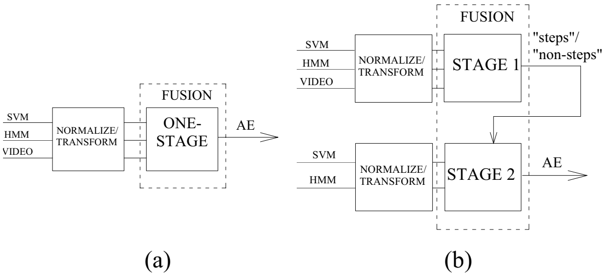


Figure 3. *One-stage (a) and two-stage (b) fusion with fuzzy integral*

The individual FMs for FI fusion are trained on development data in our work using the gradient descent training algorithm [10]. A 5-fold cross validation on development data was used to stop the training process to avoid overtraining.

4. Experiments and results

4.1. Database

In our experiments, the CLEAR'07 evaluation database is used [3]. It consists of 25 interactive seminars, approximately 30min-long that have been recorded by AIT, ITC, IBM, UKA, and UPC in their smart-rooms. In our experiments for development and testing we used only recordings of 3 sites (AIT, ITC, and UPC) because the IBM data is not included in the testing database, and the performance of the video tracking algorithm on the UKA data is very low, due to errors presented in the video recordings (heavy radial distortions in zenithal camera). In other respects, the training/testing division is preserved from CLEAR'07 evaluation scenario.

The AED evaluation uses 12 semantic classes (classes of interest), i.e. types of AEs that are: "door knock", "door open/slam", "steps", "chair moving", "spoon/cup jingle", "paper work", "key jingle", "keyboard typing", "phone ring", "applause", "cough", and "laugh". Apart from the 12 evaluated classes, there are 3 other events present in the seminars ("speech", "silence", "unknown") which are not evaluated.

The Accuracy metric [3] is used in this work and it is defined as the harmonic mean between *precision* and *recall* computed for the classes of interest, where *precision* is number of correct hypothesis AEs divided by total number of hypothesis AEs, and *recall* as number of correctly detected reference AEs divided by total number of reference AEs.

4.2. Results and discussion

The results of the two-stage fusion are presented in Figure 4. Results from the first-stage fusion for "steps"/"non-steps" detection are presented on the left side, while the final results of the AED system are on the right one. Firstly, it can be seen that fusion of SVM and HMM-based systems leads to small improvements for testing data, while in combination with video information the improvement is noticeable. It is worth to mention that 48.1% of accuracy for "steps" detection would indicate a little worse decision than random choice if the metric scores both "non-steps" meta-class and "steps" class. However, in our case, only the "steps" class is scored and thus 48.1% indicates that not only around 48.1% of "steps" are detected (recall) but also that 48.1% of all produced decisions are correct (accuracy).

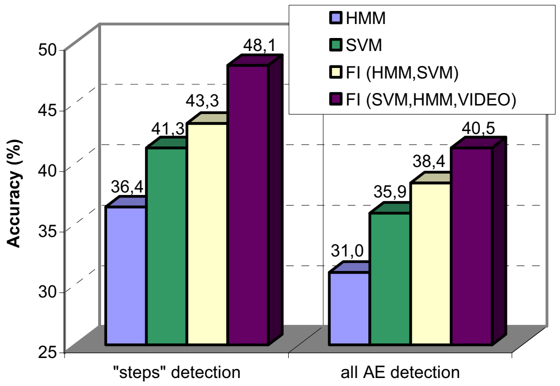


Figure 4. *Accuracy results from the first stage (left), and first with second stages (right) using FI*

The final results of detection of all 15 classes of AEs are presented in the right part of the Figure 4. It can be seen that total system accuracy benefits from better recognition of "steps" class, resulting in a final score of 40.5% accuracy.

5. Conclusions

In this work, motivated by the large amount of AED errors that occur in seminar conditions, which exhibit a large proportion of temporal overlap of sounds, we aim at improving the AED accuracy by applying fuzzy integral fusion to two complementary audio-based AED systems, built with SVM and HMM classifiers, and also a video-based AED system. After applying a video 3D tracking algorithm, video-based features that represent the movement have been extracted, and a probabilistic video-based classifier for "steps"/"non-steps" detection has been developed to improve the detection of that particular class and, consequently, to improve the AED accuracy of the whole set of 12 targeted AEs. Experimental results using the CLEAR 2007 evaluation data show that the detection rate increases by fusing the two audio information sources, and it is further improved by including video information. Future work will be devoted to extend the multimodal part of the AED system to more classes.

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] 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

[2] 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

[3] A. Temko, C. Nadeu, J-I. Biel, "Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07", in *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625, pp.354-363, Springer, 2008

[4] L. Kuncheva, Combining Pattern Classifiers, John Wiley & Sons, 2004

[5] A. Temko, D. Macho, C. Nadeu, "Fuzzy Integral Based Information Fusion for Classification of Highly Confusable Non-Speech Sounds", *Pattern Recognition*, vol. 41(5), pp.1831-1840, Elsevier, 2008

[6] A. López, C. Canton-Ferrer, J. R. Casas, "Multi-Person 3D Tracking with Particle Filters on Voxels", IEEE ICASSP'07, pp. 913-916, 2007

[7] M. Arulampalam., S. Maskell, N. Gordon, T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking", *IEEE Transaction on Signal Processing*, vol. 50, 174-188, 2002

[8] C. Hsu, C. Lin, "A Comparison of Methods for Multi-class Support Vector Machines", *IEEE Transactions on Neural Networks*, pp.415-425, 2002

[9] S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.2)", Cambridge University, 2002

[10] M. Grabisch, "A new algorithm for identifying fuzzy measures and its application to pattern recognition", IEEE International Conference on Fuzzy Systems, pp.145-50, 1995