Speech Emotion Classification using Machine Learning Algorithms

S. Casale, A. Russo, G. Scebba
Dipartimento di Ingegneria Informatica e delle Telecomunicazioni
Facoltà di Ingegneria - Università di Catania
Viale Andrea Doria 6, 95125, Catania, Italy
email: scasale@diit.unict.it, arusso@diit.unict.it

S. Serrano
Dipartimento di Fisica della Materia e Tecnologie Fisiche Avanzate
Facoltà di Ingegneria - Università di Messina
Contrada Di Dio (S. Agata), 98166, Messina, Italy
email: sserrano@ingegneria.unime.it

Abstract

The recognition of emotional states is a relatively new technique in the field of Machine Learning. The paper presents the study and the performance results of a system for emotion classification using the architecture of a Distributed Speech Recognition System (DSR). The features used were extracted by the front-end ETSI Aurora eXtended of a mobile terminal in compliance with the ETSI ES 202 211 V1.1.1 standard. On the basis of the time trend of these parameters, over 3800 statistical parameters were extracted to characterize semantic units of varying length (sentences and words). Using the WEKA (Waikato Environment for Knowledge Analysis) software the most significant parameters for the classification of emotional states were selected and the results of various classification techniques were analysed. The results, obtained using both the Berlin Database of Emotional Speech (EMO-DB) and the Speech Under Simulated and Actual Stress (SUSAS) corpus, showed that the best performance is achieved using a Support Vector Machine (SVM) trained with the Sequential Minimal Optimization (SMO) algorithm, after normalizing and discretizing the input statistical parameters.

1. Introduction

The emotional information hidden in speech is an important factor of communication and interaction between humans because while it does not alter the linguistic contents it provides feedback in communication. In spoken communication between humans it is useful to distinguish between two channels [3]: the primary channel, linked to the syntactic-semantic part, conveys linguistic information; the secondary channel conveys paralinguistic information such as tone, emotional state, and gestures. Once recognized and processed, the additional information carried over the secondary channel allows the following functions to be performed: convergence (the interlocutors try to make their way of speaking homogeneous), interaction with the primary channel (making it possible to understand, for example, whether the speaker is joking or asking a rhetorical question), increasing the degree of judgement (to understand, for example, whether the speaker is lying), avoiding misunderstanding (allowing the speaker to emphasize the main parts of a sentence), and giving additional information about the speaker (for example, geographical origin, sex or age).

Various identifiers are used in everyday language to refer to the various emotions. Plutchik and Whissel [3] identify over 120. It is currently difficult to imagine an artificial system capable of reaching such a high degree of discrimination. To overcome this obstacle, emotional states can be characterised in a continuous 2- or 3-dimensional space [8], for example by characterizing on each coordinate the valence (the positive or negative nature of the emotion), degree of activation (quantifying the excitement of the speaker) and dominance (the degree of submissiveness or strength of the speaker). There are various applications for a system capable of recognizing emotional states via speech in a range of fields including psychiatric diagnosis, the toy industry, Customer Relationship Management (CRM), home jukeboxes, Automatic Speech Recognition (ASR), Speech Synthesis, automatic learning, alarms and voicemail systems. A speech emotion recognition system can be considered as

978-0-7695-3279-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICSC.2008.43

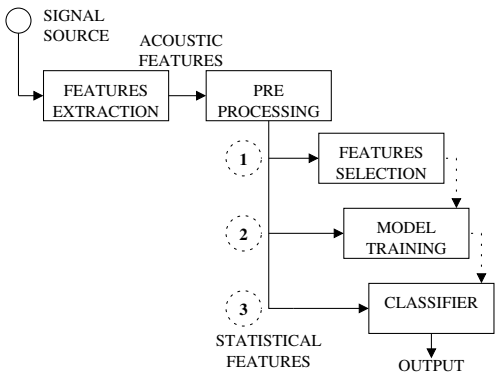


Figure 1. Block diagram of the classification system.

any kind of automatic classification system. Fig. 1 presents a generic block diagram showing the three fundamental blocks for the acquisition of information (specifically the audio signal), extraction and processing of the parameters and classification of the semantic units. There are also three different phases: the first two are used while the system is being set up to identify the statistical features to be used (phase 1) and to train the models to be used during classification (phase 2); phase 3, on the other hand, exploits the statistical features identified and the models trained to classify semantic units whose emotional contents are unknown. In this paper phase 3 was used on sentences with a known emotional content to estimate the performance of the system in terms of correct classification. The front-end for acquisition of the audio signal and extraction of the acoustic parameters was the ETSI ES 202 211 V1.1.1 standard. On the basis of the acoustic parameters, over 3800 statistical parameters were extracted for each semantic unit, as described in Section 2. To realize a classification system it is necessary to have examples of speech that contain different emotional states or states that can be identified as one of the primary emotions considered. These speech corpora were used in both the training and testing phases. More specifically, we used the two different speech corpora described in Section 3. Not all the statistical parameters obtained can be used for classification purposes. We therefore used the parameter selection technique described in Section 4.1. The various classification systems used and the results obtained are described in Section 5.

2 Features Extraction

The features are extracted according to the specifications of the speech recognition front-end algorithm of the ETSI ES 202 211 V1.1.1 standard [1]. The specification covers the computation of feature vectors from speech waveforms sampled at a rate of 16 kHz. The offset-free input signal is divided into overlapping frames of $N = 400$ samples ($25ms$). The frame shift interval (difference between the starting points of consecutive frames) is $M = 160$ samples ($10ms$). The final feature vector extracted for every frame consists of 15 coefficients: the log-energy coefficient, the 12 cepstral coefficients $C_1 - C_{12}$, the pitch period, and the voicing class. The first and second time derivates are calculated for the log-energy coefficient and the 12 MFCCs. The pitch period is also used to calculate the jitter, which is a measure of period-to-period fluctuation in fundamental frequency. Jitter is calculated between consecutive voiced periods via the formula

$$Jitter = \frac{|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}$$

where T_i is the pitch period of the i^{th} window and N is the total number of voiced frames in the segment. In all, excluding the classification of the frame, there are $13 \cdot 3 + 2 = 41$ time series per segment. Prior to subsequent processing all null elements (unvoiced frames for which the front-end is unable to calculate the pitch) are eliminated from the pitch sequence, and the initial and final frames classified as containing silence are eliminated from the MFCC sequence (thus removing any initial and final silence frames from the segment). The value of a single parameter extracted from a frame lasting a few milliseconds is of little significance to determine an emotional state. It is, on the contrary, of interest to investigate the trend taken by the parameter over time. Certain statistics such as the average, minimum and maximum values are extracted from time segments of speech signals. On the basis of the trend followed by each of the features extracted, the following sequences are calculated: local maxima; local minima; distances between local maxima; distances between local minima; distances between local minima and maxima; distances between local maxima and minima; slopes between local minima and maxima; slopes between local maxima and minima; differences between minima and maxima; differences between maxima and minima. For all these sequences the following statistical information is calculated: mean; variance; maximum; minimum; difference between maximum and minimum; 1st quartile; 2nd quartile (median); 3rd quartile; interquartile range (3rd quartile - 1st quartile) [12][10]. Two further statistical features are obtained by calculating the ratio between the number of relative minima and the number of frames and that between the number of relative maxima and the number of frames. According to the classification of the frames, the following sequences are also extracted: length of silence segments; length of unvoiced segments; length of mixed segments; length of voiced segments. For these new

sequences all the statistical information mentioned above is calculated. A final statistical feature calculated is the ratio between the number of transitions between various states and the number of frames in the segment. We therefore had $41 \cdot 10 \cdot 9 + 41 \cdot 2 + 4 \cdot 9 + 1 = 3809$ features for each segment to be classified.

3 Speech corpora

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech recognition. It is evident that research into emotional speech recognition is limited to certain emotions, because the majority of emotional speech data collections encompass 5 or 6 emotions, although there are many more emotion categories in real life. Two speech corpora were used in this research: the first, in German, is called the Berlin Database of Emotional Speech (EMO-DB) [2] and contains semantic units made up of sentences; the second, in English, is called Speech Under Simulated and Actual Stress (SUSAS) [5] and comprises semantic units made up of single words.

3.1 Berlin Database of Emotional Speech

This database comprises 6 basic emotions (anger, boredom, disgust, anxiety, happiness and sadness) as well as neutral speech. Ten professional native German actors (5 female and 5 male) simulated these emotions, producing 10 utterances (5 short and 5 longer sentences), which could be used in everyday communication and are interpretable in all applied emotions. The recorded speech material of about 800 sentences (7 emotions $\cdot$ 10 actors $\cdot$ 10 sentences + some second versions) was evaluated with respect to recognizability and naturalness in a forced-choice automated listening test by 20-30 judges. After selection, the database contained a total of 494 sentences (286 uttered by women and 208 by men).
The sentences were not equally distributed between the various emotional states:55 frightened; 38 disgusted; 64 happy; 79 bored; 78 neutral; 53 sad; 127 angry. As well as being grouped for the classification of the 7 different emotional states *(7EMOTIONS)* the sentences in the database were also grouped in such a way as to make a distinction between the following groups of states:
Activation: (anger, disgust, fear, happiness) - (boredom, sadness) - (neutral);
EmoNoEmo: (anger, boredom, disgust, fear, happiness, sadness) - (neutral);
Evaluation: (anger, boredom, disgust, anxiety, sadness) - (happiness) - (neutral).

3.2 Speech Under Simulated and Actual Stress

The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. The five stress domains include: talking styles (slow, fast, soft, loud, angry, clear, question); single tracking task or speech produced in noise (Lombard effect); dual tracking computer response task; actual subject motion-fear tasks (G-force, Lombard effect, noise, fear); psychiatric analysis data (speech in states of depression, fear, anxiety). The database contains both simulated speech under stress (*Simulated Domain*) and actual speech under stress (*Actual Domain*). A common highly confusable vocabulary set of 35 aircraft communication words makes up the SUSAS database. The words are uttered by 9 male speakers representing the three main USA dialects (General American, Boston, New York). Each style contains 2 recordings of the same word by each speaker. The audio is sampled at 8kHz with a resolution of 16 bits per sample. In this research, as in [13], only 6 of the 11 available states were used: *Angry, Fast, Lombard, Question, Slow and Soft*. Due to the short duration of the recordings which did not allow the front-end to extract the features correctly, 100 ms of silence were added at the start of each recording.

4 Features selection, discretization and transformation

4.1 Features selection

Whereas it would appear, intuitively, that a large number of features would improve the discrimination capabilities of a classification system, in reality various studies have shown that this is not always true. By reducing the size of the classification vector, the system is provided with a more compact and more easily interpretable set of data, the performance of the learning algorithm is improved and the speed of the system increased [12][9][6]. The main feature selection methods are divided into *wrapper* methods and *filter* methods. *Wrapper* methods establish the set of components by interacting with the classification algorithm: they are more accurate but require more computing time. *Filter* methods are independent as they do not require interaction with the classification algorithm and are thus faster. When there is a large amount of training data, filter models are a good choice on account of their computational efficiency and their independence of the learning algorithm. One method used to eliminate redundant and insignificant components is to identify components that are closely correlated with a class but not with each other. Analysis can be in the form of *forward selection*, starting with an empty list and at each step inserting a new attribute until the increase

in performance drops below a pre-established threshold, or by *backward elimination*, starting from a vector containing all the components and eliminating the worst step by step. There are also more complicated search methods, including the *best first* method, which keeps a list of all the subsets of components evaluated, ordered according to performance measures in such a way that a previous configuration can be revisited. *Genetic Algorithms* are a search method based on the principle of natural selection.
In this paper, of the *feature selection* techniques provided by WEKA, we used *CFSSubsetEval*. This algorithm uses as a feature evaluator *Correlation-based Feature Selection*, which tries to identify and discard components that are closely correlated with one another. To determine the best subset we used a *best-first* search strategy and a *stratified 10 cross validation* procedure. We thus had 10 different sets of selected parameters. A parameter *x* may be selected for classification in all 10 subsets, while another parameter *y* may be selected for classification in only 9 out of the 10 subsets, a further parameter *z* may be selected in only 1 of the 10 subsets, and so on. Considering the parameters selected for the various subsets, we created an aggregate of parameters *"1-10"*, which will contain all the parameters selected for classification in at least one of the subsets, the aggregate *"2-10"* which contains all the parameters selected for at least 2 of the subsets, and so on up to the aggregate *"10"* which contains only parameters selected for classification in all of the 10 subsets considered.

4.2 Features discretization

In some cases the continuous domain of the components of the feature vector selected is unsuitable for certain classification algorithms. In order to fully exploit the performance of these algorithms it may be convenient to discretize the continuous domain of the various components. There are unsupervised discretization methods such as *equal-frequency binning*, and supervised methods, such as the algorithms proposed by Fayyad and Irani [4] and Kononenko [7]. In this paper both techniques were used to identify the improvement in performance that can be obtained when classifying emotional states.

4.3 Features normalization

Starting from the original data, new attributes can be obtained so as to present the data in a form that is more appropriate for the learning scheme used. One technique often used for this purpose is data normalization. The most commonly used data normalization techniques are those which try to ensure that all the components fall within a predefined range. In this work we used the min-max and z-score techniques. The former transforms the initial range $[A_{min}, A_{max}]$ into a new range $[\hat{A}_{min}, \hat{A}_{max}]$:

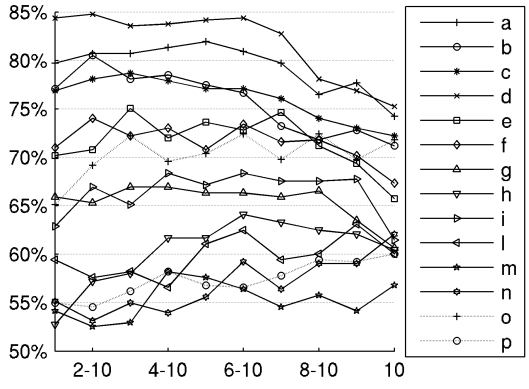
$$\hat{v} = \frac{v - A_{min}}{A_{max} - A_{min}} \left(\hat{A}_{max} - \hat{A}_{min}\right) + \hat{A}_{min}$$

The latter transforms

$$\hat{v} = \frac{v - A_{mean}}{A_{std-dev}}.$$

5 Results

Of the numerous classification algorithms provided by WEKA we preferred initially to use those which did not require excessively long processing times: *NaiveBayes* (standard probabilistic Naïve Bayes classifiers), *NaiveBayesSimple* (simple implementation of Naïve Bayes), *J48* (C4.5 decision tree learner), *RandomForest* (construct random forests), *REPTree*, (fast tree learner that uses reduced-error pruning), *Nnge* (nearest-neighbor method of generating rules using nonnested generalized exemplars), *Part* (obtain rules from partial decision trees using J48), *Ridor* (ripple-down rule learner), *RBFNetwork* (implements a radial basis function network), *SimpleLogistic* (build linear logistic regression models with built-in attribute selection), *SMO* (sequential minimal optimization algorithm for support vector classification), *IB1* (basic nearest-neighbor instance-based learner), *Hyperpipes* (extremely simple, fast learner based on hypervolumes in instance space), *VFI* (voting feature intervals method, simple and fast). We used the default settings for all the algorithms analysed, with the exception of the *IBk* algorithm, for which the value of parameter *k* was set to 2. The first step was therefore to determine which of the techniques considered offers the best performance using the EMO-DB speech corpus. The classification systems were first implemented using all the 3809 features available. In these conditions the best performance was obtained using the SMO algorithm, which yielded an average recognition of about 78%. We then analysed the performance of the various classifiers using the subsets of features selected via the process described in Section 4.1. Fig. 2 illustrates the performance of the various classification systems in terms of the average percentage of recognition of the 7 different emotional states in the database using the various aggregates of parameters. Tests were performed on the same data used during training with "stratified cross validation", more specifically 10-fold crossvalidation. The system was trained on 90% of the database and tested on the remaining 10%. This was repeated 10 times with a different 90/10 percent split each time [11]. From analysis of the results obtained it emerges that the various classification systems do not perform uniformly as the number of parameters used varies. From the 1-10 to the 5-10 aggregates the results are grouped into three categories: the first, which includes



a='bayes.BayesNet' b='bayes.NaiveBayes' c='functions.RBFNetwork'
d='functions.SMO' e='lazy.IB1' f='lazy.IBk' g='misc.HyperPipes'
h='misc.VFI' i='rules.NNge' l='rules.PART' m='rules.Ridor'
n='trees.J48' o='trees.RandomForest' p='trees.REPTree'

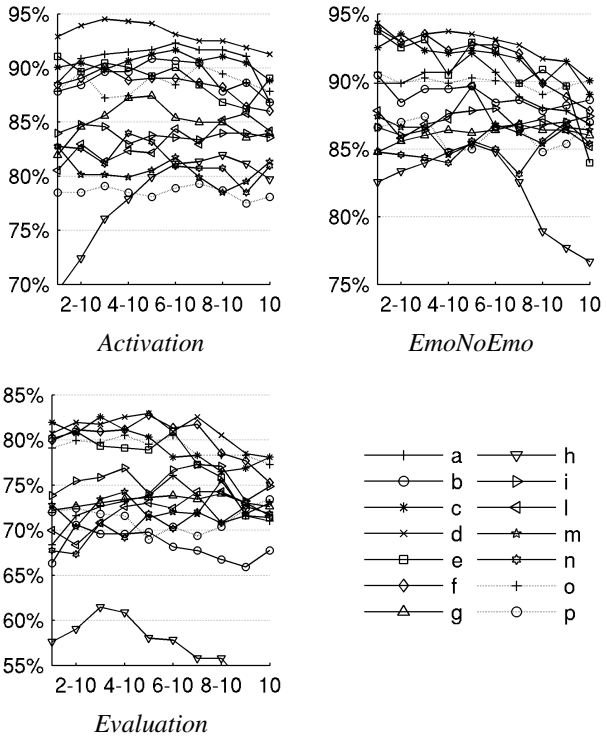
Figure 2. Mean percentage of correct classification of the 7 emotional states in the EMO-DB using the various aggregate of features and the various classification system

the SMO, Simple Logistic, BayesNet, NaiveBayesSimple, NaiveBayes, RBFNetwork classifiers, offers the best performance, with an average percentage of recognition ranging between 75% and 85%; the second, which includes the IB1, HyperPipes, RandomForest, NNge, KStar classifiers, offers intermediate performance levels, with an average recognition percentage ranging between 65% and 75%; the third, which includes Part, J48, REPTree, Ridor, VFI, offers the worst performance, with an average ranging between 52% e il 65%. From the 6-10 to the 10 aggregates the differences between the three groups tend to decrease, due to a deterioration in the performance of the classifiers belonging to the first group and, vice versa, an improvement in the performance of those belonging to the third group. In general, the performance of the classifiers belonging to the second group is in a way independent of the subset of parameters used. Irrespective of the results obtained with the subset 9-10 the SMO technique always gives the best performance. In the case of the *2-10* aggregate, the SMO system achieves an average recognition of almost 85% with an improvement of 7% regarding the case in which all features are used. Table 1 is the matrix of misclassification of emotional states in this case. Fig. 3 shows the performance of the various classification systems in terms of the average recognition percentage for the groups *Activation*, *Emonoemo* and *Evaluation* using varying subsets of parameters. As can be seen, the performance does not appear to depend on the parameter aggregate used. The *SMO* classifier gives the best performance in most of the cases considered. Its performance

Table 1. Misclassification (%) between 7 different emotional states using the *SMO* classifier and the *2-10* aggregate

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**85.04**	0.00	0.00	3.15	11.81	0.00	0.00
2	0.00	**92.41**	2.53	0.00	0.00	0.00	5.06
3	0.00	2.63	**86.84**	7.89	0.00	2.63	0.00
4	9.09	1.82	0.00	**78.18**	7.27	0.00	3.64
5	28.12	0.00	1.56	6.25	**64.06**	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	**98.08**	1.92
7	1.28	7.69	0.00	1.28	0.00	1.28	**88.46**

1 = Anger, 2 = Boredom, 3 = Disgust, 4 = Anxiety
5 = Happiness, 6 = Sadness, 7 = Neutral



a='bayes.BayesNet' b='bayes.NaiveBayes' c='functions.RBFNetwork'
d='functions.SMO' e='lazy.IB1' f='lazy.IBk' g='misc.HyperPipes'
h='misc.VFI' i='rules.NNge' l='rules.PART' m='rules.Ridor'
n='trees.J48' o='trees.RandomForest' p='trees.REPTree'

Figure 3. Mean percentage of correct classification for the groups *Activation*, *EmoNoEmo*, *Evaluation* using the various aggregates of parameters

Table 2. Misclassification (%) in the *Activation* group using the *SMO* classifier and the *3-10* aggregate

	1	2	3
1	**97.89**	1.41	0.70
2	3.05	**91.60**	5.34
3	7.69	5.13	**87.18**

1 = Anger-Disgust-Fear-Happiness
2 = Boredom-Sadness, 3 = Neutral

Table 3. Misclassification (%) in the *Emonoemo* group using the *SMO* classifier and the *1-10* aggregate

	1	2
1	**97.59**	2.41
2	23.08	**76.92**

1 = Anger-Boredom-Disgust-Fear-Happiness-Sadness
2 = Neutral

is far higher than 90% in the classification of the emotional states in the *Activation* and *EmoNoEmo* groups, dropping to values no higher than 85% for the *Evaluation* group.

Table 2 shows the percentage of misclassification between the 3 different groups called *Activation* when the *SMO* classifier is used with the *3-10* aggregate. In this case, in fact, the best average performance is achieved: correct classification in 94.5% of cases. The worst performance is observed in classifying the *Neutral* state. The phrases belonging to this group are, in fact, misclassified in over 7.5% of cases as belonging to the *Anger-Disgust-Fear-Happiness* group and in over 5% as belonging to the *Boredom-Sadness* group.

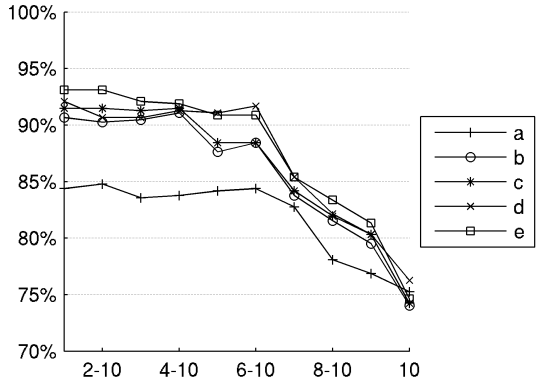
Using the 2 groups called *Emonoemo*, *SMO* again performs best, but with the *1-10* aggregate. The average performance obtained in this case is 94.3%, but there is considerable misclassification of the phrases belonging to the *Neutral* state: in over 20% of cases they are classified as belonging to the *Anger-Boredom-Disgust-Fear-Happiness-Sadness* group as shown in Table 4.

Finally, using the group called *Evaluation*, the best average

Table 4. Misclassification (%) in the *Evaluation* group using the *SMO* classifier and the *5-10* aggregate

	1	2	3
1	**94.87**	3.13	1.99
2	67.19	**31.25**	1.56
3	28.21	0.00	**71.79**

1 = Anger-Boredom-Disgust-Fear-Sadness
2 = Happiness, 3 = Neutral



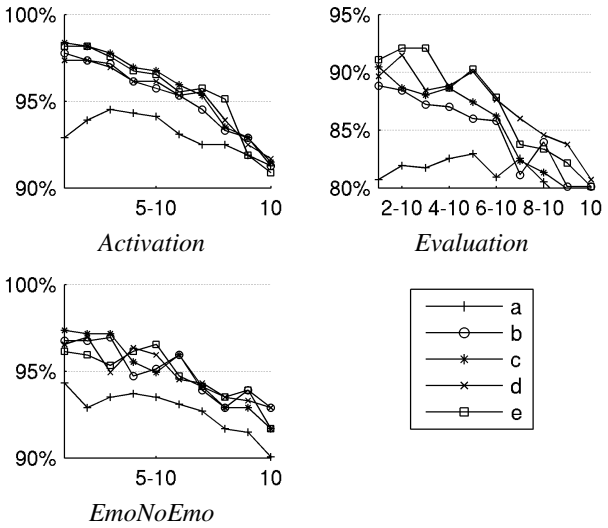
a='no discretization and normalization', b='Fayyad & Irani', c='Fayyad & Irani + normalization', d='Kononenko', e='Kononenko + normalization'

Figure 4. Mean percentage of correct classification of the 7 emotional states in the EMO-DB using the various aggregate of features, the SMO algorithm, discretization and normalization.

performance is obtained by the *SMO* classifier and the *5-10* parameter aggregate. The results given in Table 4 show that *Happiness* is misclassified in over 67% of cases as belonging to the *Anger-Disgust-Fear-Sadness* group. In addition, *Neutral* is often misclassified (in over 28% of cases), again as belonging to the *Anger-Disgust-Fear-Sadness* group.

After this first analysis we tied to improve the performance of the classification system using the feature discretization and normalization techniques outlined in Sections 4.2 and 4.3 respectively. The results obtained are shown in Fig. 4 from which it can be seen that the best performace is obtained using the *Kononenko* discretization algorithm and *z-score* normalization. Using the aggregate of parameters *1-10* the average percentage of correct recognition obtained is about 93%, an improvement of about 8% on the optimal case with no discretization and normalization. Table 5 is the matrix of misclassification of emotional states in this case.

Fig.5 shows the comparison of the mean percentage of correct classification for the different groups of emotional states using the various aggragate of features, the SMO algorithm, discretization and normalization. The best performances for the *Activation* and *EmoNoEmo* groups are obtained using the *Fayyad & Irani* discretization algorithm and *z-score* normalization. Using the aggregate of features *1-10* the mean percentage of correct classification is about 98% and 97% respectively. Instead, the best performance for the *Evaluation* group is obtained using the *Kononenko* discretization algorithm and *z-score* normalization. Using the aggregate of features *3-10* the mean percentage of cor-



a='no discretization and normalization', b='Fayyad & Irani', c='Fayyad & Irani + normalization', d='Kononenko', e='Kononenko + normalization'

Figure 5. Mean percentage of correct classification for the groups *Activation*, *EmoNoEmo*, *Evaluation* using the various aggregates of features, the SMO algorithm, discretization and normalization.

rect classification is about 93%. Tables 6,7, 8 show the misclassification of emotional states in these cases. The goodness of the emotional state classification method was confirmed by using the SUSAS speech corpus. As described in Section 3 the differences as compared with EMO-DB are essentially the language (American English vs. German), the length of the segments to be classified (single words instead of sentences) and the type of emotional states involved. Following the approach in [13], we used only 6 of the 11 states provided by SUSAS. For each of the 9 groups ($Boston_{1,2,3}$, $General_{1,2,3}$, $NewYork_{1,2,3}$) we analysed the performance of the SMO classification algorithm, first using all the 3809 features extracted for each word. Subsequently, by applying the feature selection technique, we built the 10 aggregates of parameters described in Section 4.1. We then proceeded with discretization and normalization using the techniques described in Sections 4.2 and 4.3. Once again the average percentage of recognition went from about 80% when all the features were used to well over 90% when only normalized and discretized features were used. By way of example, Fig. 6 shows the results obtained with the subsets $Boston_1$, $General_1$ and $NewYork_1$.

Table 5. Misclassification (%) between 7 different emotional states using the *SMO* classifier and the *1-10* aggregate with Kononenko discretization and normalization of the features

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	**96.06**	0.00	0.00	1.57	2.37	0.00	0.00
2	0.00	**97.47**	0.00	0.00	0.00	0.00	2.53
3	2.63	0.00	**94.74**	0.00	0.00	2.63	0.00
4	3.63	0.00	1.82	**90.90**	1.82	1.82	0.00
5	18.75	0.00	0.00	0.00	**81.25**	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	**98.08**	1.92
7	1.28	5.12	0.00	1.28	0.00	1.28	**91.02**

1 = Anger, 2 = Boredom, 3 = Disgust, 4 = Anxiety
5 = Happiness, 6 = Sadness, 7 = Neutral

Table 6. Misclassification (%) in the *Activation* group using the *SMO* classifier and the *1-10* aggregate with Fayyad & Irani discretization and normalization of the features

	1	**2**	**3**
1	**98.94**	0.36	0.70
2	0.00	**98.47**	1.53
3	2.56	1.28	**96.16**

1 = Anger-Disgust-Fear-Happiness
2 = Boredom-Sadness, 3 = Neutral

Table 7. Misclassification (%) in the *EmoNoEmo* group using the *SMO* classifier and the *1-10* aggregate with Fayyad & Irani discretization and normalization of the features

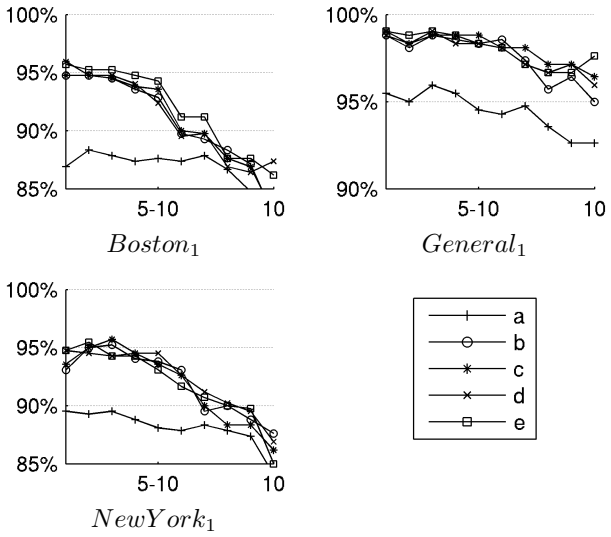
	1	**2**
1	**98.07**	1.93
2	6.41	**93.59**

1 = Anger-Boredom-Disgust-Fear-Happiness-Sadness
2 = Neutral

Table 8. Misclassification (%) in the *Evaluation* group using the *SMO* classifier and the *3-10* aggregate with Kononenko discretization and normalization of the features

	1	**2**	**3**
1	**95.16**	3.14	1.70
2	21.88	**78.12**	0.00
3	10.25	0.00	**89.75**

1 = Anger-Boredom-Disgust-Fear-Sadness
2 = Happiness, 3 = Neutral



a='no discretization and normalization', b='Fayyad & Irani', c='Fayyad & Irani + normalization', d='Kononenko', e='Kononenko + normalization'

Figure 6. Mean percentage of correct classification of the 6 emotional states in the SUSAS DB using the various aggregate of features, the SMO algorithm, discretization and normalization.

6 Conclusions

In this paper we have addressed implementation of a model of an automatic emotional state recognition system capable of working in a DSR environment, using features extracted from an audio signal by the ETSI ES 202 211 v.1.1.1 standard front-end. The experiments were carried out using two different *speech corpora*: EMO-DB, in German, and SUSAS, in American English. A *feature vector* was extracted from each audio segment, containing over 3800 staitstical components calculated on the basis of the time trend of the energy algorithm, 12 cepstral coefficients, the pitch period, jitter and voicing class. Various machine learning techniques made available by WEKA were compared. The SMO algorithm was identified as giving the best performance. Feature selection using a *correlation-based* algorithm, normalization and discretization of the features selected made it possible to enhance performance considerably. The results obtained using EMO-DB yielded over 92% correct classification. In addition, for the nine examples of dialect in SUSAS the system performed extremely well, all accuracy percentages being over 92%, and in some cases ($GENERAL_3$) 100%, meaning that all 420 recordings were correctly classified as corresponding to the 6 emotional states considered.

References

[1] Etsi es 202 211 v1.1.1 (speech processing, transmission and quality aspects (stq); distributed speech recognition; extended front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm). Technical report, European Telecommunications Standards Institute, 2003.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *InterSpeech*, pages 1517–1520, 2005.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, Jan 2001.

[4] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Artificial Intelligence*, pages 1022–1029, Aug-Sept 1993. Chambéry, France.

[5] J. Hansen and S. Bou-Ghazale. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proc. Int'l Conf. Speech Communication and Technology (Eurospeech)*, volume 4, pages 1743–1746. Sept 22-25, 1997, Rhodes, Greece.

[6] H.Lei and V. Govindaraju. Speeding up multi-class svm by pca and feature selection. In *SIAM DM'05*, April 2005. Newport Beach, Cal., USA.

[7] I. Kononenko and M. Robnik-Sikonja. Discretization of continuous attributes using relief. In *ERK '95*. Portoroz, Slovenia, 1995.

[8] K. Kroschel, S. Narayanan, and M. Grimm. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Int'l Conf. on Acoustics, Speech, and Signal Processing*, volume 4, pages 1085–1088, April 2007. Honolulu, Hawaii, USA.

[9] P. Langley and S. Sage. Induction of selective bayesian classifiers. In *Tenth Conf. Uncertainty in Artificial Intelligence*, pages 399–406, 1994. Seattle. W.A.

[10] P. Oudeyer. The production and recognition of emotions in speech: Features and algorithms. *Int'l Journal of Human-Computer Studies*, 59(1-2):157–183, 2003.

[11] R. E. Slyh, W. T. Nelson, and J. H. L. Hansen. Analysis of mrate, shimmer, jitter, and f_0 contour features across stress and speaking style in the susas databases. In *Int'l Conf. on Acoustics, Speech, and Signal Processing*, volume 4, pages 2091–2094, March 1999. Civic Plaza, Hyatt Regency - Phoenix, Arizona, USA.

[12] T. Vogt and E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In IEEE, editor, *Int'l Conf. Multimedia and Expo*, pages 474–477, Jul 2005.

[13] J. T. X. Li, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman. Stress and emotion classification using jitter and shimmer features. In *Int'l Conf. on Acoustics, Speech, and Signal Processing*, volume 4, pages 1081–1084, April 2007. Honolulu, Hawaii, USA.