EXTERNAL VS. INTERNAL SVM-RFE: THE SVM-RFE METHOD REVISITED AND APPLIED TO EMOTION RECOGNITION

Hela Daassi-Gnaba[*], *Yacine Oussar*[†]

Abstract: Support Vector Machines (SVM) are well known as a kernel based method mostly applied to classification. SVM-Recursive Feature Elimination (SVM-RFE) is a variable ranking and selection method dedicated to the design of SVM based classifiers. In this paper, we propose to revisit the SVM-RFE method. We study two implementations of this feature selection method that we call *External SVM-RFE* and *Internal SVM-RFE*, respectively. The two implementations are applied to rank and select acoustic features extracted from speech to design optimized linear SVM classifiers that recognize speaker emotions. To show the efficiency of the *External* and *Internal SVM-RFE* methods, an extensive experimental study is presented. The SVM classifiers were selected using a validation procedure that ensures strict speaker independence. The results are discussed and compared with those achieved when the features are ranked using the Gram-Schmidt procedure. Overall, the results achieve a recognition rate that exceeds 90%.

Key words: *Feature selection, classification, Support Vector Machines (SVM), emotion recognition*

Received: July 1, 2014 **DOI:** 10.14311/NNW.2015.25.004
Revised and accepted: February 12, 2015

1. Introduction

Feature selection is a fundamental issue when dealing with data separation. On one hand, the candidate features must form a large set in order to be enough informative and selective to separate the data. On the other hand, only the most relevant of them have to be involved in the design of the classifier. Irrelevant features can be considered as noisy data and may deteriorate the classification accuracy. Thus, they must be discarded. Usually, the feature selection process consists in determining the subset of variables that achieves the best classification

[*]Hela Daassi-Gnaba, Laboratoire d'Informatique Avancée de Saint-Denis (LIASD, EA 4383), Université Paris 8, 2 rue de la Liberté, 93526 Saint-Denis Cedex, France, E-mail: hela.daassi@univ-paris8.fr

[†]Yacine Oussar – Corresponding author, Laboratoire de Physique et d'Étude des Matériaux (LPEM), PSL Research University, ESPCI-ParisTech, Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 8213, 10 rue Vauquelin, 75231 Paris Cedex, France, E-mail: yacine.oussar@espci.fr

performance. This process may start by ranking the available variables according to their level of relevance and then selects the most relevant of them. However, ranking and selection may come combined in a unique procedure. Frequently, the selection methods are run previous to the classifier design. They do not take into account the classifier structure or the training algorithm. SVM-RFE [14] is simultaneously a ranking and a selection method. It is dedicated to the design of linear SVM classifiers [5]. By means of (i) a built-in regularization mechanism and (ii) the linearity in their parameters, SVM have shown a great ability for the design of classifiers with strong generalization capabilities.

In the present paper, we propose to revisit the SVM-RFE method. Since the ranking and selection performed by SVM-RFE require to train SVM classifiers, the regularization hyperparameter commonly pointed by C has to be optimized. Depending on how the optimization of this parameter interacts with the feature ranking, two different schemes can be derived. We will call these schemes and their respective implementations as *External SVM-RFE* and *Internal SVM-RFE*.

To illustrate the performance of the two implementations we propose, we are interested in an application that comes from the speech processing world [20]. Indeed, speech processing is a rich and a wide research domain with many applications in various areas. Beside speech recognition which has been a stimulating research activity for more than half a century, computational paralinguistics [26], which includes emotion recognition from speech [17, 28], has become of a great interest during the last few years. This research domain implements intensively methods that belong to the machine learning world [6]. Speech emotion recognition is in itself a challenging research field [22–25]. Its main objective is often an improvement of the speech recognition by making the human-machine interaction more natural [12]. Speech emotion recognition can be useful for applications which require an extraction of emotions from speech to generate corresponding facial expressions. This process is involved in various domains as a key factor to make the human-machine interaction friendlier and more efficient [10, 11].

One important challenge of the speech emotion recognition is the extraction of relevant features that are efficiently informative on the existing emotions. In the literature, different types of speech features are used [2, 12, 29], we focus on basic acoustic features since (i) we assume that the acoustic information is fully given by a set of both prosodic and spectral features and (ii) the use of these acoustic features is still an active research domain in emotion recognition [30]. We propose to study four separation problems: “Happy” versus “Neutral”, “Cold anger” versus “Neutral”, “Hot anger” versus “Neutral” and “Panic” versus “Neutral”.

The aim of this paper is twofold: (i) a detailed description of a dual implementation of the SVM-RFE method, (ii) an application of the proposed methods for an efficient pairwise emotion separation with optimized classifiers. The classifiers involve the most relevant features and determine the emotion from the speech signal. To compare the efficiency of the two implementations we propose, the results we obtained are discussed and compared with those achieved when the features are ranked using the Gram-Schmidt method [7] and selected according to a wrapper approach [13].

The paper is organized as follows: SVM classification is briefly reported in Sec. 2. The ranking and selection methods we propose are described in detail in Sec. 3. Sec. 4 presents: (i) the corpus used for this study, (ii) the results obtained by numerical experiments for speech emotion recognition.

2. SVM classification: a brief recall

SVM classification is used to find an optimal separation between two classes: the maximum margin hyperplane [9]. The set of examples that are sufficient to determine the maximum margin hyperplane are called the support vectors. If the data are linearly separable, a linear SVM classifier is sufficient. Otherwise, if the data are not linearly separable, SVM classification proceeds by projecting the input vectors in a high dimensional space called the feature space then a linear separation is possible. In practice, this data conversion leads to the use of a kernel function. To be a SVM kernel, a function has to verify a set of conditions listed in [9]. An SVM discriminant function is given by

$$f(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b, \tag{1}$$

where k is the kernel function, $\mathbf{x}_i$ are the support vectors, y_i are the corresponding class labels (± 1) and M is the number of support vectors. Note that α_i and b are the parameters of the classifier adjusted during the training process that leads to maximizing:

$$L(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j),$$

subject to

$$\sum_{i=1}^{M} \alpha_i y_i = 0,$$

and

$$0 \leq \alpha_i \leq C,$$

for $1 \leq i \leq M$.

The regularization parameter C controls the trade off between classification errors on training data and margin maximization.

The validation procedure used in our experiments is a particular implementation of the cross validation method [15]. It is called the Leave-One-Speaker-Out (LOSO) method. Since the data used in this study involves six speakers, this procedure consists in a data partitioning in which the validation fold contains data from the sixth speaker that does not appear in the training folds containing data from the five other speakers. This method guarantees strict speaker independence.

3. Variable ranking and selection methods

3.1 The Gram-Schmidt orthogonalization procedure

The Gram-Schmidt orthogonalization (GS) procedure is a method for ranking the variables of linear-in-their-parameters models according to their level of relevance. In the context of machine learning, this method was first introduced in [7]. Afterwards, it has been widely implemented for various purposes. The GS procedure is an iterative method. For ranking a set of $N_{\max}$ variables, it proceeds, at the first iteration, by estimating the relevance of each variable by computing the following quantities:

$$\cos^2(\mathbf{x}_k, \mathbf{y}) = \frac{(\mathbf{x}_k.\mathbf{y})^2}{\|\mathbf{x}_k\|^2 \|\mathbf{y}\|^2}, \quad k = 1, \ldots, N_{\max},$$

where $\mathbf{x}_k$ is the vector of the values of the k-th variable and the components of the vector $\mathbf{y}$ takes values (± 1) since we are dealing with a classification problem.

The most relevant variable is the input vector that leads to the largest value of this quantity. Projecting the remaining $N_{\max} - 1$ variables and the output vector $\mathbf{y}$ to the subspace orthogonal to the vector of the most relevant variable ends the first iteration. Indeed, this projection permits to avoid the selection of redundant variables. The second iteration proceeds similarly by computing the relevance of the $N_{\max} - 1$ variables, selecting the most relevant one and projecting the remaining $N_{\max} - 2$ variables. The procedure stops when all the variables are ranked. Once the features are ranked, the N most relevant of them can be selected using either a filter or a wrapper approach [13]. Although the filter approach is often computationally cost effective, we focused on the wrapper approach which usually leads to a better generalization in the data separation with an acceptable computational burden in our implementation. Fig. 1 describes a basic example with two variables $\mathbf{x}_1$ and $\mathbf{x}_2$.

In our experiments, both hyperparameter C and the value of N were optimized according to a 6-fold LOSO validation procedure described in the previous Section. For each value of N, the hyperparameter C is optimized following a list search.

We describe below in detail the feature selection procedure according to a wrapper approach when the GS procedure is used with linear SVM.

Step 1 F: f_1, f_2, ... $f_{N_{\max}}$ is the set of ranked features
using the Gram-Schmidt orthogonalization procedure
Step 2 Define a list of n values for C: C_1, C_2, ... C_n
Step 3 For all i from 1 to $N_{\max}$
Consider the f_1 to f_i ranked features
For all j from 1 to n
Set C to C_j
Compute the score of the LOSO procedure
End of loop on j
Save the best LOSO score and the corresponding C_j
End of loop on i
Select the subset of the N most relevant features
that achieves the best LOSO score

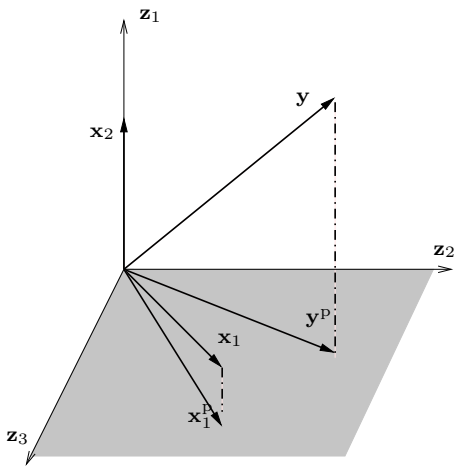

Fig. 1 *A basic illustration of ranking variables.* $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$ *is a set of 3 linearly independent vectors: they form a 3-dimensional vector space. At the first step,* $\mathbf{x}_2$ *is the most relevant variable with respect to the output* $\mathbf{y}$*. It is selected and then* $\mathbf{y}$ *and variable* $\mathbf{x}_1$ *are projected on the subspace orthogonal to* $\mathbf{x}_2$*. If there are more projected variables* $\mathbf{x}_i^{\mathrm{p}}$*, their relevance is computed with respect to* $\mathbf{y}^{\mathrm{p}}$*.*

3.2 The SVM-RFE methods

As discussed below in Sec. 4, data are linearly separable. Thereby, linear SVM are preferred for the design of classifiers. The SVM-RFE is a ranking method dedicated for the design of linear SVM classifiers [13]. This method evaluates the relevance of a variable through the change that occurs in the training cost function when this variable is withdrawn from the classifier inputs. More precisely, withdrawing the i-th variable from the set of features results in a change $\Delta J(i)$ in the training cost function J as discussed in [14]. The expression of J is given by

$$J = \sum_{i=1}^{P} e_i^2 + C\|\mathbf{w}\|^2,$$

where P is the number of examples, $e_i = (y_i - \hat{y}_i)$ is the error for example i and $\mathbf{w}$ is the vector of classifier adjustable parameters. The expression of the output of the classifier is given by

$$\hat{y}_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b,$$

where b is an adjustable parameter. Thereby, J is quadratic with respect to the parameters w_i. Accordingly, the change $\Delta J(i)$ is proportional to the square of w_i,

$$\Delta J(i) \propto w_i^2.$$

Thus, the relevance of a variable is given by the magnitude of the square of the corresponding parameter w_i. The smaller is the magnitude of w_i^2, the less relevant is the variable. The SVM-RFE method starts with considering all the

available features. A first training allows to determine the less relevant variable which is withdrawn from the set of variables. Then, the method proceeds iteratively performing a training at each iteration. The last remaining variable is the most relevant. Note that if the output is given by the expression of Eq. (1), the parameters w_i are given by $\mathbf{w} = \sum_{k=1}^{M} \alpha_k y_k \mathbf{x}_k$.

Since the SVM-RFE method necessitates to train linear SVM classifiers, a strategy to optimize the hyperparameter C must be applied. Indeed, the ranking will depend on the value of C. One might ask: (i) how this parameter influences the ranking and the classifier overall performance, (ii) how to set it effectively. To provide an answer to these questions, we propose two formulations of this ranking method that differ on the optimization scheme of C. We name them the *External SVM-RFE* and the *Internal SVM-RFE* methods.

- *External SVM-RFE*: This first formulation consists in selecting a value for hyperparameter C in a predetermined set and keeping it fixed until all the variables are ranked. The complete procedure is described below.

Step 1 Define a set of candidate values for C: C_1, C_2, ... C_n
Step 2 For all j from 1 to n
  Set C to C_j
  For all i from 1 to $N_{\max}$
    Consider the $N_{\max} - i + 1$ remaining variables
    Compute the score of the LOSO procedure
    Compute vector $\mathbf{w}$
    Withdraw the less relevant variable
  End of loop on i
  Save the subset of ranked variables that achieves
  the best LOSO score for C_j
End of loop on j
Among all the saved subsets, select the subset
that achieves the best LOSO score

Determining the subset that achieves the best LOSO score allows to select the N most relevant variables. Thus, this ranking procedure possesses a built-in selection mechanism. This method produces as many rankings as the number of values tested for hyperparamter C.

- *Internal SVM-RFE*: This second formulation consists in optimizing the value for hyperparameter C for every variable to withdraw. The complete procedure is described below.

Step 1 Define a set of candidate values for C: C_1, C_2, ... C_n
Step 2 For all i from 1 to $N_{\max}$
  Consider the $N_{\max} - i + 1$ remaining variables
  For all j from 1 to n
    Set C to C_j
    Compute the score of the LOSO procedure
  End of loop on j

Save the subset of ranked variables that achieves
the best LOSO score
and the corresponding C_j
Compute vector $\mathbf{w}$ with C_j
Withdraw the less relevant variable
End of loop on i
Among all the saved subsets, select the subset
that achieves the best LOSO score

Similarly to the first formulation, the *Internal SVM-RFE* method also possesses a built-in selection mechanism. Indeed, the subset that achieves the best LOSO score is formed by the N most relevant variables among all those ranked. Contrary to the *External SVM-RFE*, this formulation produces a unique ranking.

3.3 Discussion

Both *External* and *Internal SVM-RFE* methods are computationally more complex than the Gram-Schmidt procedure. However, simultaneously to the ranking, they perform a selection procedure according to a built-in selection mechanism.

The two *SVM-RFE* methods are sensitive to the tested values of parameter C. Practically, these values are taken in an arbitrary list of numbers generated according to a logarithmic scale.

With the *External SVM-RFE* method, a different ranking is achieved with each value of C. Thus, if the logarithmic scale is not fine enough, the feature subset with the best validation score may be a little different of the subset achieved by the optimal value of C omitted in the list.

With the *Internal SVM-RFE* method, the situation is quite different. The ranking depends on the values of C considered as optimal at each step. Thus, a small deviation is introduced in the ranking at each step, i.e. $N_{\max}$ times on the whole. To overcome this drawback, one can involve a larger list of values for parameter C. The price to pay is the increase in the computational burden. Thus, this issue makes the *Internal SVM-RFE* method numerically more expensive than the *External SVM-RFE*.

When implemented with real data, one can wonder if either of these two methods is more efficient. We propose to study the performances of both methods for ranking and selecting acoustic features to design optimized classifiers for emotion recognition.

4. Experimental results

4.1 Corpus and feature extraction

The data used in our experiments was obtained from the LDC Emotional Prosody and Transcripts database [8]. It consists of English language acted speech recordings. This database contains both audio recordings and the corresponding transcripts. The recordings deal with professional actors reading series of semantically

neutral utterances (dates and numbers) spanning fifteen distinct emotional categories. In this study, we chose six professional actors: three male (CC, CL and MF) and three female (GG, JG and MK) to guarantee speaker gender balancing. From the LDC Emotional Prosody Data, we focused on the recognition of a first emotion separation problem: "Happy" versus "Neutral" emotions, then we also paid attention to a second emotion separation problem: "Cold anger" versus "Neutral" emotions. We study further emotion separation problems in Sec. 4.4.

Pre-processing the database leads to the following outcome:

- For the first separation problem: 53 utterances for female speakers (32 Happy and 21 Neutral utterances) and 52 for male (31 Happy and 21 Neutral utterances).
- For the second separation problem: 53 utterances for female speakers (32 Cold anger and 21 Neutral utterances) and 53 for male (32 Cold anger and 21 Neutral utterances).

Many acoustic features for speech emotion recognition have been explored. However, there is still no consensus on a fixed set of features. In [21], Schuller et al. have used 276 acoustic features. In the Interspeech challenges 2009 and 2010, Schuller et al. have increased the number of acoustic features to 384 in [22] and 1582 in [23]. Later, in Audio/Visual Emotion Challenges (AVEC) in 2011 and 2012, Schuller et al. have slightly decreased the number of acoustic features from 1941 in [24] to 1841 in [25].

In our work, to reduce the constraint of computational complexity, we propose to extract basic acoustic features. In order to take into account relevant and informative input data, we considered both prosodic and spectral features. More specifically, we propose to use statistical moments from two prosodic features and four spectral features. The statistical moments are: maximum, minimum, range (maximum-minimum), mean, median and standard deviation. The selected prosodic features are the fundamental frequency contour (F_0) and the energy contour (E_n). The selected spectral features are the formant frequencies (F_1, F_2) and their bandwidths (B_1, B_2). We gathered the 6 moments from each of the 6 basic acoustic features to generate a set of $N_{\max} = 36$ variables. Theses features can be computed quickly which is convenient for real time applications. All these features were extracted with the Praat software [4] and their statistical moments were computed using the Matlab software. According to the corpus described above, 105 utterances are given to separate the "Happy" versus "Neutral" case and 106 utterances are available for the "Cold anger" versus "Neutral" case. We assume that in both cases, each utterance is fully described given the values of the 36 features.

4.2 Happy/Neutral emotion separation

Prior to the design of SVM classifiers, we were interested in testing the data linear separability. Indeed, linear separability is a desirable property when separating data since the resulting classifiers are less complicated to built than nonlinear classifiers. For this purpose, the Ho-Kashyap algorithm [16] was run to determine if the data are linearly separable regarding the training examples.

Rank	Feature
1	range F_0
2	max F_0
3	mean F_0
4	median B_2
5	median F_2
6	min F_0
7	std F_1
8	max F_1
9	max B_1
10	range B_2
11	mean F_1
12	std F_2
13	median B_1
14	std B_1
15	min F_1

Rank	Feature
1	median F_0
2	min F_0
3	std F_1
4	mean F_2
5	range F_0
6	median B_2
7	max F_0
8	range F_1
9	mean F_1
10	median F_1
11	std F_2
12	median E_n
13	max B_1

Tab. I *The N most relevant acoustic features that ensure linear separability according to the Gram-Schmidt (left) and the External SVM-RFE (right) methods with the Happy/Neutral data.*

Methods	N	C
Gram-Schmidt	15	X
External SVM-RFE	13	8

Tab. II *Conditions of linear separability for the Happy/Neutral data.*

When taking into account the whole available examples (105 utterances), the Ho-Kashyap algorithm showed that the data are linearly separable when the $N = 15$ most relevant input variables, according to the Gram-Schmidt method, over the 36 available are considered. Tab. I (left) illustrates the list of these 15 most relevant features. When using the *External SVM-RFE* method, features ranking and classifiers design are strongly linked. Therefore, for each value of C, the Ho-Kashyap algorithm is run after the classifier design: the smaller subset of relevant features that allows data linear separability is $N = 13$ corresponding to $C = 8$. Tab. I (right) illustrates the list of these 13 most relevant features. Tab. I shows that 8 features are common to these two methods. As explained in Sec. 3.3 a full implementation of the *Internal SVM-RFE* method comes with a heavy computational burden. We then skipped the study of the linear separability for this case. The results are summarized by Tab. II. Note that linear SVM were implemented using the spider software [27].

As shown by Tab. I, with the Gram-Schmidt (left) and the *External SVM-RFE* (right) methods, the most relevant features are moments related to the pitch frequency F_0. This result is consistent with the property of the pitch frequency which is known as a leading parameter in speech emotion recognition [1, 18].

Methods	Well classified [%]	N	C	M
Gram-Schmidt + linear classifier	83.8	15	X	X
Gram-Schmidt + linear SVM	89.5	15	20	19
External SVM-RFE	**94.3**	**16**	**7**	**20**
Internal SVM-RFE	91.4	11	20	20

Tab. III *Percentage of well classified emotions for the Happy/Neutral data.*

When the variables are ranked using the Gram-Schmidt method, the best recognition rate of 89.5% is achieved by involving the $N = 15$ most relevant variables and $C = 20$. As a comparison, 83.8% of the emotions were successfully recognized with a linear classifier using the same $N = 15$ most relevant features. Note that the linear classifier consists of an hyperplane whose parameters are adjusted by minimizing the least squares cost function.

The best performance was obtained with the *External SVM-RFE* having $N = 16$ and $C = 7$. With this classifier involving $M = 20$ support vectors (roughly 19% of the whole available data), 94.3% of the emotions were successfully recognized. The overall results are given by Tab. III.

As mentioned above, such linear SVM with the 16 most relevant features as inputs separates successfully any subset from the data if it is trained with all the available examples. Since we implemented the LOSO validation procedure, the behavior of the validation speaker can be slightly different from the others. Hence, the hyperplane determined during training may not successfully separate all the examples that belong to the validation speaker. As a result, the overall recognition rate is not guaranteed to be close to 100%.

Tab. IV and Tab. V summarize the results of the classification of "Happy" versus "Neutral" emotions using respectively the Gram-Schmidt and *External SVM-RFE* methods for each speaker (first column); we show the percentage of well classified items (second column), the precision (third column) and the recall for the recognition of Happy and Neutral emotions (fourth column). Note that for a given class, the precision is the fraction of well classified emotions over all those put in this class. The recall is the fraction of emotions put in this class over all those labeled in this class. Precision and recall can be computed using the confusion matrices.

These results show that a promising rate in emotion recognition can be obtained with few relevant acoustic features. This classification confirms the validity of our assumption as well as the feasibility of a numerically cost effective implementation. Indeed, the classifier is linear and uses a small set of input variables instead of the overall 36 extracted from the speech signal.

For the sake of the design of parsimonious classifiers involving less input variables ($N < 15$), nonlinear SVM classifiers using a Gaussian kernel were also implemented. The regularization parameter C, the Gaussian kernel parameter σ, as well as the value of N were simultaneously optimized according to the 6-fold LOSO validation procedure described above. The results showed that neither the classification error nor the number of relevant input variables was optimized. Hence, we consider that the nonlinear classification does not bring any improvement for the "Happy" versus "Neutral" emotions separation.

Speaker		Well classified [%]	Precision Happy	Precision Neutral	Recall Happy	Recall Neutral
Male	CC	77.8	0.89	0.67	0.73	0.86
	CL	100	1	1	1	1
	MF	100	1	1	1	1
Female	GG	88.9	0.91	0.86	0.91	0.86
	JG	83.3	0.83	0.83	0.91	0.71
	MK	88.2	1	0.78	0.8	1

Tab. IV *Percentage of well classified emotions, precision and recall obtained when the acoustic features are ranked using the Gram-Schmidt method (Happy/Neutral data).*

Speaker		Well classified [%]	Precision Happy	Precision Neutral	Recall Happy	Recall Neutral
Male	CC	100	1	1	1	1
	CL	94.1	0.91	1	1	0.86
	MF	100	1	1	1	1
Female	GG	100	1	1	1	1
	JG	83.3	0.78	1	1	0.57
	MK	88.2	1	0.78	0.8	1

Tab. V *Percentage of well classified emotions, precision and recall obtained when the acoustic features are ranked using the External SVM-RFE method (Happy/Neutral data).*

4.3 Cold anger/Neutral emotion separation

Similarly to the previous classification problem, the whole available acoustic features were ranked taking into account the Cold anger/Neutral emotion labeling. Data separability was tested using the Ho-Kashyap algorithm [16]. The latter showed that the whole available data (106 utterances) are linearly separable when considering at least the 17 most relevant features according to the Gram-Schmidt method (Tab. VI (left)) or the 18 most relevant features according to the *External SVM-RFE* method with $C = 4$ (Tab. VI (right)). Tab. VI also shows that 11 features are common to theses two methods. Tab. VII summarizes these results.

As shown by Tab. VI, with the *External SVM-RFE* method (right), the most relevant features are moments related to the pitch frequency F_0. According to the Gram-Schmidt method (left), the most relevant feature is the signal energy. Nevertheless, 4 moments related to the pitch frequency show up in the list. To a lesser extent than the previous results, the pitch frequency stands as a highly informative feature for speech emotion recognition [1, 18].

The classification results for the "Cold anger" versus "Neutral" problem are given by Tab. VIII. When the variables are ranked using the Gram-Schmidt

Rank	Feature
1	range E_n
2	max F_2
3	min B_2
4	min F_0
5	median F_0
6	min F_1
7	median F_2
8	mean B_1
9	range F_0
10	max F_0
11	std F_1
12	std F_2
13	range B_2
14	median B_2
15	median E_n
16	std E_n
17	std B_1

Rank	Feature
1	median F_0
2	min F_0
3	range F_0
4	max F_0
5	std F_1
6	median F_2
7	median B_1
8	median E_n
9	range E_n
10	mean F_1
11	median F_1
12	std E_n
13	max F_2
14	min F_1
15	std F_0
16	mean F_2
17	mean B_2
18	max E_n

Tab. VI *The N most relevant acoustic features that ensure linear separability according to the Gram-Schmidt (left) and the External SVM-RFE (right) methods with the Cold anger/Neutral data.*

Methods	N	C
Gram-Schmidt	17	X
External SVM-RFE	18	4

Tab. VII *Conditions of linear separability for the Cold anger/Neutral data.*

Methods	Well classified [%]	N	C	M
Gram-Schmidt + linear classifier	77.4	12	X	X
Gram-Schmidt + linear SVM	86.8	12	10	27
External SVM-RFE	87.7	21	2	33
Internal SVM-RFE	**90.6**	**10**	**1000**	**27**

Tab. VIII *Percentage of well classified emotions for the Cold anger/Neutral data.*

method, the best recognition rate of 86.8% is achieved by involving the $N = 12$ most relevant variables and $C = 10$. As a comparison, 77.4% of the emotions were successfully recognized with a linear classifier using the same $N = 12$ most relevant features. The best validation score consisting in 90.6% of well classified emotions was obtained with the *Internal SVM-RFE* having $N = 10$ and $C = 1000$

Speaker		Well classified [%]	Precision		Recall	
			Cold anger	Neutral	Cold anger	Neutral
Male	CC	94.1	1	0.87	0.9	1
	CL	94.4	1	0.87	0.91	1
	MF	94.4	1	0.87	0.91	1
Female	GG	82.4	0.89	0.75	0.8	0.86
	JG	55.6	0.59	0	0.91	0
	MK	100	1	1	1	1

Tab. IX *Percentage of well classified emotions, precision and recall obtained when the acoustic features are ranked using the Gram-Schmidt method (Cold anger/Neutral data).*

Speaker		Well classified [%]	Precision		Recall	
			Cold anger	Neutral	Cold anger	Neutral
Male	CC	88.2	0.9	0.86	0.9	0.86
	CL	88.9	0.85	1	1	0.71
	MF	94.4	1	0.87	0.91	1
Female	GG	88.2	0.9	0.86	0.9	0.86
	JG	88.9	0.85	1	1	0.71
	MK	94.4	0.92	1	1	0.86

Tab. X *Percentage of well classified emotions, precision and recall obtained when the acoustic features are ranked using the Internal SVM-RFE method (Cold anger/Neutral data).*

(see Tab. VIII). It involves $M = 27$ support vectors (roughly 25% of the whole available data).

Tab. IX and Tab. X illustrate the classification results for the "Cold anger" versus "Neutral" case using respectively the Gram-Schmidt and *Internal SVM-RFE* methods for each speaker (first column); it gives the correct classification rate (second column), the precision (third column) and the recall for the recognition of Cold anger and Neutral emotions (fourth column). Precision and recall can be computed using the confusion matrices.

A comparison of the results illustrated by both Tab. IX and Tab. X shows that the *Internal SVM-RFE* method leads to a ranking that better separates the female speakers. Indeed, the emotions of speaker JG have a recognition rate of 88.9% while using the Gram-Schmidt method achieves only 55.6% of well classified emotions. Although the latter performs better for male speakers. The overall performance goes to the *Internal SVM-RFE* method.

To be consistent with the previsous classification problem, nonlinear SVM classifiers using a Gaussian kernel were also implemented. The regularization parameter C, the Gaussian kernel parameter σ, as well as the value of N were optimized

according to the 6-fold LOSO validation procedure described above. Similarly, the performance was not improved. Also for this case, linear classification remains more suitable.

4.4 Further emotion separation problems

In order to validate the results obtained above, we propose to study two more cases of emotion separation. We still assume that for both cases, each utterance is fully described given the values of the 36 features outlined in Sec. 4.1.

4.4.1 Hot anger/Neutral emotion separation

The third problem we introduce consists in 105 utterances: 52 for female speakers (31 Hot anger and 21 Neutral utterances) and 53 for male (32 Hot anger and 21 Neutral utterances). Tab. XI shows the percentage of well classified emotions achieved by the two proposed methods and compared with the score obtained when the features are ranked using the Gram-Schmidt method. In this case, the *External SVM-RFE* method achieves the best separation performance.

Methods	Well classified [%]	N	C
Gram-Schmidt + linear SVM	90.5	7	20
External SVM-RFE	**93.3**	**8**	**1**
Internal SVM-RFE	87.6	11	0.02

Tab. XI *Percentage of well classified emotions for the Hot anger/Neutral data.*

4.4.2 Panic/Neutral emotion separation

The fourth problem we introduce consists in 103 utterances: 51 for female speakers (30 Panic and 21 Neutral utterances) and 52 for male (31 Panic and 21 Neutral utterances). Tab. XII shows the percentage of well classified emotions achieved by the two proposed methods and compared with the score obtained when the features are ranked using the Gram-Schmidt method. In this case, the *Internal SVM-RFE* method achieves the best separation performance.

Methods	Well classified [%]	N	C
Gram-Schmidt + linear SVM	92.2	6	20
External SVM-RFE	93.2	12	1
Internal SVM-RFE	**94.2**	**11**	**50**

Tab. XII *Percentage of well classified emotions for the Panic/Neutral data.*

4.5 Discussion

In [21], Schuller et al. presented results obtained with the Berlin Emotional Speech Database [3]. They considered seven different emotions and 276 acoustic features.

They implemented SVM classifiers with 75 relevant variables selected using the Sequential Floating Forward Search (SFFS) method [19]. They obtained a recognition rate of 87.5%.

In our work, we present results obtained with the English LDC Emotional Prosody and Transcripts Database [8]. We considered five different emotions. To reduce the computational burden and to design optimized classifiers, we proposed two different implementations of a feature ranking method dedicated to SVM classifiers. We extracted 36 basic features. On the "Happy" versus "Neutral" problem, a recognition rate of 94.3% was obtained with a subset formed by the 16 most relevant features. On the "Panic" versus "Neutral" problem, a recognition rate of 94.2% was achieved with a subset formed by the 11 most relevant features among the 36 extracted. Tab. XIII summarizes the results obtained with the four separation problems we considered.

Separation problem	*External SVM-RFE* [%]	*Internal SVM-RFE* [%]
"Happy" vs. "Neutral"	**94.3 (N=16)**	91.4
"Cold anger" vs. "Neutral"	87.7	**90.6 (N=10)**
"Hot anger" vs. "Neutral"	**93.3 (N=8)**	87.6
"Panic" vs. "Neutral"	93.2	**94.2 (N=11)**

Tab. XIII *Percentage of well classified emotions for all the studied problems.*

Our results are slightly better than those from Schuller et al. noticed above. In fact, they present improvements by the recognition rate as well as by the smaller size of the relevant feature subsets. Since the implementation conditions are quite different, the approach we propose can be considered as complementary to the other existing methods and does not claim to systematically outperform them whatever the data and the number of different labels.

We recall that the aim of our study is to propose and compare two different implementations of the SVM-RFE method for the design of optimized linear SVM classifiers. The results we obtain are promising since they show the effectiveness of our approach: the two methods perform differently for a given set of data. Thereby, they must be taken into account both.

5. Conclusion

When dealing with the design of classifiers from data, the feature ranking and selection is a key step to achieve promising recognition rates. Within this framework, the presented study proposes two implementations of the SVM-RFE method entitled *External SVM-RFE* and *Internal SVM-RFE*, respectively. Theses two methods determine two different optimization schemes for the regularization hyperparameter C. For *External SVM-RFE*, the complete ranking and selection scheme is run for each value of C taken from a predetermined set of candidate values. The *Internal SVM-RFE* proceeds by optimizing the hyperparameter C at each iteration of the ranking and selection scheme.

We selected emotion recognition as an interesting application to illustrate the efficiency of both methods and also their differences to design optimized classifiers. The study of data separability shows that considering a small set of relevant features is sufficient to ensure data linear separability. Several numerical experiments involving linear SVM classifiers were conducted. Four different separation problems were considered. Promising recognition rates over 90% were achieved. Furthermore, a small number of support vectors for each of the optimized classifiers confirms their good generalization capabilities.

The results show that the best performance is obtained with either method according to the separation problem. For the "Happy" versus "Neutral" and "Hot anger" versus "Neutral" problems, the best performance was achieved by the *External SVM-RFE* method while for the "Cold anger" versus "Neutral" and "Panic" versus "Neutral" emotions separation the best recognition rate was obtained with the *Internal SVM-RFE* method. Thus, according to the available data either method may be more efficient. Therefore, for a given study, both methods must be considered, implemented and their performances compared.

References

[1] BÄNZIGER T., SCHERER K.R. The role of intonation in emotional expression. *Speech Communication*. 2005, 46(3-4), pp. 252-267, doi: 10.1016/j.specom.2005.02.016.

[2] BÁRTŮ M. Speech disorder analysis using Matching Pursuit and Kohonen Self-Organizing Maps. *Neural Network World*. 2012, 22(6), pp. 519-533, doi: 10.14311/NNW.2012.22.032.

[3] *Berlin Emotional Speech Database*. Available from: http://www.expressive-speech.net/

[4] BOERSMA P., WEENINK D. Praat: doing phonetics by computer 5.0.32 [software]. 2008-08-12 [accessed 2009-07-15]. Available from: http://www.praat.org/

[5] BOSER B.E., GUYON I.M., VAPNIK V.N. A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT)*, Pittsburgh, PA, USA. NY, USA: ACM, 1992, pp. 144-152, doi: 10.1145/130385.130401.

[6] CASALE S., et al. Speech emotion classification using machine learning algorithms. In: *Proceedings of the IEEE International Conference on Semantic Computing*, Santa Clara, CA, USA. MA, USA: IEEE, 2008, pp. 158-165, doi: 10.1109/ICSC.2008.43.

[7] CHEN S., BILLINGS S.A., Luo W. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*. 1989, 50(5), pp. 1873-1896, doi: 10.1080/00207178908953472.

[8] *Corpus LDC*. Available from: http://www.ldc.upenn.edu/

[9] CRISTIANINI N., SHAWE-TAYLOR J. *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[10] DAASSI-GNABA H., LOPEZ KRAHE J. Universal combined system: speech recognition, emotion recognition and talking head for deaf and hard of hearing people. In: *Proceedings of the 10th Association for the Advancement of Assistive Technology in Europe (AAATE)*, Florence, Italy. Netherlands: IOS Press, 2009, pp. 503-508, doi: 10.3233/978-1-60750-042-1-503.

[11] DAASSI-GNABA H., OUSSAR Y. Enhanced emotion recognition by feature selection to animate a talking head. In: *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium. Louvain-La-Neuve, Belgium: i6doc.com, 2012, pp. 513-518.

[12] EL AYADI M., KAMEL M.S., KARRAY F. Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition*. 2011, 44(3), pp. 572-587, doi: 10.1016/j.patcog.2010.09.020.

[13] GUYON I., ELISSEEFF A. An Introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003, 3, pp. 1157-1182.

[14] GUYON I., et al. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002, 46(1-3), pp. 389-422, doi: 10.1023/A:1012487302797.

[15] HASTIE T., TIBSHIRANI R., FRIEDMAN J. *The Elements of Statistical Learning*. New-York: Springer, 2009.

[16] HO Y.C., KASHYAP R.L. An algorithm for linear inequalities and its applications. *IEEE Transactions on Electron Computer*. 1965, 14(5), pp. 683-688, doi: 10.1109/PGEC.1965.264207.

[17] INGALE A.B., CHAUDHARI D.S. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*. 2012, 2(1), pp. 235-238.

[18] KWON O.W., et al. Emotion recognition by speech signals. In: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland. Bonn, Germany: Wolfgang Hess, 2003, pp. 125-128.

[19] PUDIL P., NOVOVIČOVÁ J., KITTLER J. Floating search methods in feature selection. *Pattern Recognition Letters*. 1994, 15(11), pp. 1119-1125, doi: 10.1016/0167-8655(94)90127-9.

[20] RABINER L.R., SCHAFER R.W. *Introduction to Digital Speech Processing*. Hanover: Now Publishers Inc, 2007, doi: 10.1561/2000000001.

[21] SCHULLER B., et al. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: *Proceedings of the 6th Annual Conference of the International Speech Communication Association (ISCA), INTERSPEECH*, Lisboa, Portugal. Bonn, Germany: Wolfgang Hess, 2005, pp. 805-809.

[22] SCHULLER B., STEIDL S., BATLINER A. The INTERSPEECH 2009 emotion challenge. In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association (ISCA), INTERSPEECH*, Brighton, United Kingdom. Bonn, Germany: Wolfgang Hess, 2009, pp. 312-315.

[23] SCHULLER B., et al. The INTERSPEECH 2010 paralinguistic challenge. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (ISCA), INTERSPEECH*, Makuhari, Chiba, Japan. Bonn, Germany: Wolfgang Hess, 2010, pp. 2794-2797.

[24] SCHULLER B., et al. AVEC 2011–The first international audio/visual emotion challenge. In: *Proceedings of the 4th International Conference Affective Computing and Intelligent Interaction (ACII)*, Memphis, Tennessee, USA. Berlin, Germany: Springer, 2011, pp. 415-424.

[25] SCHULLER B., et al. AVEC 2012–The continuous audio/visual emotion challenge. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*, Santa Monica, California, USA. New York ACM, 2012, pp. 449-456.

[26] SCHULLER B., BATLINER A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Chichester: Wiley, 2014.

[27] The Spider 1.71 [software]. 2006-07-26 [accessed 2007-10-01]. Available from: http://people.kyb.tuebingen.mpg.de/spider/

[28] TAHON M., DELABORDE A., DEVILLERS L. Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (ISCA), INTERSPEECH*, Florence, Italy. Bonn, Germany: Wolfgang Hess, 2011, pp. 3121-3124.

[29] TUCKOVA J., SEBESTA V. Prosody optimisation of a Czech language synthesizer. *Neural Network World*. 2008, 18(4), pp. 291-308.

[30] YANG N., et al. Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion. In: *Proceedings of the 4th IEEE Workshop on Spoken Language Technology (SLT)*, Miami, Florida, USA. Miami IEEE, 2012, pp. 455-460, doi: 10.1109/SLT.2012.6424267.