Inclusion of Video Information for Detection of Acoustic Events using the Fuzzy Integral

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
{butko, temko, climent, ccanton}@gps.tsc.upc.edu

Abstract. When applied to interactive seminars, the detection of acoustic events from only audio information shows a large amount of errors, which are mostly due to the temporal overlaps of sounds. Video signals may be a useful additional source of information to cope with that problem for particular events. In this work, we aim at improving the detection of steps by using two audio-based Acoustic Event Detection (AED) systems, with SVM and HMM, and a video-based AED system, which employs the output of a 3D video tracking algorithm. The fuzzy integral is used to fuse the outputs of the three detection systems. Experimental results using the CLEAR 2007 evaluation data show that video information can be successfully used to improve the results of audio-based AED.

Keywords: Acoustic Event Detection, Fuzzy Integral, Multimodality, Support Vector Machines, Hidden Markov Models, Video 3D Tracking.

1. Introduction

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 (Classification of Events, Activities and Relationships) international evaluation database consists of several interactive seminars which, among other things, contain "meeting", "coffee break", "question/answers" activities. The evaluation campaign showed that in that seminar conditions AED is a challenging problem. In fact, 5 out of 6 submitted systems showed accuracy 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 efficient algorithms that work at the signal level, the model level or 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 including video information in our existing audio-based detection systems 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.

The information about movements and positions of people in a meeting room may be correlated with acoustic events 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 acoustic events in the CLEAR’07 evaluation database, in this work we use video 3D tracking information in order to improve the detection of that particular class.

In our work, *late fusion* is used by combining the decisions from several information sources: two audio-based AED systems, with SVM and HMM, and a VIDEO-based AED system. Fusion is carried out with the Fuzzy Integral (FI) [4] [5], a fusion technique which is able to take into account the interdependences among information sources. Unlike non-trainable fusion operators (*mean, product* [4]) the statistical FI approach can be more beneficial in our challenging task. From the results, FI fusion shows better accuracy than either the single classifiers or the classical Weighted Arithmetical Mean (WAM) fusion operator [4].

The rest of this paper is organized as follows: Section 2 describes video and audio-based systems of AED. The 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 acoustic events 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. Person tracking and multi-object tracking

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 of the people inside the room at a given time *t*. Two main factors are to be taken into account when implementing a particle filter: the likelihood function and the propagation strategy.

Likelihood function $p(z_t|x_t)$ can be defined as the likelihood of a particle belonging to the volume that corresponds to a person. For a given particle j occupying a voxel x_t, its likelihood is formulated as:

$$p(z_t \mid x_t^j) = \frac{1}{\left|C(x_t^j, q)\right|} \sum_{p \in C(x_t^j, q)} d(x_t^j, p) \quad (1)$$

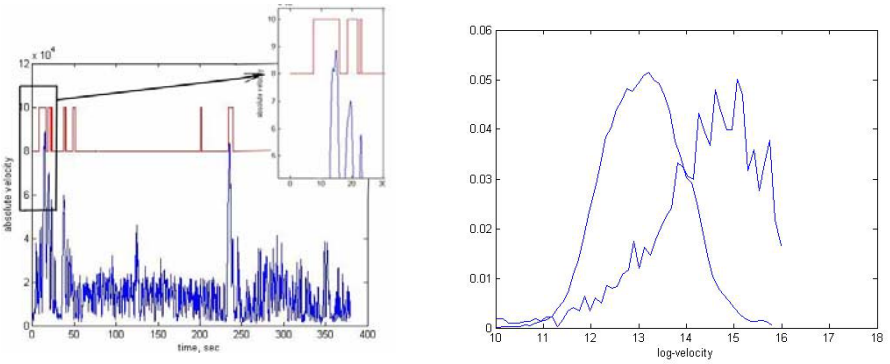
where $C(\cdot)$ stands for the neighborhood over a connectivity q domain on the 3D orthogonal grid and $|C(\cdot)|$ represents its cardinality. Typically, connectivity in 3D discrete grids can be 6, 14 and 26 and in our research $q=26$ provided accurate results. Function $d(\cdot)$ measures the distance between a foreground voxel p in the neighborhood and the particle.

Challenges in 3D multi-person tracking from volumetric scene reconstruction are basically twofold. First, finding an interaction model in order to avoid mismatches and target merging. Several approaches have been proposed [8] but the joint PF presented in [9] is the optimal solution to multi-target tracking using PFs. However, its computational load increases dramatically with the number of targets to track since every particle estimates the location of all targets in the scene simultaneously. The proposed solution is to use a split PF per person, which requires less computational load at the cost of not being able to solve some complex cross-overs. However, this situation is alleviated by the fact that cross-overs are restricted to the horizontal plane in our scenario (see Fig.1).

Let us assume that there are M independent PF trackers, being M the number of humans in the room. Nevertheless, they are not fully independent since each PF can consider voxels from other tracked targets in either the likelihood evaluation or the 3D re-sampling step resulting in target merging or identity mismatches. In order to achieve the most independent set of trackers, we consider a blocking method to model interactions. Many blocking proposals can be found in 2D tracking related works [9] and we extend it to our 3D case.

Fig. 1. Particles from the tracker A (yellow ellipsoid) falling into the exclusion zone of tracker B (green ellipsoid) will be penalized

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.



(a) (b)

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

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:

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

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 $h(n) = [-m \ldots -2\ -1\ 0\ 1\ 2 \ldots m]$ (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.

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

The normalized histograms of the logarithm of velocity for "steps" and "non-steps" obtained from development seminars are depicted in Fig. 2 (b), from which 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 Expectation-Maximization

algorithm (EM). During detection on testing data the final decision for "steps"/ "non-steps" classes is made using the Bayesian rule:

$$P(w_j \mid x) = P(x \mid w_j)P(w_j), j=\{1,2\}. \quad \textbf{(3)}$$

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. According to the results shown in Fig. 3, the best detection of "steps" on development data is achieved with a 2-sec-long derivative filter and a Hamming window.

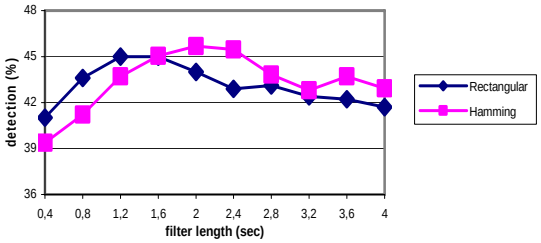


Fig. 3. Detection of "steps" on the development database as a function of the length of the derivative filter (in seconds)

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 1vs1 scheme on the isolated AEs, from two databases of isolated acoustic events, 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) [10] 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

We formulate the goal of acoustic event detection in a way similar to speech recognition: to find the event sequence that maximizes the posterior probability of the event sequence $W = (w_1, w_2, ..., w_M)$, given the observations $O = (o_1, o_2, ..., o_T)$:

$$W_{max} = argmax\ P(W|O) = argmax P(O|W)P(W). \quad (4)$$

We assume that $P(W)$ is the same for all event sequences.

For building and manipulating hidden Markov models HTK toolkit is used [11]. 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 20-ms with shift 10-ms. There is one HMM for each acoustic event class, with five emitting states and fully connected state transitions. We also used a similar HMM 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 during a cross-validation procedure on the development data. Each HMM is trained on all signal segments belonging to the corresponding event class in the development seminar data, using the standard Baum-Welch training algorithm. During testing the AED system finds the best path through the recognition network and each segment in the path represents a detected AE.

3. Fusion of information sources

3.1. The 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. It is convenient to organize the output of all classification systems in a decision profile:

$$DP(x) = \begin{bmatrix} d_{1,1}(x) \ldots d_{1,n}(x) \ldots d_{1,N}(x) \\ \ldots \\ d_{j,1}(x) \ldots d_{j,n}(x) \ldots d_{j,N}(x) \\ \ldots \\ d_{z,1}(x) \ldots d_{z,n}(x) \ldots d_{z,N}(x) \end{bmatrix} \quad (5)$$

where a row is classifier output and a column is a support of all classifiers for a class.

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 (the supports for class c_n, i.e. a column from decision profile) and before defining how FI combines information sources, let's look to the conventional WAM fusion operator. A final support measure for the class c_n using WAM can be defined as:

$$M_{WAM} = \sum_{i \in Z} \mu(i) h_i$$
$$\text{where } \sum_{i \in Z} \mu(i) = 1 \text{ (additive)}, \ \mu(i) \geq 0 \text{ for all } i \in Z \tag{6}$$

The WAM operator combines the score of z competent information sources through the weights of importance expressed by $\mu(i)$. The main disadvantage of the WAM operator is that it implies preferential independence of the information sources.

Let's denote with $\mu(i,j) = \mu(\{i,j\})$ the weight of importance corresponding to the couple of information sources i and j from Z. If μ is not additive, i.e. $\mu(i,j) \neq [\mu(i) + \mu(j)]$ for a given couple $\{i,j\} \subseteq Z$, we must take into account some interaction among the information sources. Therefore, we can build an aggregation operator starting from the WAM, adding the term of "second order" that involves the corrective coefficients $\mu(i,j) - [\mu(i) + \mu(j)]$, then the term of "third order", etc. Finally, we arrive to the definition of the FI: assuming the sequence h_i, $i=1,..,z$, is ordered in such a way that $h_1 \leq ... \leq h_z$, the Choquet *fuzzy integral* can be computed as

$$M_{FI}(\mu, h) = \sum_{i=1}^{z} [\mu(i,...,z) - \mu(i+1,...,z)] \ h_i \tag{7}$$

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

$\mu(\emptyset) = 0, \mu(Z) = 1$, Boundary

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

For instance, as an illustrative example let's consider the case of 2 information sources with unordered system outputs $h_1=0.4$ and $h_2=0.3$, and corresponding fuzzy measures $\mu(1)=0.6$ and $\mu(2)=0.8$. Note that $\mu(0)=0$ and $\mu(1,2)=1$. In that case, the Choquet *fuzzy integral* is computed as $M_{FI}(\mu,h) = (\mu(1,2) - \mu(1))h_2 + \mu(1)h_1 = 0.36$.

3.2 Synchronization and normalization of system outputs

In order to fuse 3 information sources (SVM-based, HMM-based, 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 through 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 a 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 then decoded by the trained HMM models of winning and each non-winning AE class in order to obtain the corresponding log-likelihood values which form vector of scores. In case of VIDEO-based AED system we obtain scores for the two classes "steps" and "non-steps" as the distance between the values of log-velocity and the decision boundary. To make the scores of VIDEO-based and HMM-based systems positive *min-max* normalization [12] is used.

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

$$q_i\big|_{normalized} = \exp(k * q_i) / \sum_i \exp(k * q_i) \qquad (8)$$

where the coefficient k controls the distance between the components of the vector $[q_1, q_2, \ldots, q_N]$. 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.

4. Experiments and results

4.1. Database and metric

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 (Athens Information Technology), ITC (Instituto Trentino di Cultura), IBM, UKA (Universität Karlsruhe), and UPC (Universitat Politècnica de Catalunya) 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. 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 the fuzzy integral can be done either by transforming (extending) the score for "non-steps" from the VIDEO-based system to the remaining 14 classes which do not include "steps" or, vice-versa, transforming (restricting) the scores of 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 Fig. 4.

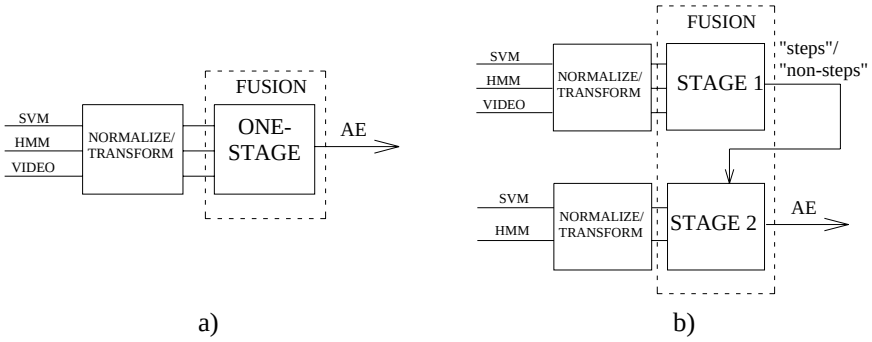


Fig. 4. One-stage (a) and two-stage (b) fusion with the fuzzy integral

For one-stage fusion (Fig. 4 (a)) the score V of "non-steps" of the VIDEO-based system is equally distributed among the remaining 14 classes assigning to each of them score V before applying *soft-max* normalization. At the first stage of the two-stage approach, all the classes not labeled as "steps" form the "non-steps" meta-class. The final score of "non-steps" is chosen as maximum value of scores of all the classes that formed that meta-class.

For the weights in WAM operator we use uniform class noise model with the

detection system [4]. The individual FMs for the fuzzy integral fusion are trained on development data in our work using the gradient descent training algorithm [13]. The 5-fold cross validation on development data was used to stop the training process to avoid overtraining. The tricky point was that during training the algorithm minimizes the total error on development data. As the number of data per each class is non-uniform distributed, during the training process the number of detection mistakes for the most representative classes ("speech", "silence") is decreased at the expense of increasing errors on the classes with lower number of representatives. The final metric scores, however, only 12 classes which are the classes with much smaller number of representatives than e.g. "speech". This way, the FI with the trained FM measure tends to detect correctly the classes that are not scored by the metric. To cope with this problem, we firstly fixed the FM of the classes of no interest ("speech", "unknown", and "silence") to be in the equilibrium state [13] and, secondly, calculate the cross-validation accuracy only for the classes of interest.

4.3. Results and discussion

The results of first-stage fusion for "steps"/"non-steps" detection are presented in Fig. 5. It can be seen that fusion of SVM and HMM-based systems leads to a small improvement, 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 scored 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 (precision). On the first stage the FI fusion gives superior results in comparison with WAM fusion. This indicates that a certain interaction between information sources for "steps" detection exists that can not be captured by WAM fusion operator.

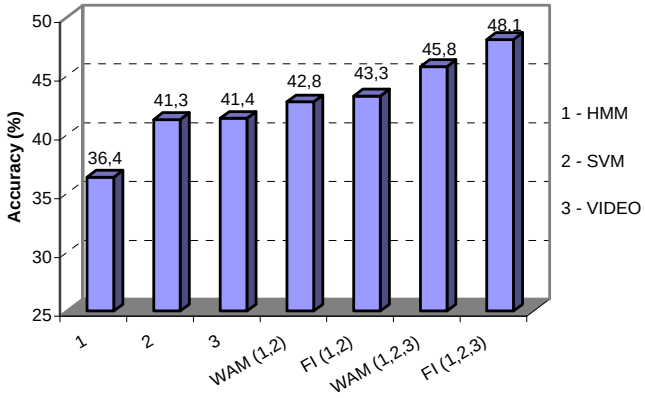


Fig. 5. Accuracy of "steps" detection on the first stage using the fuzzy integral

The final results of detection of all 15 classes of AEs are presented in Fig. 6. It can be seen that total system accuracy benefits from better recognition of "steps" class.

Again in this experiment the FI fusion shows better performance then WAM, resulting in a final accuracy of 40.5%.

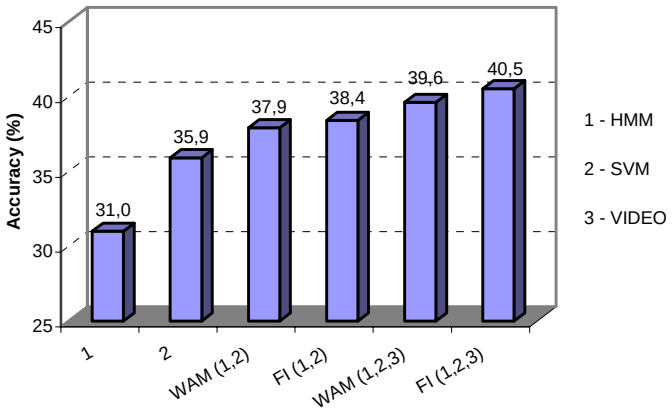


Fig. 6. Total system accuracy based on the first and the second stage fusion

One-stage fusion explained in the previous subsection showed lower scores - only 37.9% for WAM and 38.4% with FI. This fact may indicate that in our particular case spreading no-information for classes with missing scores can be harmful and, conversely, to compress the scores of many classes to binary problems can be more beneficial. However, the way of extending/compressing of the scores should be studied in more depth to further support this statement.

5. Conclusions

In this work, by using data from interactive seminars, we have shown that video signals can be a useful additional source of information to cope with the problem of acoustic event detection. Using an algorithm for video 3D tracking, video-based features that represent the movement have been extracted, and a probabilistic classifier for "steps"/"non-steps" detection has been developed. The fuzzy integral was used to fuse the outputs of both that video-based detector and two audio-based AED systems which use either SVM or HMM classifiers. Results show that video information helps to detect acoustic “steps” events, and future work will be devoted to extend the multimodal AED system to more classes.

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. The fourth author has been partially supported by the Spanish Ministerio de Educación y Ciencia, under project TEC2007-66858/TCM.

6. 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. O. Lanz ”Approximate Bayesian Multibody Tracking”, *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 28(9), pp. 1439-1449, 2006
9. Z. Khan, T. Balch, F. Dellaert, “Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model”, International Conference on Intelligent Robots and Systems, 2003
10. C. Hsu, C. Lin, “A Comparison of Methods for Multi-class Support Vector Machines*”, IEEE Transactions on Neural Networks*, pp.415-425, 2002
11. S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book (for HTK Version 3.2)”, Cambridge University, 2002
12. L. Shalabi, Z. Shaaban, B. Kasasbeh, “Data Mining: A Preprocessing Engine”, Journal of Computer Science, vol. 2 (9), pp. 735-739, 2006
13. 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