FUZZY INTEGRAL BASED INFORMATION FUSION FOR CLASSIFICATION OF HIGHLY CONFUSABLE NON-SPEECH SOUNDS

Andrey Temko[1*], Dušan Macho[2], Climent Nadeu[1]

[1]TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici D5, Jordi Girona 1-3, 08034 Barcelona, Spain

[2]Motorola Inc., Schaumburg, Illinois, USA.

ABSTRACT

Acoustic event classification may help to describe acoustic scenes and contribute to improve the robustness of speech technologies. In this work, fusion of different information sources with the Fuzzy Integral (FI), and the associated Fuzzy Measure (FM), are applied to the problem of classifying a small set of highly confusable human non-speech sounds. As FI is a meaningful formalism for combining classifier outputs that can capture interactions among the various sources of information, it shows in our experiments a significantly better performance than that of any single classifier entering the FI fusion module. Actually, that FI decision-level fusion approach shows comparable results to the high-performing SVM feature-level fusion and thus it seems to be a good choice when feature-level fusion is not an option. We have also observed that the importance and the degree of interaction among the various feature types given by the FM can be used for feature selection, and gives a valuable insight into the problem.

Keywords: Acoustic event classification, audio features, fuzzy integral and fuzzy measure, feature-level and decision-level information fusion, feature selection, interaction of information sources, Choquet integral.

1. INTRODUCTION

In context-aware systems such as smart rooms or intelligent personal devices, acoustic event classification (AEC) can provide support for a high-level analysis of the underlying acoustic scene. On the other hand, AEC can contribute to improve the performance and robustness of speech technologies such as speech and speaker recognition, speech enhancement, and speaker localization.

[*] Corresponding author. Tel: +34-93-401-1627, Fax: +34-93-401-6447. E-mail address: temko@gps.tsc.upc.es (A.Temko)

A rich variety of information sources is obtained in our work by extracting a set of ten different kinds of features and using them as inputs of ten different SVM classifiers, whose outputs are combined to give a final classification score. Besides the above-mentioned fusion of information sources at the decision level, and for clarity purposes, we will also consider information fusion at the feature level, i.e. an early integration of information sources, and will be carried out by the SVM. These two kinds of fusion are depicted in Fig. 1.

Usual combinations of classifier outputs like sum, product, max, min, weighted arithmetical mean (WAM), etc [4], assume that each output represents an independent source of information that can be treated separately. Often, this is not the case, and an approach that considers the interactions among the classifier outputs is needed. Over the past several years there have been a number of successful applications of the Fuzzy Integral (FI) [4] [5] in decision-making and pattern recognition using multiple information sources (e.g. [6] [7] [8]). FI is a meaningful formalism for combining classifier outputs which can capture interactions among the various sources of information. Moreover, the Fuzzy Measure (FM), which is associated with the FI, furnishes a measure of importance for each subset of information sources, allowing feature selection and giving a valuable insight into the classification problem itself.

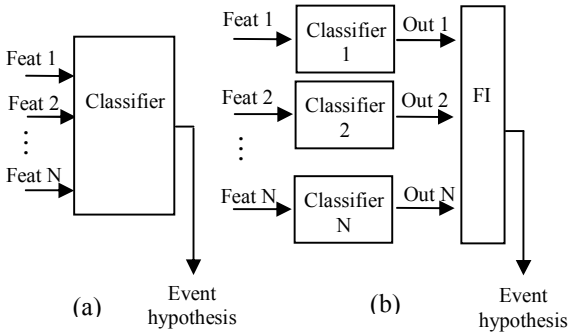

Figure 1. Fusion at the feature level (a) and at the decision level (b).

Both feature-level fusion and decision-level fusion are compared in our AEC experiments. As a default classifier we use the SVM classifier, which helps to overcome the problem of the high-dimensionality [9] of the input feature space. In the experiments, fusion of several information sources with the FI formalism showed a significant improvement with respect to the best single information source. The results also indicate that the decision-level fusion by FI outperforms WAM and has similar or better results than feature-level fusion with the SVM. Additionally, the number of feature sets was reduced to half with no loss in the SVM-based classification rate by using the information displayed by the FM

scores.

In previously reported classification experiments with 16 types of meeting-room acoustic events [1], we tested several classifiers based on either Gaussian Mixture Models (GMM) or Support Vector Machines (SVM) [2]. In those experiments, the SVM approach yielded significantly higher accuracies than the GMM approach. This was mostly due to the small size of the training database since, as SVM is a discriminant classifier, it has lower size requirements for training data than the GMM generative classifier. Furthermore, the SVM classifier is not as sensitive as the GMM one to the presence of irrelevant features [3], so it is appropriate to use it with a large and diverse feature set, as was done in those tests.

In this paper, we focus on the classification of a particular type of acoustic events, a set of five human vocal-tract non-speech sounds (cough, laughter, sneeze, sniff and yawn), which were found responsible for a large part of errors in the classification of meeting-room acoustic events in [1]. In fact, those sounds contributed with 70% of the total classification error, in spite of accounting only for 30% of the acoustic events included in the testing database. Additionally, it was observed in [1] that those human non-speech sounds were mainly confused among themselves. Using the same small database and keeping SVM as the basic classifier, the work presented in this paper is intended to reduce the classification error rate of the above mentioned set of highly confusable human non-speech sounds by turning to the fusion of different information sources that in our case consists of the combination of classifier outputs.

Finally, as the FI aggregation may be appropriate when the feature-level fusion is difficult (e.g. due to the different nature of the involved features), or when it is beneficial to preserve the application or technique dependency (e.g. when fusing well established feature-classifier configurations), we have also conducted experiments to combine Hidden Markov Models (HMM) that use frame-level features with the SVM using signal-level features, and have witnessed an additional improvement. As smart rooms are usually equipped with a network of microphones and video cameras that provide multimodal information, fusion of information with the FI may find a useful application in such a framework.

The rest of the paper is organized as follows: Section II gives the basics of FI and FM. Audio features investigated in this work are presented in Section III. Section IV presents the experiments and discussion. Finally, conclusions are given in Section V.

2. 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) [10]:

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

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 DP) 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:

$$M_{WAM} = \sum_{i \in Z} \mu(i) h_i \qquad (1)$$

where $\sum_{i \in Z} \mu(i) = 1$ (additive), $\mu(i) \geq 0$ *for all* $i \in Z$

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 [11].

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. In this way, 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* [1][6][12] can be computed as

$$M_{FI}(\mu,h) = \sum_{i=1}^{z} [\mu(i,...,z) - \mu(i+1,...,z)]\ h_i \qquad (2)$$

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 has to meet the following conditions:

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

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

where $S, T \subseteq Z$.

To illustrate the FI, let us consider a case of two information sources with outputs h_1 and h_2, and assume that $h_1<h_2$. Consequently, we have corrective coefficients of the second order only: $\mu(1,2)-[\mu(1)+\mu(2)]$. According to (2), FI is computed as

$$M_{FI}(\mu,h) = [\mu(1,2)-\mu(2)]\,h_1 + \mu(2)\,h_2$$

which, after a slight manipulation, results in

$$M_{FI}(\mu,h) = [\mu(1,2)-(\mu(2)+\mu(1))]\,h_1 + \mu(1)\,h_1 + \mu(2)\,h_2$$

where the first term corresponds to the "second order" correction mentioned above.

For *Z* information sources there are a total of 2^Z FM parameters that can be arranged in a lattice with the usual ordering of real numbers [7]. The lattice representation shows the monotonicity of the FM and particular values involved in the FI calculation. An example of lattice representation of FM defined for 4 information sources is shown on Fig. 2. The lattice consists of *Z+1* layers with each node representing a particular subset of *Z*. Two nodes in adjacent layers are connected only if there are set-inclusion

relationships between the two subsets of Z whose measures they represent. The red line on the Fig. 2 shows the values used for the FI calculation given the following ordering of classifiers' scores: $h_1<h_4<h_2<h_3$.

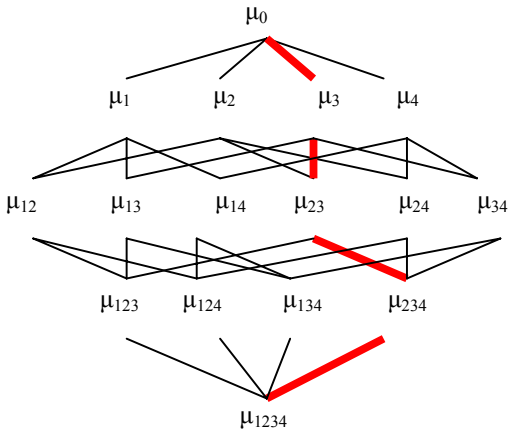

Figure 2. Lattice representation of Fuzzy Measure for 4 information sources.

Indeed, the large flexibility of the FI aggregation operator is due to the use of FM that can model interaction among criteria. And although the FM $\mu(i)$ provides an initial view about the importance of information source i, all possible subsets of Z that include that information source should be analysed to give a final score. For instance, we may have $\mu(i) = 0$, suggesting that element i, $i \notin T$, is not important; but if, at the same time, $\mu(T \cup i) >> \mu(T)$, this actually indicates i is an important element for the decision. For calculating the *importance* of the information source i, the Shapley score [6] [11] is used. It is defined as:

$$\phi(\mu, i) = \sum_{T \subseteq Z \setminus i} \frac{(|Z| - |T| - 1)!|T|!}{|Z|!}[\mu(T \cup i) - \mu(T)] \qquad (3)$$

Generally, (3) calculates a weighted average value of the marginal contribution $\mu(T \cup i) - \mu(T)$ of the element i over all possible combinations. It can be easily shown that the information source importance sums to one.

Another interesting concept is interaction among information sources. As long as the fuzzy measure is not additive, there exists some correlation among information sources. When $\mu(i, j) < \mu(i) + \mu(j)$ the information sources i and j express negative synergy and can be considered redundant. On the contrary,

when $\mu(i,j) > \mu(i) + \mu(j)$, the information sources *i* and *j* are complementary and express positive synergy. For calculating the interaction indices, instead of the marginal contribution of element *i* in (3), the contribution of a pair of information sources *i* and *j* is defined as the difference between the marginal contribution of the pair and the addition of the two individual marginal contributions [11], or equivalently:

$$(\Delta_{i,j}\mu)(T) = \mu(T \cup i,j) - \mu(T \cup i) - \mu(T \cup j) + \mu(T) \qquad (4)$$

and the *interaction* indices are calculated as:

$$I(\mu; i, j) = \sum_{T \subseteq Z \setminus i,j} \frac{(|Z| - |T| - 2)!|T|!}{(|Z| - 1)!}(\Delta_{i,j}\mu)(T)] \qquad (5)$$

We can see the index is positive as long as *i* and *j* are negatively correlated (complementary), and negative when *i* and *j* are positively correlated (competitive).

As was mentioned in [11], FI has very good properties for aggregation: it is continuous, non-decreasing, ranges between a minimum and a maximum value, and coincides with WAM (discrete Lebesgue integral) as long as the FM is additive. Actually, it was shown in [11] that the ordered weighted average, the WAM, and the partial minimum and maximum operators are all particular cases of FI with special FM. In fact, FI can be seen as a compromise between the evidence expressed by the outputs of the classification systems and the competence represented by the FM's knowledge of how the different information sources interact [4].

As the FM is a generalization of a probability measure, we can calculate a measure of uncertainty associated to FM analogously to the way the entropy is computed from the probability [13], that is:

$$H(\mu) = \sum_{i=1}^{z} \sum_{T \subseteq Z \setminus i} \gamma_T \, g[\mu(T \cup i) - \mu(T)] \qquad (6)$$

where $\gamma_T = (|Z| - |T| - 1)!\ |T|!\ /\ |Z|!$, $g(x) = -x \ln x$, and $0 \ln 0 = 0$ by convention.

When normalized by $\ln|Z|$, $H(\mu)$ measures the extent to which the information sources are being used in calculating the aggregation value of $M_{FI}(\mu, h)$. When that *entropy* measure is close to 1, all criteria are used almost equally; when it is close to 0, the FI concentrates almost on only one criterion [14].

It is obvious that FI completely relies on the FM. The better the FM describes the real competence and interaction among all classification systems, the more accurate results can be expected. There are two methods of calculating the FM known to the authors (if it is not provided by an expert knowledge): one based on fuzzy densities [4], and the other based on learning the FM from training data [7] [8]. In our work, we have used the latter method: a supervised, gradient-based algorithm of learning the FM, with additional steps for smoothing the unmodified nodes:

Step 1. Initialize the FM to the equilibrium state $\mu(i) = 1/Z$, where Z is the number of information sources and FM is additive, i.e. $\mu(i, j) = \mu(i) + \mu(j)$

Step 2. For a data point x with label c_n

Step 2.1. Obtain the *DP(x)* and calculate the FIs M_n for each of N classes (i.e. for each column of the *DP*).

Step 2.2. Calculate the error for each of N classes: $\varepsilon_n = c_n - M_n$, where c_n is one for the correct class and zeros for the others.

Step 2.3. For each of N classes, update the FM μ values that were used in the calculation of M_n (e.g. in Fig. 2, those that are on the red line, i.e. μ_{234}, μ_{23} and μ_3) using the formula derived from a mean-squared-error criterion [7]. Note that for each class the order of classifiers may differ, what implies that different μ values are used for the calculation of M_n.

Step 2.4. Verify the monotonicity condition for the μ values that were used in the calculation of M_n.

Step 3. Due to the scarcity of data, verify the monotonicity condition of the unmodified μ values and smooth their values. The smoothing is based on the average values of the upper and lower neighbours of the current node.

For a detailed description of the algorithm and the exact parameter update equations, we refer to [7] [8].

3. AUDIO FEATURES

Although the best feature sets for AEC in [1] consisted of combinations of features used in automatic speech recognition and other perceptual features, in the current work we only focus on the latter, since their contribution to vocal-tract sounds is not so well-established. 10 types of features were chosen with a substantial degree of redundancy in order to use FM to find out their relative importance and their degree

of interaction. We use the following notation in feature definition:

$s(n)$ – signal value at the time index n;

N – frame length;

$f(i)$, $a(i)$ – frequency value at the frequency bin i and the corresponding Discrete Fourier Transform (DFT) amplitude, respectively;

$x(k)$, $y(k)$ – value of mel-scaled logarithmic filter-bank energy (log FBEs) at the sub-band frequency index k corresponding to the current and previous frame, respectively;

The following types of frame-level acoustic features, with the number of features in parenthesis, are investigated:

1. Zero crossing rate (1). It measures the number of zero crossings of the waveform within a frame and is calculated as:

$$ZCR = \sum_{n=0}^{N-1} I\{s(n)s(n-1) < 0\} \tag{7}$$

where the indicator function $I\{A\}$ is 1 if its argument A is true and 0 otherwise.

2. Short-time energy (1). Total signal energy in a frame calculated as:

$$STE = \sum_{n=0}^{N-1} s(n)s(n) \tag{8}$$

3. Fundamental frequency (1). A simple cepstrum-based method was used to determine the pitch in the range [70, 500] Hz [15]. When the signal is unvoiced, a zero value is used.

4. Sub-band log energies (4). The 4 sub-bands are equally distributed along the 20 mel-scaled FBEs (5 per sub-band). The energy of each sub-band is calculated as:

$$SBE(j) = \sum_{k=5j}^{5j+N-1} x(k) \quad \text{for } j = 0,...,3 \tag{9}$$

where $N=5$ is the number of log FBEs per sub-band.

5. Sub-band log energy distribution (4). Percentage distribution of the total log frame energy among the above defined 4 sub-bands.

6. Sub-band log energy correlations (4). This new type of feature is a measure of correlation of log FBEs between two adjacent frames and within each of the above defined 4 sub-bands. It is computed as the maximum absolute value of the cross-correlation function between the two sequences *x(k)* and *y(k)*:

$$SBC(j) = \max_{d} \left[abs \left(\frac{\sum_{k=5j}^{5j+N-1} [(x(k) - mx(j)) \cdot (y(k-d) - my(j))]}{\sqrt{\sum_{k=5j}^{5j+N-1} (x(k) - mx(j))^2} \sqrt{\sum_{k=5j}^{5j+N-1} (y(k-d) - my(j))^2}} \right) \right] \quad (10)$$

for $j = 0,...,3$

where *mx*(*j*) and *my*(*j*) are the means of the corresponding sub-band spectra, *d=0,1…,N-1* are mel-scaled sub-band frequency delays, and *N=5* is the number of log FBEs per sub-band.

7. Sub-band log energy time differences (4). It measures the changes of spectra in time and is calculated as difference of log energies between two adjacent frames for the above defined 4 sub-bands:

$$SBD(j) = \sum_{k=5j}^{5j+N-1} (x(k) - y(k)) \quad \text{for } j = 0,...,3 \quad (11)$$

where *N=5* is the number of log FBEs per sub-band.

8 .Spectral centroid (1). The centroid is a measure of the spectral "brightness" of the spectral frame and is defined as the linear average frequency weighted by DFT amplitudes, divided by the sum of the amplitudes:

$$CE = \frac{\sum_{\forall i} f(i)\, a(i)}{\sum_{\forall i} a(i)} \quad (12)$$

9. Spectral roll-off (1). It is a measure of the skewness of the spectral shape and is defined as a frequency bin f_c below which the *c* percentage of the spectral amplitudes is concentrated (in our case c=95):

$$\sum_{i=0}^{f_c} a(i) = \frac{c}{100} \sum_{\forall i} a(i) \tag{13}$$

10. Spectral bandwidth (1). A measure of spreading of the spectrum around the spectral centroid:

$$BW = \sqrt{\frac{\sum_{\forall i} (f(i) - CE)^2 a^2(i)}{\sum_{\forall i} a^2(i)}} \tag{14}$$

where *CE* is the *spectral centroid* of the frame.

Therefore, 22 acoustical measures are extracted from each frame, using 16ms/8ms frame length/shift. Then, from the whole time sequence of each acoustical measure in an event, four statistical parameters are computed: mean, standard deviation, autocorrelation coefficient at the second lag, and entropy. Those four statistical values per acoustical measure are used to represent the whole acoustic event.

4. EXPERIMENTS AND DISCUSSION

4.1. Experimental Setup

4.1.1. Database

Due to the lack of an acceptable corpus, the acoustic event database used in this work has been assembled using different sources. Part of the database was taken from the seminar recordings employed within the CHIL project [16]. The other part has been found in a large number of Internet websites.

Table 1. Sound Classes and Number of Samples Per Class

	Event	Number
A	Cough & Throat	119
B	Laughter	37
C	Sneeze	40
D	Sniff	37
E	Yawn	12

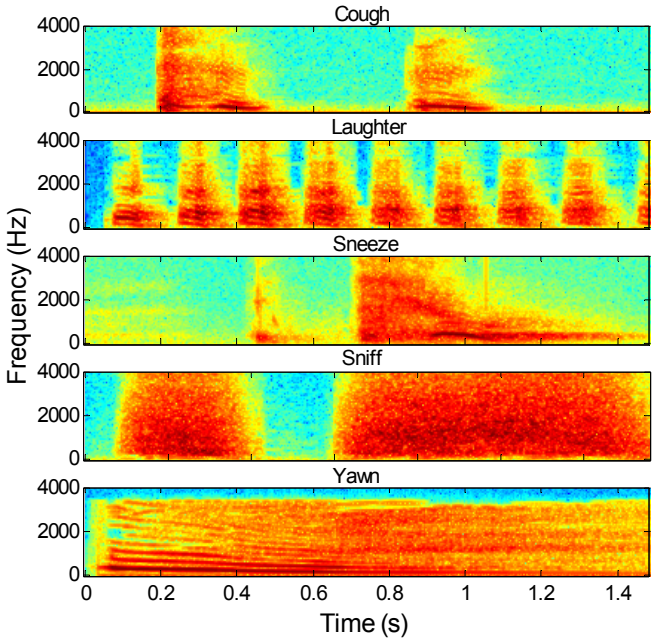

Figure 3. Sample spectrograms of acoustic events from **Table 1**.

All sounds were down-sampled to 8 kHz. The fact that the acoustic events were taken from different sources makes the classification task more complicated due to the presence of several (sometimes unknown) environments and recording conditions. Table 1 shows the five acoustic classes considered in this work and Fig. 3 shows their sample spectrograms. Notice that each realization of "cough" and "sniff" (there are two in the depicted time interval) shows a rather stationary behaviour, and "laughter" is almost periodic. Conversely, both "sneeze" and "yawn" have more spectral change. Actually, the "sneeze" sound results from the concatenation of two very different waveforms, and the "yawn" sound shows a decreasing pitch in its first segment.

There is a high variation in the number of samples per class, which represents an additional difficulty. In order to achieve a reasonable testing scenario, the data has been approximately equally split into the training and testing parts in such a way that there was the same number of representatives from the two data sources in the training and testing part. 10 runs were done in all the experiments.

4.1.2. SVM setup

The training data for each binary SVM classifier were firstly normalized anisotropicly to be in the range from –1 to 1, and the obtained normalizing template was then applied also to the testing data that are fed to that classifier. In the experiments with the SVM we used the Gaussian kernel. Leave-one-out cross validation [2] was applied to search for the optimal kernel parameter σ. To cope with the data imbalance we introduced different generalization parameters (C_+ and C_-) for positively- and negatively-labelled training samples: $C_+ = K\frac{A_-}{A_+}$, $C_- = K\frac{A_+}{A_-}$ where A_+ and A_- are the number of positive and negative training samples, respectively. In this way, the training errors of the two classes contribute equally to the cost of misclassification [1]. K was set to value 10 for all experiments as it was done in [1]. The MAX WINS (pairwise majority voting) [18] scheme was used to extend the SVM to the task of classifying several classes. The softmax function was applied to the class densities, which were calculated with pairwise majority voting, in order to obtain probability-like values.

4.1.3. Metrics

For comparison of the results, three metrics are used. One is the overall system accuracy, which is computed as the quotient between the number of correct hypothesis (outputs) given by the classifier for all the classes and the total number of instances in the testing set. The other two metrics are the mean per class recall and the mean per class precision, which are defined as:

$$\text{Rec} = \frac{1}{|C|}\sum_{c\in C}\frac{|h_{corr}(c)|}{|r(c)|},\ \text{Prec} = \frac{1}{|C|}\sum_{c\in C}\frac{|h_{corr}(c)|}{|h(c)|} \qquad (15)$$

where $|.|$ denotes cardinality of a set, C is the set of classes, c is a specific class, $r(c)$ is the number of reference (manually-labelled testing) instances and $h(c)$ is the number of hypothesis instances for class c. The subscript $_{corr}$ refers to a correct hypothesis. Due to the imbalance in amount of data per class, we think that the recall measure is more meaningful than the overall accuracy, but we use both of them for our comparisons, together with the precision measure.

4.2. Shared, semi-shared, and individual fuzzy measure

The Fuzzy Measure can be defined for all classes (shared FM), as we did in all experiments reported below in this article, or it can be defined for each class separately (individual FM), or for a group of classes (we call it semi-shared FM). When shared FM is used, it is learned using the error of all classes.

Thus, one FM covers all class-classifier dependences. When using individual FM, the error of a given class contributes to change only its own FM. In that way, the various FMs allow different order of importance of classifiers for each class. In that individual FM case, enough data should be available to train each class FM. As an intermediate solution, semi-shared FM may be used, assigning each FM to a group of similar classes. Table 2 shows the results obtained for each case.

Table 2. FI Result for Individual, Semi-shared and Shared Fuzzy Measure

	FI (ind)	FI (semi-sh) 15vs234	FI (semi-sh) 35vs124	FI (sh)
Prec	81.24±3.1	80.16±2.0	81.75±2.4	81.22±1.8
Rec	80.80±2.4	79.88±1.5	81.11±1.4	81.47±1.2
Acc	83.02±1.9	82.76±1.6	83.79±1.1	83.88±0.6

As it can be seen from Table 2 shared, semi-shared and individual FMs show similar results, although shared FM is preferable than individual FM in our case when a small database is available, due to slightly better average recall and lower standard deviation of the results. Columns 2 and 3 show that, when using semi-shared FM, one should define FM for a meaningful group of classes, as done in column 3 when classes are grouped into two sets (3-5 and 1-2-4) according to the degree of non-stationarity of the corresponding types of sounds. On the contrary, in column 2, classes are divided simply to have an equal amount of data for each group, and the fusion performance is lower.

4.3. Feature and Decision Level Information Fusion

In this section, the two ways of information fusion mentioned in the Introduction are compared. For the feature-level fusion (see Fig.1 (a)), all ten types of features were used to feed the input of one SVM classifier. For the decision-level fusion (see Fig. 1 (b)), ten independent SVM-based classifiers were trained, one for each feature type. The ten input criteria, represented by these ten classifiers, were then combined by WAM operator and FI with learned shared FM. For the weights in WAM operator we use uniform class noise model with the weights computed as $\mu_i = E_i^{E_i}(1 - E_i)^{1-E_i}$ where E_i is the training error of class c_i [10].

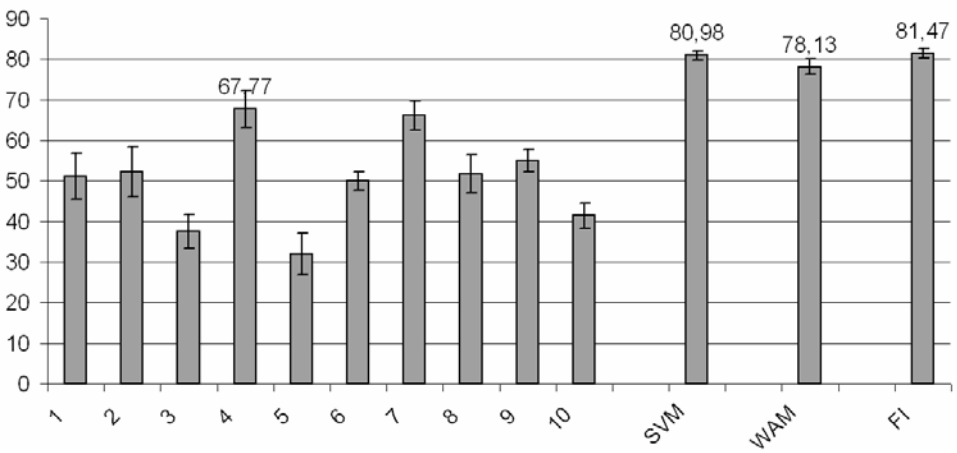

Figure 4. Recall measure for: the 10 SVM systems running on each feature type, the combination of the 10 features at the feature-level with SVM, and the fusion on the decision-level with WAM and FI operators.

As we can see from Fig. 4, all fusion approaches show a strong improvement in comparison to the SVM with the best single feature type (number 4). As expected, feeding all the features to the SVM classifier increased significantly the performance (SVM, 10 feature types). Interestingly enough, the fusion at the decision-level by FI showed comparable results to the powerful SVM classifier, which uses all the features. To gain an insight into the way FI works, we compare in Table 3 the individual recall score of the best feature type (column 2) for a given class, and the FI score (column 3) for the same class. Notice that, for the most represented class (A), the FI performance is lower, whereas for two less represented classes (C and D) it is higher. As the FM was trained using the errors of the particular classes as cost functions, we observe that, at the expense of accepting more errors for the most represented classes, the FI can recover a few errors for infrequent classes and thus obtain higher recall.

Table 3. Comparison of Individual Recall Scores for Each Class

Class	Best score	FI
A(119)	0.85	0.81
B(37)	0.61	0.61
C(40)	0.95	1.00
D(37)	0.77	1.00
E(12)	0.67	0.67

However, the accuracy and precision measures for both FI and WAM were slightly worse than that of the SVM: Accuracy=83.9 and Precision=81.2 for FI, versus Accuracy=84.8 and Precision=84.5 for SVM. Notice also that FI fusion has approximately a 10 times higher computation cost than SVM feature-level fusion (10 independent SVM classifiers vs. one), and therefore the latter would seem preferable in this case.

4.4. Feature ranking and selection

An information source consists of two parts: a classifier and a set of features. When using the same classifier for each information source, we can interpret the FM as the importance of features for the given classification task and we can use it for feature ranking and selection.

The information about both the importance of each feature type and the interaction among different feature types can be extracted applying the Shapley score to the FM. Using this approach, Fig. 5 shows that in our case the new feature type 6 (SBE correlations) is the most important followed by feature type 7 (SBE time difference). As both feature types measure the changes of the spectral envelope along the time, we can conclude that that information is of high importance. The only other feature type with importance score above the average is number 4 (SBE). Interestingly that although the new feature type has the highest overall importance, individual accuracy is only around 50% as it can be seen from Fig. 4. We also observed that without calculating the maximum absolute value in (10) the new feature individual accuracy increases to around 63% while the fusion result decreases to 79.4 %.

On the other hand, Fig. 6 shows the interaction among the feature types in our task; it can be seen that feature types 6 (SBE time correlation) and 7 (SBE time difference) express a negative interaction, which coincide with their similar character. As an extreme case, the light cell (4,5) has a large negative value and thus indicates a high competitiveness (redundancy) of the mentioned feature types. Therefore it would be better to consider only one of the two feature types. Actually, feature type 4 (SBE) and the feature type 5 (SBE Distribution) become roughly the same feature after using the SVM normalization. In a similar way, as feature types 1 (ZCR) and 8 (Sp. Centroid) are both targeting the “main” frequency, their cell is also rather light. Also, from the two lighter cells in the bottom of the Fig. 6, one can conclude that feature type 9 is redundant if feature types 8 and 10 are considered. On the contrary, feature types 4 and 6, or 4 and 7, or 4 and 10 seem to be highly complementary, and thus are preferable to be considered together.

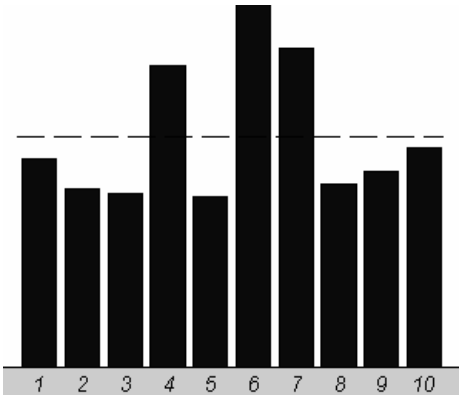

Figure 5. Importance of features extracted from FM. Dashed line shows the average importance level.

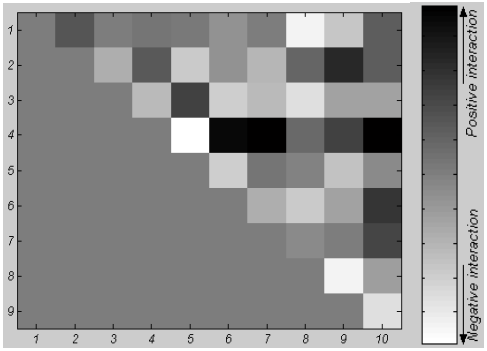

Figure 6. Interaction of features extracted from FM.

In the following AEC tests, we use the information from Fig. 5 and Fig. 6 to perform the feature selection. In the first test, we select the 5 best feature types according to the individual feature type importance (Method 1), while to select the 5 best features in the second test, both the individual feature type importance and the interaction indices are used (Method 2, see [19] for a detailed description). The selected features are then fed to the SVM classifier.

The performance of the SVM with all features is considered as a baseline in this part. It can be seen from the results in Table 4 that Method 1 did not lead to a better performance, while Method 2 obtained a slight improvement over the baseline.

Note that the recall score in the last column of Table 4 is much lower than the one shown in Fig. 4 (and last column in Table 2) for the FI technique when using the whole set of features, in spite of the fact that

Table 4. Classification Results Using Feature Selection Based on FM

	Support Vector Machines			FI
	Baseline	Method 1	Method 2	
Features	10 (all)	5(1,4,6,7,10)	5(4,6,7,8,10)	Method 2
Prec	84.50±2.1	82.76±1.3	86.14±1.7	81.74±2.4
Rec	80.98±1.1	75.31±2.1	80.14±1.6	74.79±2.5
Acc	84.83±2.3	83.97±2.2	85.86±1.4	83.79±1.6

FI technique apparently should benefit from a feature selection based on FM. There are two reasons for this behaviour. First, the 5 features have been selected according to a FM computed from 10 information sources, while FI scores in Table 4 result from a different FM, since it has been trained using only data corresponding to the 5 selected features. The second reason is based on the measure of uncertainty defined by (6). As it was mentioned in Section 2, if that entropy measure is close to 1 almost all information sources are equally used. In fact, for the 10 features case, it is 0.86, meaning that to achieve the results shown in Fig. 4, the FI operator uses in average 8-9 out of 10 information sources, so preserving only 50% of all features is not sufficient.

4.5. Fusion of different classifiers using FI

In previous sections we showed that the FI decision-level fusion obtains comparative results to the feature-level fusion using the SVM classifier. Indeed, from the computational cost point of view the feature-level fusion is preferred. However, when the resulting feature space has a too high dimensionality or when features are conveyed by different data types (strings, matrices, etc) the feature-level fusion is not an option.

On the other hand, it may be beneficial to combine the outputs of different well-established classification configurations for a given task; for example, the output of a SVM classifier which is discriminative but uses features from the whole signal with the output of a HMM generative classifier which considers time localized features. Based on that, we have tested with the FI formalism the combination of a SVM classifier that uses statistical (event-level) features with a HMM classifier that uses acoustic (frame-level) features. In these experiments, the best 5 feature types selected in the previous subsection by Method 2 are used with the SVM classifier. For HMM, we use a standard configuration coming from speech recognition: a 3 state left-to-right continuous density HMM model per class, with 8 Gaussians per state, and 13 frequency-filtered filter-bank energies (FFBE) [20] as features.

The first four columns in Table 5 show the performance of the SVM classifier and several HMM classifiers, where ΔFFBE means the time derivatives of FFBE features [20]. HMM-ΔFFBE gives low performance because the time derivatives only carry information about dynamics of sound but lack the basic static representation of the audio signal. The low score resulting from the HMM classifier when it uses as features both FFBE and their time derivatives (fourth column in Table 5), indicates that the amount of data we use is not enough to train the 26-dimensional vector data properly. Then, we decided to fuse the outputs of the previous classifiers: SVM, HMM-FFBE and HMM-ΔFFBE. From the second last column of Table 5 an improvement can be observed by FI fusion of the SVM and HMM-FFBE outputs. A further improvement is obtained by fusion of the SVM output with two information sources that separately give much lower individual performances, but use different features, as it is shown in the last column of Table 5.

Note from Fig. 4 that a much higher improvement was observed by fusing a larger number of information sources (10). Actually, the higher is the number of information sources the larger is the degree of interaction between them, and thus the better is the performance expected from the FI with an appropriately-learned FM. However, the difficulty of learning FM increases with the number of information sources. From our experience in this work, we would suggest to apply the FI formalism to fuse a number of information sources between 3 and 10.

5. CONCLUSION

In this work, we have carried out a preliminary investigation about the fusion of a relatively large number of information sources with the FI approach. We have shown an improvement over the baseline SVM approach in the task of classifying a small set of human vocal-tract non-speech sounds. By interpreting an information source as a specific combination of a classifier and a set of features, we have been able to carry out different types of tests.

Table 5. Individual performance of SVM, HMM on FFBE with and without time derivatives, and FI fusion

	SVM (1)	HMM-FFBE (2)	HMM-ΔFFBE (3)	HMM-FFBE+ΔFFBE	FI(1,2)	FI (1,2,3)
Prec	86.14±1.7	69.28±3.5	51.06±4.7	66.70±3.2	88.23±1.8	**89.47±1.9**
Rec	80.14±1.6	67.36±2.7	60.73±3.8	59.31±2.5	81.43±1.5	**82.43±1.0**
Acc	85.86±1.4	84.48±2.1	52.59±4.6	79.17±2.6	87.07±2.2	**87.93±1.8**

In the experiments, fusion of several information sources with the FI formalism showed a significant improvement with respect to the best single information source. Moreover, the FI decision-level fusion approach showed comparable results to the high-performing SVM feature-level fusion. The experimental work also indicated that the FI may be a good choice when feature-level fusion is not an option.

We have also observed that the importance and the degree of interaction among the various feature types given by the FM can be used for feature selection, and it gives a valuable insight into the problem. Actually, the information about importance and degree of interaction can be computed specifically for each class, thus allowing the selection of the set of features which are best fitted to a given class.

The use of semi-shared and individual FMs is a promising approach. Unfortunately, we were not able to explore it more due to the low size of the evaluation corpus. Future work will be devoted to the application of the FI to multi-microphone classification (and also detection) of acoustic events with a much larger dataset recorded in meeting rooms [17]. Also, a further improvement can come from the fact that signals captured from microphones placed at different positions in the room may carry different information about the acoustic events taking place in it.

6. ACKNOWLEDGEMENTS

The authors wish to thank Enric Monte for his valuable help and encouraging discussions. This work has been partially sponsored by the EU-funded project CHIL, IP506909, and by the Spanish Government-funded project ACESCA (TIN2005-08852).

7. REFERENCES

[1] A. Temko, C. Nadeu, "Classification of Acoustic Events using SVM-based Clustering Schemes", Pattern Recognition, volume 39, issue 4, pp.682-694, Elsevier, April 2006

[2] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[3] J. Weston, J. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik: “Feature Selection for SVMs”, Proc. of NIPS, 2000.

[4] L. Kuncheva, “’Fuzzy’ vs ‘Non-fuzzy’ in combining classifiers designed by boosting”, IEEE Transactions on Fuzzy Systems, 11 (6), pp. 729-741, 2003.

[5] M. Sugeno, Theory of fuzzy integrals and its applications, PhD thesis, Tokyo Institute of Technology, 1974.

[6] M. Grabisch, “Fuzzy integral in multi-criteria decision-making”, Fuzzy Sets & Systems 69, pp. 279-298, 1995

[7] S. Chang and S. Greenberg, "Syllable-proximity evaluation in automatic speech recognition using fuzzy measures and a fuzzy integral", Proc. of the 12th IEEE Fuzzy Systems Conf., pp. 828- 833 2003.

[8] M. Grabisch, "A new algorithm for identifying fuzzy measures and its application to pattern recognition". Proc. of 4th IEEE Int. Conf. on Fuzzy Systems, Yokohama, Japan, pp.145-50, 1995

[9] Y. Wu, E. Chang, K. Chang, J Smith., "Optimal Multimodal Fusion for Multimedia Data Analysis", Proc. ACM Int. Conf. on Multimedia, New York, pp.572-579, Oct. 2004.

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

[11] J-L. Marichal, "Behavioral analysis of aggregation in multicriteria decision aid, Preferences and Decisions under Incomplete Knowledge", Studies in Fuzziness and Soft Computing, Vol. 51, Physica Verlag, Heidelberg, pp. 153-178, 2000.

[12] M. Grabisch, "The Choquet integral as a linear interpolator", 10th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004), Perugia (Italy), pp.373-378, July 2004

[13] J-L. Marichal, "Entropy of discrete Choquet capacities", European Journal of Operational Research, 137 (3), pp. 612-624, 2002.

[14] I. Kojadinovic, J-L. Marichal, M. Roubens, "An axiomatic approach to the definition of the entropy of a discrete choquet capacity", 9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002), Annecy (France), pp.763–768, 2002.

[15] A. Noll, "Cepstrum Pitch Determination", Journal of the Acoustical Society of America, Vol. 41, No. 2, pp. 293-309, 1967.

[16] "Evaluation Packages for the First CHIL Evaluation Campaign", CHIL project Deliverable D7.4, downloadable from http://chil.server.de/servlet/is/2712/, Mar. 2005.

[17] A. Temko, C. Nadeu, J-I. Biel, "Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07", CLEAR'07 Evaluation Campaign and Workshop, Baltimore, MD, USA, to appear in Multimodal Technologies for Perception of Humans, LNCS, Springer

[18] C. Hsu, C. Lin, "A Comparison of Methods for Multi-class Support Vector Machines", IEEE Transactions on Neural Networks, Vol. 13, pp.415-425, 2002.

[19] L. Mikenina, H. Zimmermann, “Improved feature selection and classification by the 2-additive fuzzy measure”, Fuzzy Sets and Systems, 107:2, pp.197-218, 1999.

[20] C. Nadeu, D. Macho, J. Hernando, “Frequency and time filtering of filter-bank energies for robust HMM speech recognition”, Speech Communication, Vol. 34, pp. 93-114, 2001.