International Scholarly Research Network
ISRN Signal Processing
Volume 2011, Article ID 753819, 15 pages
doi:10.5402/2011/753819

Research Article

Classification of Emotional Speech Based on an Automatically Elaborated Hierarchical Classifier

Zhongzhe Xiao,[1,2] Emmanuel Dellandrea,[1] Weibei Dou,[3] and Liming Chen[1]

[1] *Université de Lyon, CNRS, Ecole Centrale de Lyon, LIRIS, UMR5205, 69134, France*
[2] *School of Physical Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China*
[3] *Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Emmanuel Dellandrea, emmanuel.dellandrea@ec-lyon.fr

Received 4 November 2010; Accepted 15 December 2010

Academic Editors: Y. H. Ha, R. Palaniappan, and F. Palmieri

Current machine-based techniques for vocal emotion recognition only consider a finite number of clearly labeled emotional classes whereas the kinds of emotional classes and their number are typically application dependent. Previous studies have shown that multistage classification scheme, because of ambiguous nature of affect classes, helps to improve emotion classification accuracy. However, these multistage classification schemes were manually elaborated by taking into account the underlying emotional classes to be discriminated. In this paper, we propose an automatically elaborated hierarchical classification scheme (ACS), which is driven by an evidence theory-based embedded feature-selection scheme (ESFS), for the purpose of application-dependent emotions' recognition. Experimented on the Berlin dataset with 68 features and six emotion states, this automatically elaborated hierarchical classifier (ACS) showed its effectiveness, displaying a 71.38% classification accuracy rate compared to a 71.52% classification rate achieved by our previously dimensional model-driven but still manually elaborated multistage classifier (DEC). Using the DES dataset with five emotion states, our ACS achieved a 76.74% recognition rate compared to a 81.22% accuracy rate displayed by a manually elaborated multistage classification scheme (DEC).

1. Introduction

Speech emotion analysis has attracted growing interest within the context of increasing awareness of the wide application potential of affective computing [1, 2]. Current machine-based techniques for vocal emotion recognition only consider classification problems of a finite number of discrete emotion categories [3] whereas the kinds of emotional states and their number are typically application dependent. These affective categories can be the six basic emotional states but also some nonbasic emotional classes, including for instance deception [4], certainty [5], stress [6], confidence, confusion, and frustration [7].

Most works in the literature made use of acoustic correlation of vocal emotion expressions and explore prosodic features, including pitch, energy, formants, cepstral, voice quality [8–10], and more recently Harmonic and Zipf features [11]. Moreover, the majority of them rely on one or several global one-step classifiers, including SVM, neural networks, and GMM, using the same feature set for all the emotional states, while studies on emotion taxonomy suggests that some discrete emotions are very close to each other on the dimensional emotion space, and there is confusion of emotion class borders. Indeed, Banse and Sherer evidenced [12] that acoustic correlates between fear and surprise or between boredom and sadness are not very clear, thus making an accurate emotion classification by a single step global classifier very hard. Implementing an intuition that hardly separable classes should be divided at last, Schuller et al. [13] proposed an interesting multilayer-SVM architecture. Some authors [14, 15] also tested an ensemble classification scheme making a voting by several base classifiers. However, only a very slight improvement of classification accuracy was observed. In our previous work [16], we proposed an effective multistage classification scheme driven by the dimensional emotion model which

hierarchically combines several binary classifiers. At each stage, a binary class classifier made use of a different set of the most discriminative features and discriminated emotional states according to different emotional dimensions. However, all these hierarchical classification schemes, including our own ones which is based on an empiric mapping of the discrete emotion states onto the dimensional emotion model, were manually elaborated to take into account the various emotional states under consideration whereas in practice, the types of emotion considered are rather application or dataset dependent. Clearly, we need an automatic way for building such a hierarchical classification scheme for machine-based emotion analysis, especially when the number of emotions changes and their types vary.

In this paper, we propose an automatically elaborated hierarchical classification scheme, which is driven by an evidence theory-based feature selection (ESFS), for the purpose of application-dependent emotions' recognition. Experimented on the Berlin dataset with 68 features and six emotion states, this automatically elaborated hierarchical classifier (ACS) showed its effectiveness, displaying a 71.38% accuracy rate compared to a 71.52% classification rate achieved by our previously dimensional model-driven but still manually elaborated multistage classifier (DEC). Using the DES dataset with five emotion states, our ACS achieved a 76.74% recognition rate compared to an 81% accuracy rate displayed by the manually elaborated multistage classification scheme (DEC). So far as we know, the best classification rates displayed in the literature are, respectively, 66% [17] and 76.15% [18] on the same DES dataset.

The remainder of this paper is organized as follows. Section 2 briefly introduces our evidence theory-based embedded feature selection scheme, the ESFS. We describe in Section 3 our ESFS-based algorithm for automatically deriving a hierarchical classification scheme (ACS) for application-dependent emotion analysis. Section 4 presents the experimental results of our ACS both on the Berlin and the DES datasets compared to the ones by our previously empirically elaborated but dimensional emotion model-driven hierarchical classification schemes (DECs). Finally, we summarize and conclude our work in Section 5.

2. ESFS: An Evidence Theory-Based SFS

Feature subset selection is an important subject when training classifier in Machine Learning (ML) problems. Practical ML algorithms are known to degrade in prediction accuracy when faced with many features that are not necessary [19, 20]. The current feature selection methods can be categorized into three broad classes according to their dependence to the underlying classifer [21]: filter approach, wrapper approach, or embedded one. In this section, we describe a novel embedded feature selection method, called ESFS, which is similar to the wrapper method SFS, since it relies on the simple principle to incrementally add the most relevant features. As an embedded method, our ESFS also carries out the selection of an optimal feature subset together with the classifier construction [22, 23]. Its originality concerns the use of mass functions from the evidence theory that allows merging elegantly information sources carried out by features, in an embedded way, thereby leading to a lower computational cost than the original SFS. We first introduce some basics of the evidence theory then describe our ESFS.

2.1. Introduction to the Evidence Theory. In our feature selection scheme, the term "belief mass" from the evidence theory is introduced into the processing of features. Dempster and Shafer wanted in the 1970s to calculate a general uncertainty level from the Bayesian theory. They developed the concept of "uncertainty mapping" to measure the uncertainty between a lower limit and an upper limit [24, 25]. Similar to the probabilities in the Bayesian theory, they presented a combination rule of the belief masses (or mass function) $m()$.

The evidence theory was completed and presented by Shafer in [26]. It relies on the definition of a set of n hypothesis Ω which have to be exclusive and exhaustive. In this theory, the reasoning concerns the frame of discernment 2^{Ω} which is the set composed of the 2^n subsets of Ω [27]. In order to express the degree of confidence we have in a source of information for an event A of 2^{Ω}, we associate to it an elementary mass of evidence $m(A)$. The elementary mass function or belief mass which presents the chance of being a true statement is defined as

$$m: \ 2\Omega \longrightarrow [0,1], \tag{1}$$

which satisfies

$$m(\Phi) = 0, \qquad \sum_{A \subseteq 2^{\Omega}} m(A) = 1. \tag{2}$$

The belief function is defined if it satisfies $\mathrm{Bel}(\Phi) = 0$ and $\mathrm{Bel}(\Omega) = 1$ and for any collection $A_1 \cdots A_n$ of subsets of Ω

$$\mathrm{Bel}(A_1 \cup \cdots \cup A_n) \geq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \phi}} (-1)^{|I|+1} \mathrm{Bel}\left(\bigcap_{i \in I} A_i\right). \tag{3}$$

The belief function shows the *lower* bound on the chances, and it corresponds to the mass function with the following formulae:

$$\begin{aligned} \mathrm{Bel}(A) &= \sum_{B \subseteq A} m(B), \quad \forall A \subset \Omega, \\ m(A) &= \sum_{A \subseteq 2^{\Omega}} (-1)^{|A-B|} \mathrm{Bel}(B), \end{aligned} \tag{4}$$

where $|X|$ means the number of elements in the subset.

The doubt function is defined as

$$\mathrm{Dou}(A) = \mathrm{Bel}(\neg A). \tag{5}$$

And the *upper* probability function is defined as

$$P^*(A) = 1 - \mathrm{Dou}(A). \tag{6}$$

The true belief in A should be between $\mathrm{Bel}(A)$ and $P^*(A)$.

The Dempster's combination rule can combine two or more *independent* sets of mass assignments by using orthogonal sum. For the case of two mass functions, let m_1 and m_2 be mass functions on the same frame Ω, the orthogonal sum is defined as $m = m_1 \oplus m_2$, to be $m(\phi) = 0$, and

$$m(A) = K \sum_{X \cap Y = A} m_1(X) \cdot m_2(Y),$$
$$K = \frac{1}{1 - \sum_{X \cap Y = \phi} m_1(X) \cdot m_2(Y)}. \tag{7}$$

For the case with more than two mass functions, let $m = m_1 \oplus \cdots \oplus m_n$, and it satisfies $m(\phi) = 0$ and

$$m(A) = K \sum_{\bigcap A_i = A} \prod_{1 \le i \le n} m_i(A_i),$$
$$K = \frac{1}{1 - \sum_{\bigcap A_i = \phi} \prod_{1 \le i \le n} m_i(A_i)}. \tag{8}$$

This definition of mass functions from the evidence is used in our model in order to represent the source of information given by each acoustic feature here and to combine them easily and to consider them as a classifier whose recognition value is given by the mass function.

2.2. The ESFS Scheme. An exhaustive search of the best subset of features, leading to explore a space of 2^n subsets, is impractical; we thus turn to a heuristic approach for the feature selection as does SFS. However, different from SFS which is wrapper-based approach, our evidence theory-based feature-selection technique, ESFS, makes use of the concept of belief mass from the evidence theory as a classifier and its combination rules to fuse various audio features, leading to an embedded feature-selection method. Moreover, as compared to the original SFS, the range of subsets to be evaluated in the forward process in ESFS is extended to multiple subsets for each size, and the feature set is reduced according to a certain threshold before the selection in order to decrease the computational burden caused by the extension of the subsets in the evaluation.

A heuristic feature selection algorithm can be characterized by its stance on four basic issues that determine the nature of the heuristic search process [28]. First, one must determine the starting point in the space of feature subsets, which influences the direction of search, and the operators used to generate successor states. The second issue concerns the search strategy. As an exhaustive search in a space of 2^n feature subsets is impractical, one needs to provide a more realistic approach such as greedy methods to traverse the space. At each point of the search, one considers local changes to the current state of the features, selects one, and iterates. The third issue concerns the strategy used to evaluate alternative subsets of features. Finally, one must decide on some criterion for halting the search.

As illustrated in Figure 1, we can summarize our embedded ESFS using belief masses by the following four steps while answering the previous four questions:

(i) computation of the belief masses of the single features from the training set,

(ii) evaluation and ordering of the single features to decide the initial set for potential best features,

(iii) combination of features for the generation of the feature subsets, making use of operators of combination,

(iv) selection of the best feature subset.

2.2.1. Calculation of the Belief Masses of the Single Features. Before the feature selection starts, all features are normalized into [0,1]. For each feature,

$$\mathrm{Fea}_n = \frac{\mathrm{Fea}_{n0} - \min(\mathrm{Fea}_{n0})}{\max(\mathrm{Fea}_{n0}) - \min(\mathrm{Fea}_{n0})}, \tag{9}$$

where Fea_{n0} is the set of original value of the nth feature and Fea_n is the normalized value of the nth feature.

By definition of the belief masses, the mass can be obtained in different ways which can represent the chance for a statement to be true. In our work, the PDFs (probability density functions) of the features of the training data are used for calculating the masses of the single features.

The curves of PDFs of the features are obtained by applying polynomial interpolation to the statistics of the distribution of the feature values from the training data.

Taking the case of a 2-class classifier as an example, the two classes are defined as subset A and subset A^C. First, the probability densities of the features in each of the 2 subsets are estimated from the training samples by the statistics of the values of the features in each class. We define the probability density of the kth feature Fea_k in subset A as $\Pr_k(A, f_k)$ and the probability density in subset A^C as $\Pr_k(A^C, f_k)$, where the f_k is the value of the feature Fea_k. According to the probability densities, the masses of feature Fea_k on these 2 subsets can be defined to meet the requirement in (2) as

$$m_k(A, f_k) = \frac{\Pr_k(A, f_k)}{\Pr_k(A, f_k) + \Pr_k(A^C, f_k)},$$
$$m_k\left(A^C, f_k\right) = \frac{\Pr_k(A^C, f_k)}{\Pr_k(A, f_k) + \Pr_k(A^C, f_k)}, \tag{10}$$

where at any possible value of the kth feature f_k, $m_k(A, f_k) + m_k(A^C, f_k) = 1$.

In the case of N classes, the classes are defined as $A_1, A_2, \ldots, A_N$. The masses of feature F_k of the ith class A_i can be obtained as

$$m_k(A_i, f_k) = \frac{\Pr_k(A_i, f_k)}{\sum_{n=1}^{N} \Pr_k(A_n, f_k)}, \tag{11}$$

which satisfies

$$\sum_{i=1}^{N} m_k(A_i, f_k) = 1. \tag{12}$$

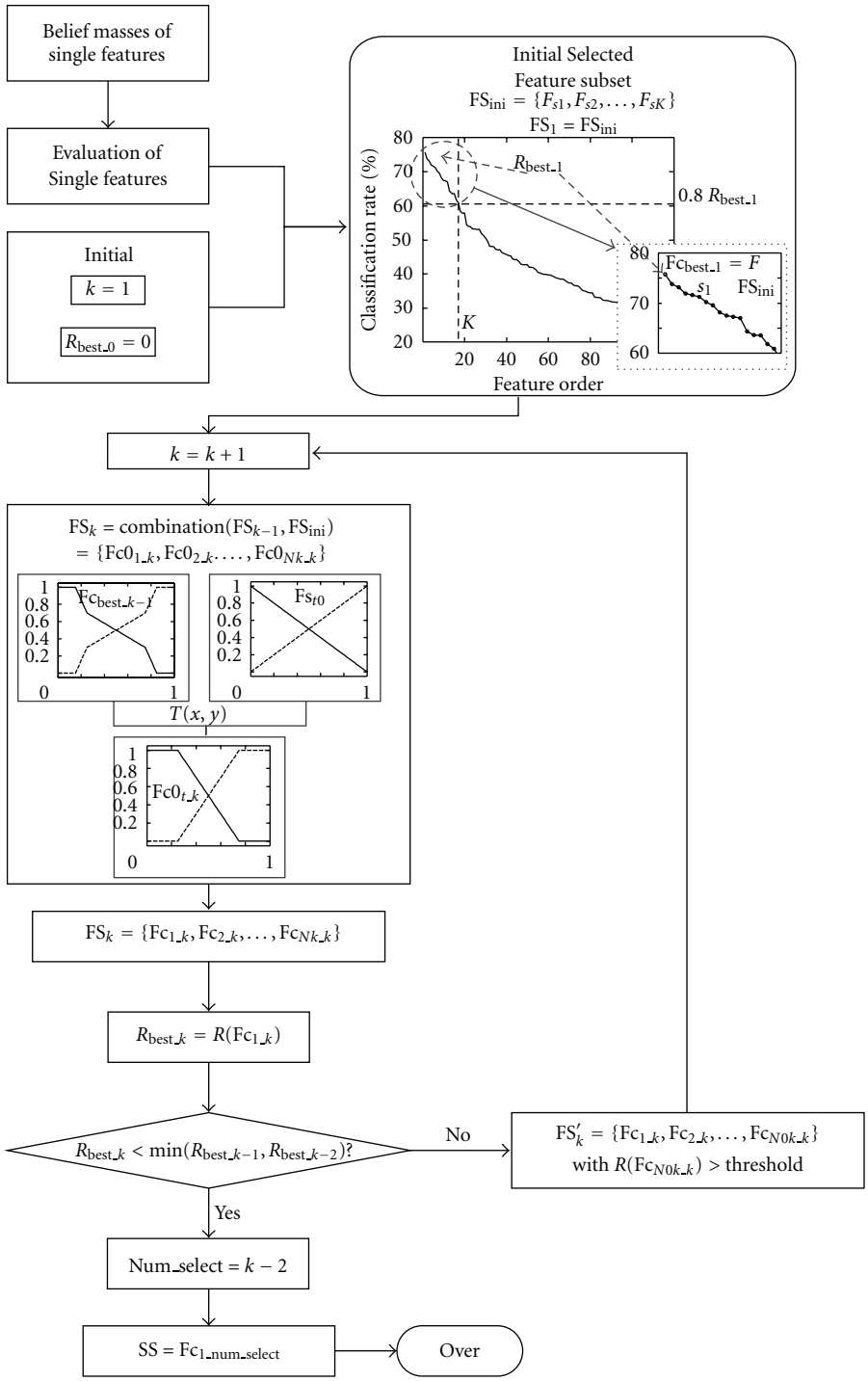


Figure 1: ESFS feature selection scheme.

2.2.2. Evaluation of the Single Features and Selection of the Initial Set of Potentially Best Features. Once computed the belief masses associated with each single feature from the training data, they are used to build a simple classifier. Indeed, given the belief masses associated with a single feature, a data sample can be simply assigned to the class having the highest belief mass. Using classification accuracy rate by these singles feature-based classifiers, all the features can then be sorted in a descending order as $\{F_{s1}, F_{s2}, \ldots, F_{sN}\}$, where N is the number of features in the whole feature set. In order to reduce the computational burden in the feature selection, an initial feature set FS_{ini} is selected with the first K best features in this ordered feature set, using for instance a threshold for cutting off classification accuracy rates, leading to

$$FS_{ini} = \{F_{s1}, F_{s2}, \ldots, F_{sK}\}. \quad (13)$$

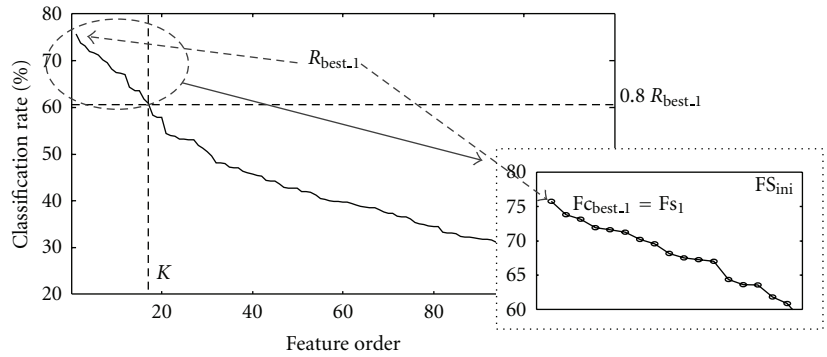


Figure 2: An example of ordered single features in FS_{ini}.

The threshold of the classification rates is decided according to the best classification rate as

$$R_{\text{single}}(F_{s\_K}) \geq \text{thres}_{\_1} * R_{\text{best}_1}, \tag{14}$$

where $R_{\text{best\_1}} = R_{\text{single}}(F_{s\_1})$, as illustrated in Figure 2. In our work on vocal emotion analysis, the threshold value $\text{thres}_{\_1}$ is set to 0.8 according to a balance between the overall performance and the calculation time by experiments. This threshold may vary with the underlying problem. Out of 68 features considered in our work, around 30 features are kept in our vocal emotion analysis problem while setting the threshold to 0.8.

Only the features selected in the set FS_{ini} will attend to the latter steps of the feature selection process. The elements (features) in FS_{ini} are considered as subsets of features having the size 1 at the same time.

2.2.3. Combination of Features for the Generation of the Best Feature Subsets. Having the best feature subsets with size $k-1$ ($k \geq 2$), the generation of a new feature subset of size k is achieved by computing a new composite feature through an operator of combination, thus fusing a feature subset of size $k-1$ with one from the initial feature set FS_{ini}. All these composite features are then sorted in a descending order according to their classification accuracy and the best ones are selected using a threshold as we did for the selection of the initial feature set FS_{ini}.

We note the set of all the feature subsets in the evaluation with size k as FS_k and the set of the selected feature subsets with size k as FS'_k. Thus, FS_1 equals to the original whole feature set, and $\text{FS}'_1 = \text{FS}_{\text{ini}}$. From $k = 2$, the set of the feature subsets FS_k is noted as

$$\begin{aligned} \text{FS}_k &= \text{Combine}(\text{FS}'_{k-1}, \text{FS}_{\text{ini}}) \\ &= \{\text{Fc0}_{1\_k}, \text{Fc0}_{2\_k}, \ldots, \text{Fc0}_{Nk\_k}\}, \end{aligned} \tag{15}$$

where the function "Combine" aims to generate new composite features by combining features from each of the two sets FS'_{k-1} and FS_{ini} with all the possible combinations except the ones in which a feature from FS_{ini} already appears in the composite feature from FS'_{k-1}; Fc0_{n_k} represents the generated new composite features using an operator of combination, and Nk is the number of elements in the set FS_k.

The creation of a new composite feature from two other features is achieved by combining the belief masses of the two features, making use of an operator of combination. The fusing process works as follows.

Assume that N classes are considered in the classifier. For the ith class A_i, the preprocessed mass m^* for the new composite feature Fc0_{t_k}, which is generated from a composite feature Fc_{x_k-1} in FS'_{k-1} and a feature Fs_y from FS_{ini}, $\text{Fc0}_{t\_k} = \text{Combine}(\text{Fc}_{x\_k-1}, \text{Fs}_y)$, is calculated as

$$m^*(A_i, \text{fc0}_{t\_k}) = T\Big(m(A_i, \text{fc}_{x\_k-1}), m\Big(A_i, \text{fs}_y\Big)\Big), \tag{16}$$

where the f_x is the value of the feature F_x and $T(x, y)$ is an operator of combination. The commonly used existing operators for fusing two elements, the triangle norms, are used in our work to combine the features. These operators will be explained in details in next subsection. The sum of m^*s may not be 1 depending upon the combination operator being used. In order to meet the definition of belief masses, the m^*s can then be normalized as the masses for the new composite feature

$$m(A_i, \text{fc0}_{t\_k}) = \frac{m^*(A_i, \text{fc0}_{t\_k})}{\sum_{n=1}^{N} m^*(A_n, \text{fc0}_{t\_k})}. \tag{17}$$

The performance of the new composite feature may be better than both its two base features used in the combination, as illustrated in Figure 3. However, the new composite feature may also perform worse than any of the two original features, in which case the new composite feature will be eliminated in the feature selection process.

All these new composite features can also be sorted in a descending order according to their classification accuracy on the training dataset as we did for the single original audio features

$$\begin{aligned} \text{FS}_k &= \{\text{Fc0}_{1\_k}, \text{Fc0}_{2\_k}, \ldots, \text{Fc0}_{Nk\_k}\} \\ &= \{\text{Fc}_{1\_k}, \text{Fc}_{2\_k}, \ldots, \text{Fc}_{Nk\_k}\}. \end{aligned} \tag{18}$$

The best composite feature having size k is noted as $\text{Fc}_{\text{best\_}k} = \text{Fc}_{1\_k}$, and its classification accuracy recorded as

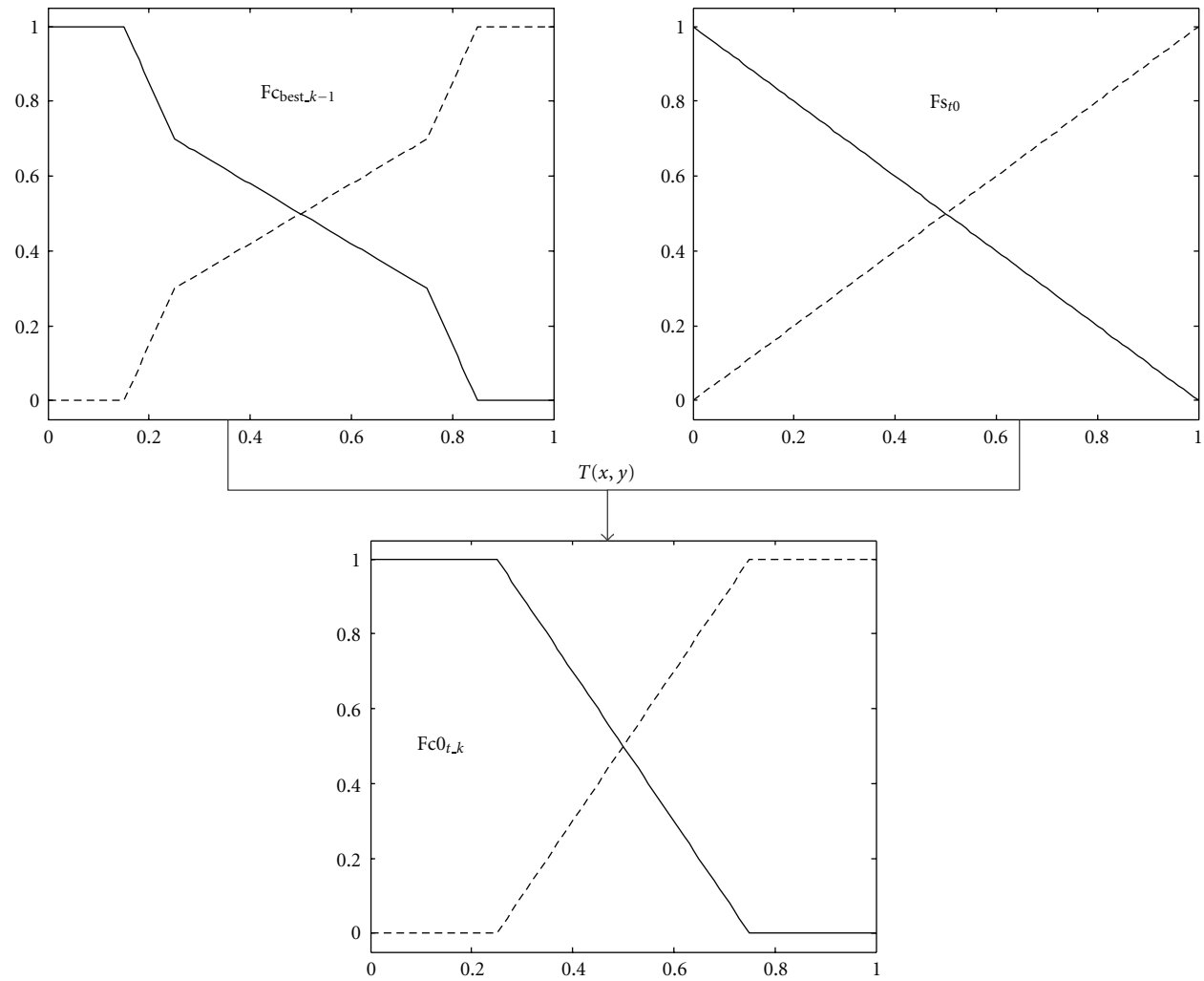


Figure 3: Example of masses of a new composite feature in the case of 2 classes.

$R_{\text{best\_}k}$. Similar to the selection of FS_{ini} from the single original features, a threshold is set to select a number of composite features having the size k for the next step of forward selection. The set of these selected composite features is noted as

$$\text{FS}'_k = \{\text{Fc}_{1\_k}, \text{Fc}_{2\_k}, \ldots, \text{Fc}_{N0k\_k}\}, \tag{19}$$

which satisfies $R(\text{Fc}_{N0k\_k}) \geq \text{thres}_{\_k} * R_{\text{best\_}k}$. In order to simplify the selection, the threshold value $\text{thres}_{\_k}$ is set in our work to the same value as 0.8 in every step without any adaptation to each step.

2.2.4. Stop Criterion and the Selection of the Best Feature Subset. The stop criterion of ESFS occurs when the best classification rate begins to decrease while increasing the size of the feature subsets. In our work, in order to avoid missing the real peak of the classification performance, the forward selection stops when the classification performance continues to decrease in two steps, $R_{\text{best\_}k} < \min(R_{\text{best\_}k-1}, R_{\text{best\_}k-2})$. The number of the selected composite features is noted as Num_select, and the selected composite features are

$$\begin{aligned}\text{SS} = \text{Fc}_{1\_\text{Num\_select}} &= \text{Combine}\left(\text{Fc}_{x\_\text{Num\_select}-1},\ \text{Fs}_y\right) \\ &= \cdots = \text{Combine}\left(\text{Combine}\cdots\text{Combine}\left(\text{Fs}_p, \text{Fs}_q\right)\right).\end{aligned} \tag{20}$$

2.3. Operators of Combination. Aggregation and fusion of different information sources are basic concerns in many systems and applications. There exists different fusion approaches, including evidence theory, possibility theory, or fuzzy set theory, but all these approaches can be summarized as application of some numerical aggregation operators. Generally speaking, the aggregation operators are mathematical functions consisting of reducing a set of numbers into a unique representative number [29].

Since the combination of the masses of the features in our feature selection scheme amounts to combine two features, the commonly used existing operators for two elements, the

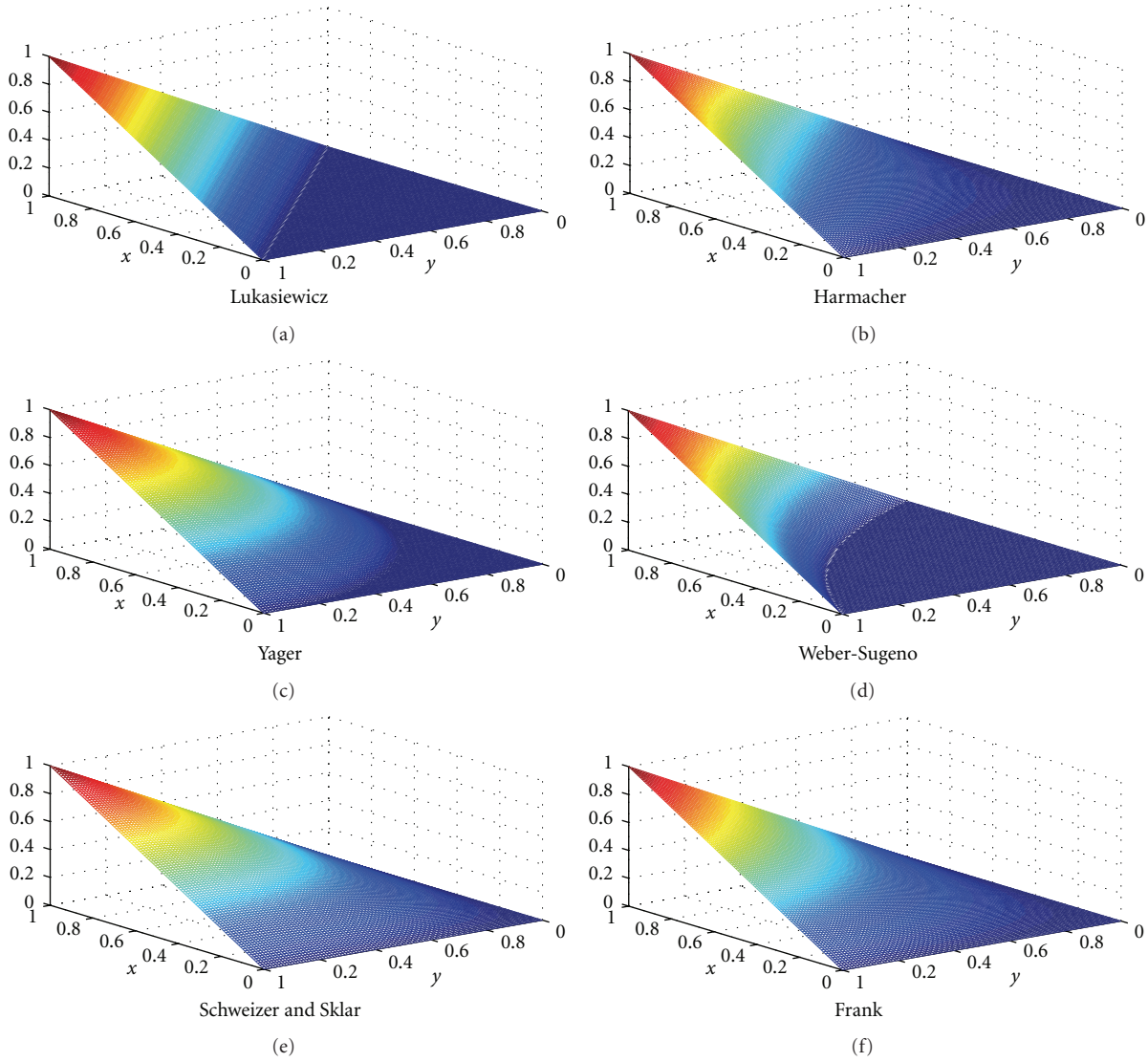


Figure 4: The property curve surfaces of the operators.

triangle norms, are used in our work to fuse the features as in (16).

The triangular norm (abbreviated as t-norm) is a kind of binary operation used in the framework of probabilistic metric spaces and in multivalued logic which was first introduced by Menger [30] in order to generalize the triangular inequality of a metric. The current concept of a t-norm and its dual operator (t-conorm) is developed due to Schweizer and Sklar [31, 32]. The t-norms generalize the conjunctive "AND" operator and the t-conorms generalize the disjunctive "OR" operator. These properties enable them to be used to define the intersection and union operations [29, 33].

The definitions of a t-norm and a t-conorm are as follows.

t-norm. A t-norm is a function $T : [0,1] \times [0,1] \to [0,1]$, having the following properties

(1) $T(x,y) = T(y,x)$ (T1) commutativity,

(2) $T(x,y) \le T(u,v)$, if $x \le u$ and $y \le v$ (T2), monotonicity (increasing),

(3) $T(x,(T(y,z)) = T(T(x,y),z)$ (T3), associativity,

(4) $T(x,1) = x$ (T4), one as a neutral element.

A well-known property of t-norms is

$$T(x,y) \le \min(x,y). \tag{21}$$

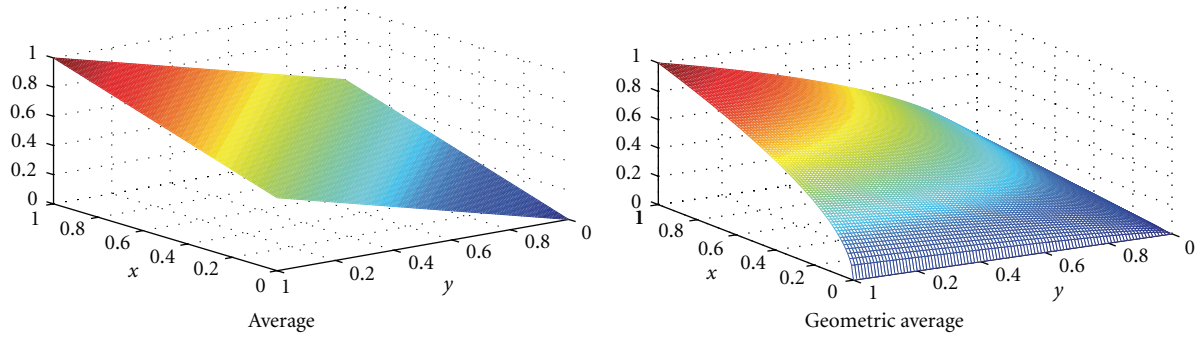
t-conorm. Formally, a t-conorm is a function $S[0,1] \times [0,1] \to [0,1]$, having the following properties:

(1) $S(x,y) = S(y,x)$ (S1), commutativity,

(2) $S(x,y) \le S(u,v)$, if $x \le u$ and $y \le v$ (S2), monotonicity (increasing),

(3) $S(x,(S(y,z)) = S(S(x,y),z)$ (S3), associativity,

(4) $S(x,0) = x$ (S4) Zero as a neutral element.



(a) (b)

Figure 5: The property curve surfaces of average and geometric average.

A well known property of t-conorms is

$$S(x, y) \geq \text{Max}(x, y). \tag{22}$$

We say that a t-norm and a t-conorm are dual (or associated) if they satisfy the DeMorgan law.

$$1 - T(x, y) = S(1 - x, 1 - y). \tag{23}$$

The minimum is the biggest t-norm; and its dual is the smallest t-conorm.

Six parameterized t-norms, namely Lukasiewicz, Hamacher, Yager, Weber-Sugeno, Schweizer and Sklar, and Frank, which are frequently proposed in the literatures [34], were tested with different parameters in our work. They are defined as follows:

(1) Lukasiewicz

$$T(x, y) = \max(x + y - 1,\ 0), \tag{24}$$

(2) Hamacher

$$T(x, y) = \frac{x \cdot y}{\gamma + (1 - \gamma) \cdot (x + y - x \cdot y)}, \quad \gamma \geq 0, \tag{25}$$

(3) Yager

$$T(x, y) = \max\left(1 - \left[(1 - x)^p + (1 - y)^p\right]^{1/p}\right), \quad p > 0, \tag{26}$$

(4) Weber-Sugeno

$$T(x, y) = \max\left(\frac{x + y - 1 + \lambda_T \cdot x \cdot y}{1 + \lambda_T}, 0\right), \quad \lambda_T > -1, \tag{27}$$

(5) Schweizer and Sklar

$$T(x, y) = 1 - \left[(1 - x)^q + (1 - y)^q - (1 - x)^q (1 - y)^q\right]^{1/q}, \quad q > 0, \tag{28}$$

(6) Frank

$$T(x, y) = \log_s\left[1 + \frac{(s^x - 1) \cdot (s^y - 1)}{s - 1}\right], \quad s > 0,\ s \neq 1. \tag{29}$$

Figure 4 depicts the various surfaces associated with the aforementioned six combination operators. The z-axis represents the output of the operators from the two inputs x and y. As we can see from these curves, the Yager, Weber-Sugeno, and Schweizer and Sklar operators have convex surfaces, while the Lukasiewicz and Frank operators have flatter or even concave surfaces. For each operator, the degree of convexity or concavity is affected by the parameters. The difference in the shape of the surfaces may influence the performance when they are applied in the classification.

In addition to these t-norm operators, the average and the geometric average of the features are also used for the combination of the features.

(i) Average

$$A(x, y) = \frac{(x + y)}{2}. \tag{30}$$

(ii) Geometric average

$$G\_a(x, y) = \sqrt{x \cdot y}. \tag{31}$$

The property curve surfaces of average and geometric average are displayed in Figure 5.

It should be noticed that since a step of normalization is applied in calculating the masses of the combined new features in (17), the "associativity" property of the t-norms is not effective in our case. Furthermore, in order to ensure the performance of the final combined new feature, the order of the selected features cannot be moved randomly.

2.4. Discussion. ESFS can be used either as an embedded feature selection method, as we do in the next section when building the hierarchical classification scheme for emotion analysis, or as a simple filter method for selection of relevant features which can then be embedded into classifiers. Used as a filter method, we carried out experiments aiming at

comparing the behavior of our ESFS with other filter feature-selection techniques, including Fischer filter method, PCA, and SFS. Using Berlin dataset for emotional speech recognition and Simplicity dataset for visual object recognition, our ESFS displayed better performance, showing its effectiveness in the selection of relevant features [35].

3. ESFS-Based Hierarchical Classification Scheme for Vocal Emotion Recognition

The fuzzy neighborhood relationship between some emotional classes, for instance between sadness and boredom, as evidenced by studies on acoustic correlates, leads to unnecessary confusion between emotion states when a single global classifier is applied using the same set of features. While several previous works have shown the effectiveness of multistage classification schemes on vocal emotion analysis, the elaboration of these hierarchical classification schemes were intuitive and manual. On the other hand, the number of emotions and their types to be recognized are typically dataset or application dependent. The empirically built hierarchical classification structure thus needs to be adjusted when the emotional space changes. In this section, we propose an automatically elaborated Hierarchical Classification Scheme (ACS) which is driven by our evidence theory-based feature selection technique ESFS. While keeping at least similar performance, the main goal here is to avoid unnecessary repeated work for manually building a new multistage classification scheme each time the vocal emotions to be analyzed change.

Basically, our ESFS, when applied as an embedded feature selection technique to an application specific vocal emotion recognition problem, automatically divides in an optimal way the set of emotional states to be recognized into two disjoint subsets of emotional states, leading to a hierarchical classifier represented by a binary tree whose root is the union of all emotion classes, while leaves are single emotion classes and intermediate nodes composite emotional classes discriminated by a subclassifier. Each of these subclassifiers is based on our ESFS introduced in the previous section, thus extracting the best features to best discriminate two composite emotional classes.

The generation process of an ACS is shown in Figure 6. The N discrete emotional classes concerned in the classification problem are first assigned to a frame of discernment $\Omega = \{E_1, E_2, \ldots, E_N\}$, where E_n stands for the nth emotional state in the frame of discernment Ω. For example, the frame of discernment associated with the Berlin database is $\Omega_{\text{Belin}} = \{\text{Anger, Happiness, Fear, Neutral, Sadness, Boredom}\}$ while the frame of discernment associated with the DES dataset is $\Omega_{\text{DES}} = \{\text{Anger, Happiness, Neutral, Surprise, Sadness}\}$. The frame of discernment describes the initial affect space under study. Using our embedded ESFS, the affect space will be recursively divided into two complementary subaffect spaces which best describe the affect space with respect to the training data, until the subaffect spaces become simple emotional classes.

The hierarchical classifier is thus expressed by a binary tree. The initial frame of discernment is set as the root node

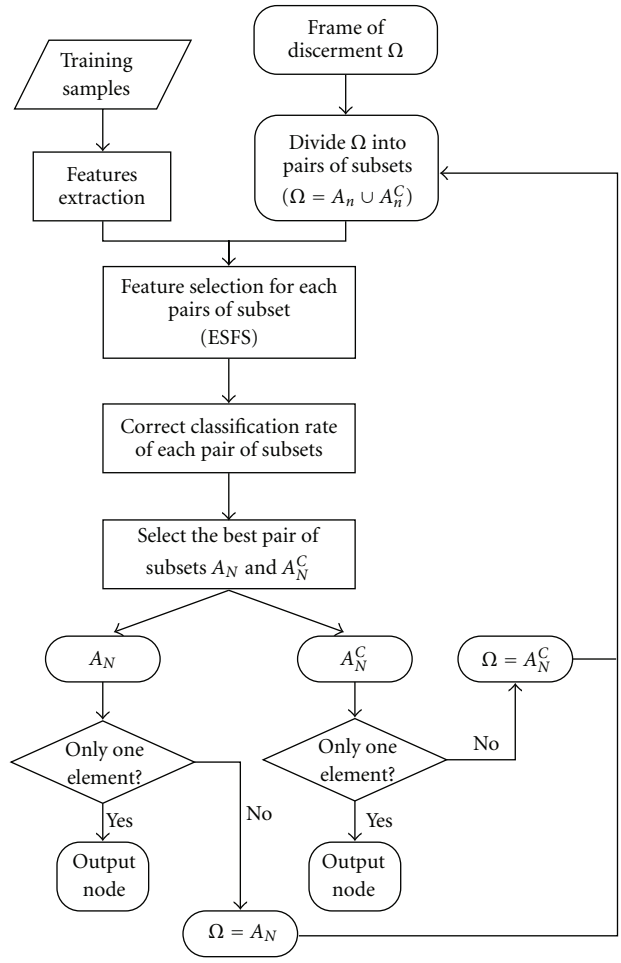


Figure 6: Generation of a hierarchical classifier.

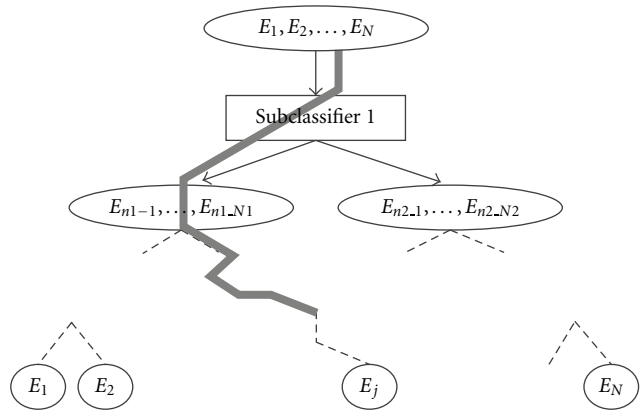


Figure 7: Hierarchical classifier and recognition route.

of this binary tree. The main steps for generating an ACS are listed as follows.

3.1. The Algorithm

Step 1. The hierarchical structure is composed of several binary subclassifiers. The Ω is divided into pairs of nonempty subsets exhausting all possible partitioning of the initial

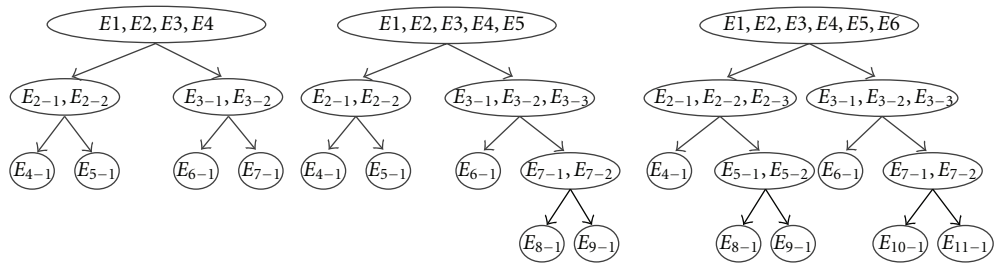


Figure 8: Typical balanced ACS classifiers for 4, 5, 6 classes.

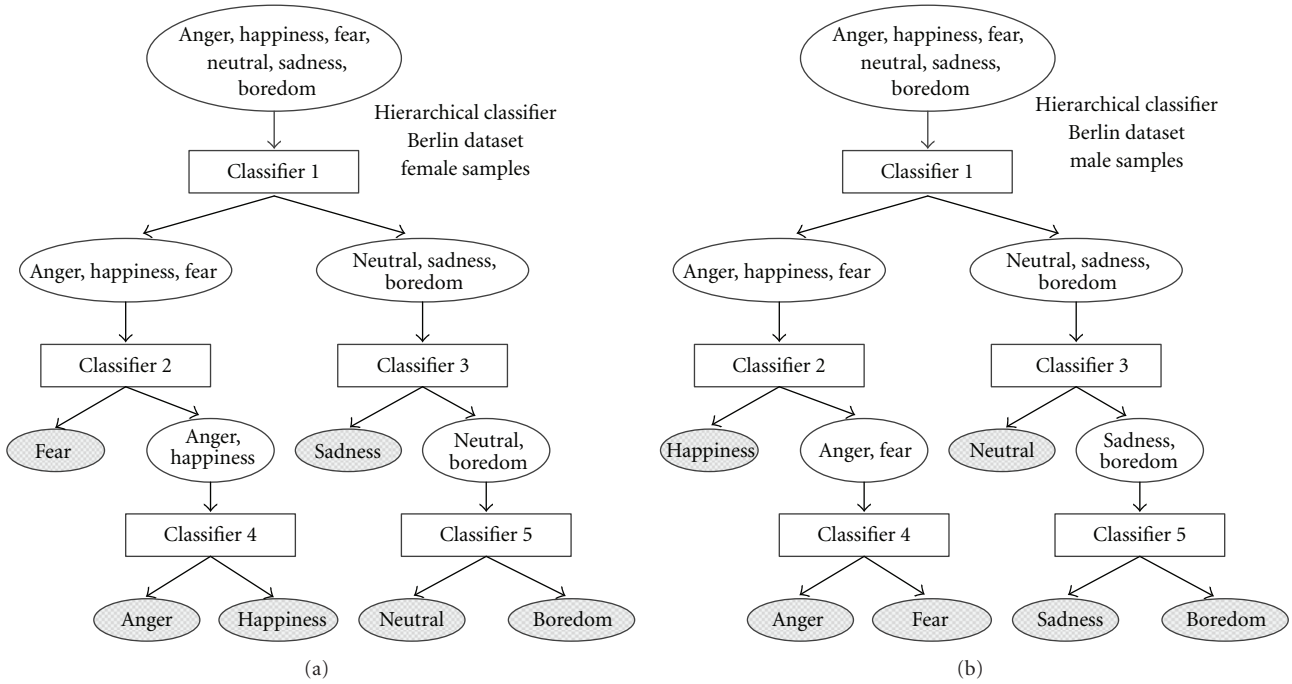


Figure 9: Hierarchical classifiers, respectively, for male and female for the Berlin dataset.

Table 1: list of pairs of subsets for 4 classes: E_1, E_2, E_3, and E_4.

Index of pairs	A_n	A_n^C
1	E_1	E_2, E_3, E_4
2	E_2	E_1, E_3, E_4
3	E_3	E_1, E_2, E_4
4	E_4	E_1, E_2, E_3
5	E_1, E_2	E_2, E_3
6	E_1, E_3	E_2, E_4
7	E_1, E_4	E_2, E_3

affect space. The two subsets in each pair are complements to each other, and each subset represents a class with respect to the initial affect space Ω

$$\Omega = A_n \cup A_n^C, \quad A_n \neq \phi,\ A_n^C \neq \phi. \tag{32}$$

All the possible pairs of complements are evaluated using ESFS to decide which partitioning of the initial affect space is the best from the viewpoint of classification accuracy. In order to avoid repeated partitioning, the pairs are defined to ensure that the number of elements in subset A_n is not larger than the number of elements in A_n^C. For example, in the case with 4 classes, 7 pairs of subsets can be evaluated as listed in Table 1.

Our feature combination and selection process (ESFS) is applied to each pair of the subsets and the belief masses of the training samples in the subsets can be obtained. All these pairs can then be sorted by their classification accuracy rates.

Step 2. The two subsets in the pair with the highest classification rate (assuming it is the nth pair of subsets) are assigned as the children nodes: A_n as the left child node and A_n^C as the right child node.

The two children nodes of A_n and A_n^C are then processed in the same way as in Step 1. The numbers of elements in the children nodes A_n/A_n^C are counted. Note the subsets A_n or A_n^C as A^*.

(i) If $\text{Size}_{A^*} = 1$ (only one element in the subset), this node is marked as a leaf node.

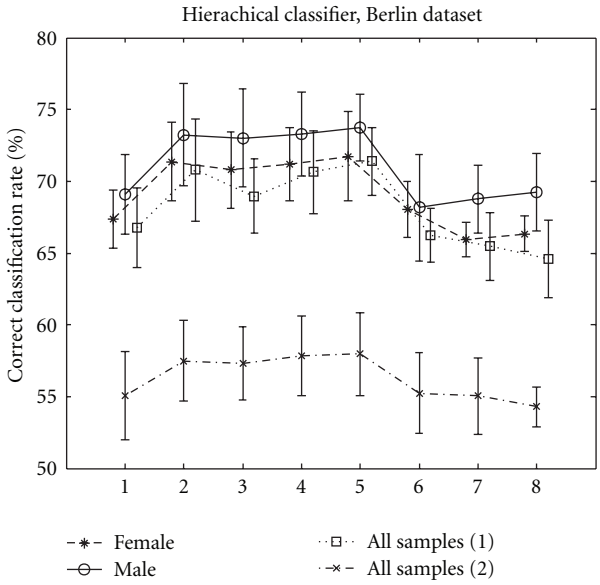


Figure 10: Classification rate with HCS on Berlin dataset, The indexes on the X axis stand for the operators: (1) Lukasiewicz, (2) Hamacher, (3) Yager, (4) Weber-Sugemo, (5) Schweizer and Sklar, (6) Frank, (7) Average, (8) Geometric Average. "All samples (1)" refers to the classification on the mixed genders samples with gender classification, and "All samples (2)" refers to the classification on the mixed genders samples without gender classification.

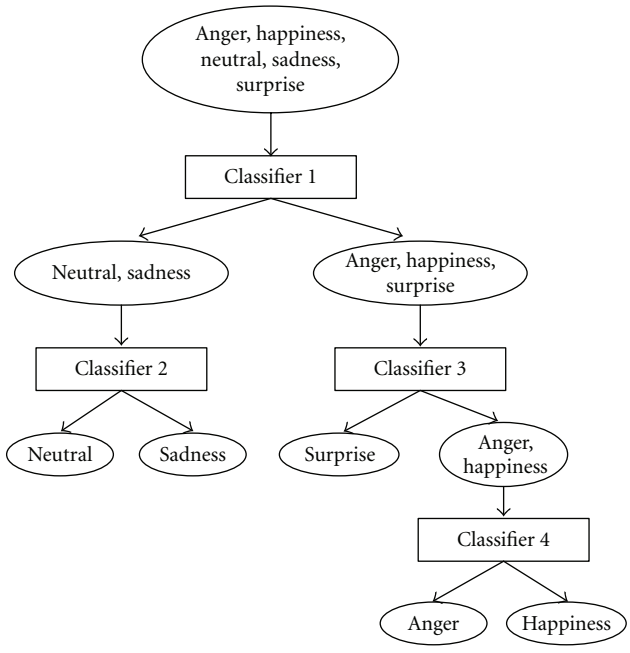


Figure 11: ACS classification scheme for the DES dataset.

(ii) If $\text{Size}_{A^*} > 1$ (the subset can be further partitioned), the frame of discernment is updated as $\Omega = A^*$, and the construction of the binary tree continues with Step 1.

Step 3. When the number of leaf nodes equals to the number of emotional classes, the generation process of the binary tree stops. The information about the binary tree is stored in the model of the classifier.

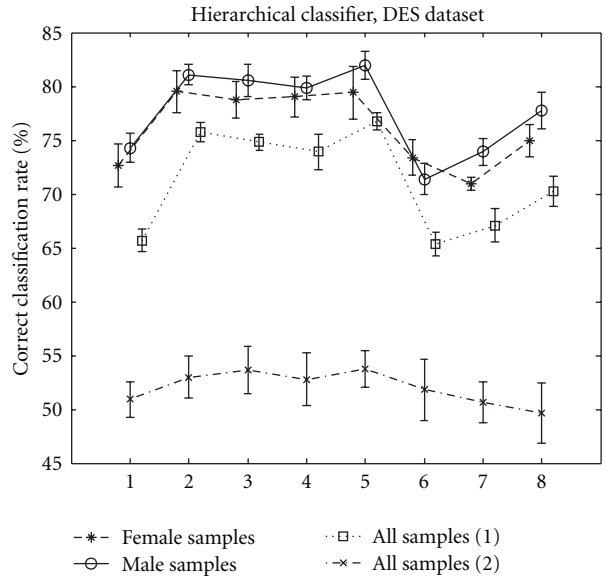


Figure 12: Classification rate with ACS on the DES dataset.

3.2. Practice and Improvement. In practice, we want our ACS resulted from the previous scheme to be as balanced as possible. Indeed, the overall classification accuracy rate of a multistage hierarchical classifier is approximately the product of the classification rates at each stage. Assuming the different stages in the classifier have classification accuracy rates close to each other as R_{stage}, for an n stage classifier, the overall classification rate can be approximated by R^n_{stage}. Thus, too many stages may lead to dramatic degrading of the overall classification accuracy rate. In order to reach classification accuracy as higher as possible, one needs to reduce the depth of the tree-based hierarchical classifier so that it is a balanced structure.

In our work, balanced pair of subsets is put forward. For each pair of subsets A_n and A^C_n, a subset distance is calculated as the difference of the number of elements of the two subsets

$$D_n = \text{Size}_{A^C_n} - \text{Size}_{A_n}, \tag{33}$$

with

$$\text{Size}_{A^*} = |A^*|, \tag{34}$$

where the $|X|$ means the number of elements in the set X.

Because the subsets A_n/A^C_n are defined so that A_n has alsways fewer or the same number of elements than A^C_n, D_n always satisfies

$$D_n \geq 0. \tag{35}$$

when the nth pair of subsets satisfies $D_n \leq 1$, it is defined as a balanced pair of subsets.

If the pair of subsets with the highest classification rate (assuming that it is the n_1th pair and the classification rate

Table 2: Thresholds according to the highest classification rate (%).

R_{n1}	95.8	94.8	93.6	92.6	**91.4**	**90.2**	**89**	87.8	86.6
thre_diff	2	2.5	3	3.5	**4**	**4.5**	**5**	5.5	6

is R_{n1}) is a balanced pair, the generation of the binary tree continues normally; if it is not a balanced pair, it will be compared with the balanced pair having the best classification rate (assuming that it is the n_2th pair and the classification rate is R_{n2}). If we have only five or six classes in our applications as it is the case for the Berlin and DES datasets, there should be two or three stages in a balance binary tree. We thus set a threshold thre_diff defining an acceptable difference on classification accuracy which can be measured by $(R_{n1} - \text{thre\_diff})^2 \geq (R_{n1})^3$, assuming the number of the stages does not exceed three. The approximate values of thre_diff related to R_{n1} are listed in Table 2. The best classification rate R_{n1} in the first stage of the hierarchical structure is normally around 90% in the experiments, so the most commonly selected thre_diff is between 4% and 5%. When the number of classes increases in the classification problems, the thresholds should be adjusted according to the number of classes.

If $R_{n1} - R_{n2} <$ thre_diff, the binary tree with the balanced pair is assumed to have better overall performance in the classification, and the n_1th pair is selected instead of the n_1th pair. However, when an unbalanced pair has much better recognition rate than the balanced one, it will be still selected.

The common structure of the ACS generated by this approach is shown in Figure 7. The grey doubled line illustrates the possible recognition route of an audio sample.

If the number of affect classes varies from 3 to 7 as it is the case for most affect recognition problems currently studied, Figure 8 illustrates some typical hierarchical classifier schemes with balanced pairs of subsets.

4. Experimental Results

The effectiveness of our approach is experimented both on the Berlin and DES datasets. In the following, we first introduce the audio features. Then, our experimental results are presented and discussed.

4.1. The Feature Set. We consider the same set of 68 features as in [11, 16], covering popular frequency and energy-based features as well as our newly introduced features, namely harmonic features for a better description of voice timbre pattern, and Zipf features for a better rhythm and prosody characterization. They are the following.

Frequency-Based Features

1–20. Mean, maximum, minimum, and median value and the variance of F0 and the first 3 formants.

Energy-Based Features

21–23. Mean, maximum, and minimum value of energy

24. Energy ratio of the signal below 250 Hz

25–32. Mean, maximum, median, and variance of the values and durations of energy plateaus

33–40, 42–49. Statistics of gradient and durations of rising and falling slopes of energy contour

41–50. Number of rising and falling slopes of energy contour per second

Harmonic Features

51–63. Mean, maximum, variance and normalized variance of the 4 areas

64–66. The ratio of mean values of areas 2∼4 to area 1

Zipf Features

67. Entropy feature of Inverse Zipf of frequency coding

68. Resampled polynomial estimation Zipf feature of UFD (Up-Flat-Down) coding

4.2. Experimental Results on the Berlin Dataset. Hold out cross-validation with 10 iterations were carried out on the Berlin dataset. In each of the iterations, 50% of samples were used as training set and the other 50% of samples as test set. Out of the seven basic emotions in the Berlin dataset, we excluded "disgust" as there are only 8 samples of "disgust" in the male samples, which is much less than the other emotional classes. Moreover, the acoustic features for this emotion was shown to be inconsistent [12]. The influence of gender information on the emotion classification accuracy was also highlighted. For each classification scheme, three experimental settings, using only the female speech samples, the male speech samples, and a combination of all the samples (mixed samples), respectively, were evaluated and compared. Figure 9 illustrates the two hierarchical classification schemes automatically generated by the previous ESFS driven ACS, respectively for male and female.

Figure 10 displays the best classification rates achieved by the eight combination operators that we tested. The error bars in the figure show the root mean square errors of the classification rates.

The best classification accuracy is 71.75% ± 3.10% for the female samples, 73.77% ± 2.33% for the male samples, and 71.38% ± 2.33% for the mixed genders with gender classification and 57.95% ± 2.87% without gender classification. These results are quite closed to the ones achieved by our manually elaborated but driven by the dimensional emotion model multistage classification scheme DEC [11, 16]. All of the best results are obtained with Schweizer and Sklar operator. From the two curves "All samples (1)" and "All samples (2)", we can see that a preprocessing of the audio samples for gender classification obviously improve the overall classification performance for the mixed gender samples. We also have discovered that the fusion operators

Table 3: Comparison between the DEC and HCS on the two datasets (%).

		Female samples	Male samples	Mixed without gender information	Mixed with gender information
Berlin dataset	DEC	71.89 ± 2.97	75.75 ± 3.15	68.60 ± 3.36	71.52 ± 3.85
	ACS	71.75 ± 3.10	73.77 ± 2.33	57.95 ± 2.87	71.38 ± 2.33
DES dataset	DEC	85.14 ± 2.02	87.02 ± 1.44	57.34 ± 1.79	81.22 ± 1.27
	ACS	79.54 ± 1.95	81.96 ± 1.27	53.75 ± 1.71	76.74 ± 0.83

Hamacher, Yager, Weber-Sugemo, and Schweizer and Sklar having properties of convex curve surfaces perform better.

4.3. Experimental Results on the DES Dataset. Our ACS described in Section 3 was also benchmarked on the DES dataset and Figure 11 illustrates the automatically generated hierarchical classification scheme which proves to be the same for both the two genders. Similar to the hierarchical classification schemes generated on the Berlin dataset, the first stage of the ACS scheme proceeds in the arousal dimension (or energy/active dimension) with the separation between neutral and sadness versus anger and happiness and surprise. For the three active emotions, surprise is separated from anger and happiness in the second stage which is still in the arousal dimension. Anger and happiness are separated in the appraisal dimension in the last stage in the hierarchical framework as for the female samples in Berlin dataset.

Holdout cross-validations with 10 iterations are used in our experiments on DES dataset. In order to compare with previous works [17, 36–38], for each iteration of experiments, 90% of segments were used as training set and the remaining 10% used as testing set. The training set and testing set were selected randomly in each group. Figure 12 shows the best classification rates of the eight combination operators that we experimented. The classification rates for the case of all speech samples from both genders are obtained by adding an automatic gender classifier [39] as we did in the experiments on the Berlin dataset. The error bars in the figure show the root mean square errors of the classification rates.

The indexes on the X axis stand for the operators (1) Lukasiewicz, (2) Hamacher, (3) Yager, (4) Weber-Sugemo, (5) Schweizer and Sklar, (6) Frank, (7) Average, and (8) Geometric Average. "All samples (1)" refers to the classification on the mixed genders samples with gender classification, and "All samples (2)" refers to the classification on the mixed genders samples without gender classification.

The best result is 79.54% ± 1.95% for the female samples with Hamacher operator when $\gamma = 3$, 81.96% ± 1.27% for the male samples with Schweizer and Sklar operator when $q = 1$, 76.74% ± 0.83% for the mixed genders with gender classification, and 53.75% ± 1.71% without gender classification with Schweizer and Sklar operator when $q = 0.6$. Significant improvement of up to 23% is obtained for the mixed genders with the gender classification as a preprocessing step.

Experimented on the same DES dataset with the same 90% data of training and 10% data of testing in cross-validation, the best result obtained in the literature by Ververidis and Kotropoulos [17] is 66% for only male samples using a one step GMM (Gaussian Mixture Model) classifier for all the five emotions and 76.15% by Schuller et al. [18]. Significant improvement in classification accuracy is achieved with our automatically generated hierarchical classifier.

4.4. Synthesis and Comparison. In Table 3, we synthesize and compare the performances between the automatically generated hierarchical classification schemes ACSs and the early empirical built hierarchical DECs on both the Berlin and DES datasets. Almost the same results are obtained by the two kinds of hierarchical classifiers for the Berlin dataset (71.52% versus 71.38%), while the empirically built hierarchical DEC classifier for DES dataset performs slightly better than the automatically derived one (81.22% versus 76.74%).

As we can see from the table, the automatically derived ACS offer very closely performance as compared to the empirical DEC while providing the advantage to avoid repeated empiric work when the emotion classification problem changes.

5. Concluding Remarks

In this paper, we have introduced a new embedded feature selection scheme ESFS which is then used as the basis for automatically deriving hierarchical classification schemes called ACS in this paper. Such a hierarchical classifier is represented by a binary tree whose root is the union of all emotion classes, leaves are single-emotion classes, and nodes are subsets containing several emotion classes obtained by a subclassifier. Each of these subclassifiers is based on a new embedded feature selection method, ESFS, which allows to easily represent classifiers characterized by their mass function which is the combination of the information given by an appropriate feature subset, each subclassifier having its own one. Benchmarked on the Berlin and DES datasets, our approach has shown its effectiveness for vocal emotion analysis, leading to closely similar performance as compared to our previous empiric dimensional emotion model-driven hierarchical classification scheme (DEC).

Many issues need to be further studied. For instance, from machine learning point of view, the automatically derived ACS consists of successively dividing the initial set of class labels into two disjoint subsets of class labels by the most optimal binary classifier according to ESFS. Unfortunately, the number of such disjoint subset pairs increases exponentially. When this is feasible with a set of 4 or 6 class labels as it was the case with the Berlin and DES datasets, we cannot do it anymore if the cardinality of class labels is a much bigger number. Therefore, some

heuristic rules also need to be found in order to be able to automatically derive the ACS that we proposed in this paper.

Another issue in machine recognition of vocal emotions is fuzzy and subjective character of vocal emotion. The judgment on emotional state conveyed by an utterance may be between some emotional states or even multiple according to person. Thus, ambiguous or multiple judgments also need to be addressed. A preliminary attempt of this issue has been studied in [40, 41].

References

[1] http://emotion-research.net.

[2] R. Picard, *Affective Computing*, MIT Press, Cambridge, Mass, USA, 1997.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] J. Hirschberg, S. Benus, J. M. Brenier et al., "Distinguishing deceptive from non-deceptive speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1833–1836, September 2005.

[5] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, pp. 1837–1840, September 2005.

[6] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, Geneva, Switzerland, September 2003.

[7] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson, "Children's Emotion Recognition in an Intelligent Tutoring Scenario," in *Proceedings of the 8th European Conference on Speech Communication and Technology (INTERSPEECH '04)*, 2004.

[8] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 474–477, July 2005.

[9] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space?" in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4501–4504, April 2008.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[11] Z. Xiao, E. Dellandrea, L. Chen, and W. Dou, "Recognition of emotions in speech by a hierarchical approach," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, Amsterdam, The Netherlands, September 2009.

[12] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[13] B. Schuller, G. Rigol, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine—belief network architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. I577–I580, May 2004.

[14] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 864–867, July 2005.

[15] D. Morrison and L. C. De Silva, "Voting ensembles for spoken affect classification," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1356–1365, 2007.

[16] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Multi-stage classification of emotional speech motivated by a dimensional emotion model," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 119–145, 2010.

[17] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 1500–1503, July 2005.

[18] B. Schuller, S. Reiter, and G. Rigoll, "Evolutionary feature generation in speech emotion recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 5–8, July 2006.

[19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[20] I. Guyon and A. Elisseff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[21] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," in *Proceedings of International Conference on Machine Learning and Cybernetics*, vol. 4, pp. 2537–2542, 2004.

[22] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif, USA, 1993.

[23] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[24] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967.

[25] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society B*, vol. 30, 1968.

[26] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, USA, 1976.

[27] G. Fioretti, "Evidence theory: a mathematical framework for unpredictable hypotheses," *Metroeconomica*, vol. 55, no. 4, pp. 345–366, 2004.

[28] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[29] M. Detyniecki, *Mathematical aggregation operators and their application to video querying*, Doctoral thesis, University of Paris 6, France, LIP6 research report 2001/002, November 2000.

[30] K. Menger, "Statistical metrics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 8, pp. 535–537, 1942.

[31] B. Schweizer and A. Sklar, "Statistical metric spaces," *Pacific Journal of Mathematics*, vol. 10, pp. 313–334, 1960.

[32] B. Schweizer and A. Sklar, *Probabilistic Metric Spaces*, North Holland, New York, NY, USA, 1983.

[33] R. Fuller, "OWA operators in decision making," in *Exploring the Limits of Support Systems*, C. Carlsson, Ed., vol. 3 of *TUCS General Publications*, pp. 85–104, 1996.

[34] W. Dou, *Segmentation of multispectral images based on information fusion: application for MRI images*, Ph.D. thesis, Université de Caen, 2006.

[35] H. Fu, Z. Xiao, E. Dellandréa, W. Dou, and L. Chen, "Image categorization using ESFS: a new embedded feature selection method based on evidence theory," in *Proceedings of the International Conference on Advanced Concepts Intelligent Vision Systems (ACIVS '09)*, Bordeaux, France, September 2009.

[36] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. I593–I596, Montreal, Canada, May 2004.

[37] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proceedings of 12th European Signal Processing Conference*, pp. 341–344, Austria, September 2004.

[38] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '05)*, pp. 2871–2874, May 2005.

[39] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of Intelligent Information Systems*, vol. 24, no. 2-3, pp. 179–198, 2005.

[40] Z. Xiao, *Recognition of emotion in audio signals*, Ph.D. thesis, Ecole Centrale de Lyon, 2008.

[41] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Ambiguous classification of emotional speech," in *Proceedings of the International Workshop on EMOTION—Satellite of International Conference on Language Resources and Evaluation (LREC '08)*, 2008.