Speaker-independent negative emotion recognition

Margarita Kotti, Fabio Paternò
ISTI-CNR, Via G. Moruzzi, 1, 56124 Pisa, Italy
Email: {margarita.kotti,Fabio.Paterno}@isti.cnr.it

Constantine Kotropoulos
Dept. of Informatics AUTH, Thessaloniki 54124, Greece
Email: costas@aiia.csd.auth.gr

***Abstract*—This work aims to provide a method able to distinguish between negative and non-negative emotions in vocal interaction. A large pool of 1418 features is extracted for that purpose. Several of those features are tested in emotion recognition for the first time. Next, feature selection is applied separately to male and female utterances. In particular, a bidirectional Best First search with backtracking is applied. The first contribution is the demonstration that a significant number of features, first tested here, are retained after feature selection. The selected features are then fed as input to support vector machines with various kernel functions as well as to the K nearest neighbors classifier. The second contribution is in the speaker-independent experiments conducted in order to cope with the limited number of speakers present in the commonly used emotion speech corpora. Speaker-independent systems are known to be more robust and present a better generalization ability than the speaker-dependent ones. Experimental results are reported for the Berlin emotional speech database. The best performing classifier is found to be the support vector machine with the Gaussian radial basis function kernel. Correctly classified utterances are 86.73%±3.95% for male subjects and 91.73%±4.18% for female subjects. The last contribution is in the statistical analysis of the performance of the support vector machine classifier against the K nearest neighbors classifier as well as the statistical analysis of the various support vector machine kernels impact.**

I. Introduction

Next-generation pervasive computing environments resort to human-centered designs instead of computer-centered designs. Traditional human-computer interaction (HCI) designs ignore emotions [1]. Such monolithic interactions are frequently perceived as cold, incompetent, and socially inept. In today's HCI, an agent would be able to engage emotionally if it has empathy, i.e. it can understand what a person's emotional state might dispose him or her to do, and how that disposition might be affected by different actions that the agent might take. hence, tools for user interface support and evaluation need to include models of emotion.

We are interested to exploit the affective information in order to create an HCI evaluation tool. Our goal is to assist detecting problems that arise from dissatisfactory interactions, since inability to rectify negative feelings may impede concentration, cognitive capacity and decision making [2]. Thus, by discriminating negative from non-negative emotions HCI designers will be able to recognize which parts of the interface are problematic, in the sense that they evoke negative emotions, and consequently improve them, thus boosting the quality of HCI. Furthermore, by discriminating the negative emotions from the non-negative ones, we move from categorical emotion representation to dimensional descriptors [3], that have attracted the interest of the research community, albeit there is no general agreement against categorical labeling in cognitive theory. Other possible applications include games, call-center management, educational software, life-support systems, commercial products, virtual guides, customer service, in-car driver interfaces, surveillance systems, conference room research, art, etc [4] [5].

Previous work has underlined the need to recognize the negative emotions. Anger was detected in recordings from a German voice portal in [6]. Support vector machines (SVMs) and Gaussian mixture model (GMM)-based classifiers were applied to pitch, energy, duration, and spectral-related features. No feature selection was exploited. GMM-based classifiers on pitch and energy succeeded the best speaker-dependent performance (i.e. F_1=0.70), when 90% of the utterances were used for training and the remaining 10% for testing. The detection of annoyance and frustration was studied in [5] by utilizing a dialogue travel database developed under the DARPA communication project. A combination of prosodic, speaking style and language model information features was employed. Features selected by a brute-force iterative feature selection algorithm were provided as input to decision trees. A speaker-dependent 75%/25% training/testing schema was adopted. The set of annoyance and frustration is recognized at a rate varying between 64.5% and 85.4%, depending on the extracted features as well as the application of word recognition. Interest detection is studied in [4]. In particular, three levels of interest were identified, ranging from curiosity to disinterest. The following acoustic features were extracted: formants, pitch, frame energy, envelope, Mel-frequency cepstral coefficients (MFCC), harmonics-to-noise ratio, jitter, and shimmer. Speaker-independent leave-one-speaker-out experiments were conducted. Feature selection was performed by sequential forward floating search at each iteration independently. SVMs with a polynomial kernel were used as classifiers. A mean accuracy of 69.2% for all the three levels of interest was reported, when the audio channel is exploited only. A multi-cue, dynamic approach in audiovisual sequences was presented in [7]. Recognition was performed by a simple recurrent neural network. The activation-valence emotional space was examined [3]. The speech features were related to prosody, such as the pitch and the rhythm. No feature selection was applied. The emotional classes were 4. That is, 1 for the neutral state and another 3 for all the quadrants defined on the activation-valence space but the positive/passive one. The latter was neglected. The SAL database was used for evaluation.

The speaker-dependent emotion classification accuracy was measured to be 73% for the audio only, when the ratio of the training and test datasets was 3 to 1.

There are several novel contributions in this paper. To begin with, a large pool of 1418 features is extracted. Several features are proposed here for the first time within the context of emotion recognition. Feature selection by a Best First strategy is applied next to male and female utterances separately. The first contribution is the demonstration that a significant number of the features, that are first tested here for emotion recognition, are retained after feature selection. The selected features are fed as input to SVMs as well as K nearest neighborhood (KNN) classifiers. Various kernel functions are tested for SVMs, such as the polynomial, the multilayer perceptron, and the Gaussian radial basis function (RBF). The second contribution is in conducting speaker-independent experiments. A thorough literature survey has revealed only a few works where speaker-independent emotion recognition assessment is conducted. The SVM with the Gaussian RBF kernel is found to be the best performing classifier. The third contribution is related to statistical analysis of the classifiers' performance. In particular, the KNN and the SVM with the Gaussian RBF kernel are compared with means of Q-statistic whereas one-way analysis of variance (one-way ANOVA) followed by Tukey's method is applied in order to compare if the different SVM kernels lead to statistically different performance gains.

The outline of the paper is as follows. In Section II, the database is described and the extracted features are summarized in Section III. Feature selection is addressed in Section IV. The experimental procedure is detailed in Section V and conclusions are drawn in Section VI.

II. Database

We are currently working on designing the collection protocol of a database of vocal interactions derived from a VoiceXML application. Thus, for the time being, we had to confine ourselves to an already publicly available database. We have resorted to the widely used Berlin emotional speech database (EMODB), whose emotional quality is ensured by human annotators. In the EMODB, 10 subjects simulate 7 emotional states, namely: anger, fear, joy, sadness, disgust, boredom, and neutral. The subjects are divided into 5 actors and 5 actresses. Each of them utters 10 utterances in German. The set of negative emotions comprises the states of anger, fear, sadness, disgust and boredom. Such emotional states possess a negative valence in the activation-valence emotion description space [3]. Non-negative emotions include the joy and the neutral state. The recordings are mono-channel, sampled at a frequency of 16 KHz, and their audio format is PCM. The audio samples have been quantized in 16 bits. The full database comprises approximately 30 minutes of speech. The database consists of 535 utterances with 233 utterances uttered by male subjects whereas the remaining 302 ones are uttered by female subjects. The EMODB is publicly available [8].

III. Feature Extraction

Cognitive scientists have not yet identified the optimal set of features that reliably discriminates among the emotional states [1]. Here, a large number of 1418 features has been extracted. Our aim is two-fold. On the one hand, we attempt to compute a multitude of features so that an exhaustive feature set is created. On the other hand, several features are investigated for discrimination between negative and non-negative emotions for the first time.

The features are related to:

- the pitch contour
- the formants contours
- the energy contour
- the features derived from the sound description toolbox [9], that is general audio description including the MPEG-7 audio framework, such as
  - spectral features
  - temporal features
  - short-time energy
  - MPEG-7 low level descriptors (i.e. AudioPower, AudioFundamentalFrequency, AudioSpectrumSpread, AudioSpectrumFlatness, LogAttackTime, TemporalCentroid and AudioSpectrumRolloff frequency)
  - MFCCs
  - total loudness and specific loudness sensation coefficients (SLSC)
- the Teager energy operator on autocorrelation
- the Fujisaki's model parameters [10]
- the jitter and the shimmer.

Pitch is computed based on an autocorrelation method, while the method to estimate formants relies on the linear prediction analysis. The first and second order feature differences are also computed in order to capture the feature temporal evolution. Thus, a multivariate time series is obtained and a dynamic model is feasible [11]. To reduce the dependency on the spoken phonetic content, several statistics of the features are extracted [4] [11], such as the maximum, the minimum, the variance, the mean, the median, the skewness, the interquartile range, and the 90th percentile. Finally, normalization takes place, since the different features may possess different scales. It aims to modify the mean and standard deviation (st. dev.) of the features values in order to ensure an equal contribution of each feature to the feature selection algorithm. Moreover, the normalization helps to eliminate the outliers. All features are subject to min-max normalization hereafter. Min-max normalization is expected to boost performance, since it preserves all original feature relationships and does not introduce any bias in the features. Let min_F and max_F be the minimum and the maximum of feature F. Min-max normalization maps the interval $[min_F, max_F]$ into a new interval $[new_{min_F}, new_{max_F}]$. Thus, any value n from the original interval is mapped into value new_n according to: $new_n = \frac{n - min_F}{max_F - min_F}(new_{max_F} - new_{min_F}) + new_{min_F}$. Here, $new_{min_F} = 0$ and $new_{max_F} = 1$. To the best of the authors' knowledge, the MPEG-7 descriptors, the Teager energy operator on autocorrelation, the total loudness, and the SLSC are investigated within the context of emotion recognition for the first time here.

IV. Feature Selection

Feature selection is applied next, because a small feature set requires less memory, fewer computations and generally possesses greater generalization abilities than a large feature set. Indeed by employing large feature sets the possibility to include features with less discriminating power increases not to mention the excess computation time and storage requirements. Initial experiments without feature selection, i.e. by exploiting all extracted features, demonstrate a performance of 64.15%±3.59% for male subjects and 62.22%±9.33% for female subjects.

Feature selection is applied separately to male and female utterances. It is widely accepted by the research community that the two genders convey their emotions in profoundly different ways [1] [12]. Some features are gender-dependent, e.g. the pitch. It is well known that in general female speech has higher pitch than male speech, due to the increase in mass of a male's vocal folds. Furthermore, the smaller vocal tract dimensions in women rather than men produces higher formant frequencies for women. Moreover, this approach is compatible to our aim to build an HCI evaluation tool, where the evaluator can specify his or her gender. However, for the general case, a gender recognizer is also available [13].

The feature selection algorithm is the *Best First*. Best First searches the feature space by greedy hill climbing. That is, a search is conducted in the feature lattice in order to find the node that is optimal with respect to the evaluation function. Here, the evaluation function is the *Pearson correlation*. So, the feature subsets that are highly correlated within the same class (i.e., the negative emotions or the non-negative ones), while possessing a low correlation across the classes are preferred. Backtracking and bidirectional search are also exploited. The selected features for the male utterances are listed in Table I and for female utterances in Table II. In particular, 61 features are retained for male utterances and 50 for the female ones.

Exactly 6 features coincide between the male and female selected feature subsets, verifying the importance of the speaker gender for the emotion recognition task. However, one can easily check that the two sets of selected features contain common categories, a result that was also reported in [2]. To be more specific, the selected features for both male and female utterances are related to the pitch contour, the energy contour, the MPEG-7 low level descriptors, the SLSC, the MFCCs, and the Fujisaki's model parameters. It is worth noticing that pitch and energy are among the most commonly used features in speech emotion recognition research. The MFCCs have also been tested previously. There has been limited work on Fujisaki's model parameters. MPEG-7 low level descriptors and SLSC are tested within the context of speech emotion recognition for the first time in this paper.

TABLE I
SELECTED FEATURES WHICH DISCRIMINATE BEST THE NEGATIVE EMOTIONS FROM THE NON-NEGATIVE ONES IN MALE UTTERANCES.

pitch existence in the utterance (expressed in %)
median of durations in the plateaux of pitch at maxima
mean of pitch values within the plateaux at maxima
maximum of pitch
variance of pitch
interquartile range of pitch
skewness of the second order differences of pitch
interquartile range of durations in the plateaux of energy at minima
mean of energy values within the plateaux at minima
median of energy values within the plateaux at maxima
mean of durations in the rising slopes of energy contours
position of the first energy maximum
st. dev. of energy in the rising slopes of energy contours
energy below 2800 Hz
energy in the frequency band 600-1500 Hz
energy in the frequency band 250-1500 Hz
minimum of the Teager energy operator on autocorrelation
interquartile range of the short-time energy
interquartile range of the short-time energy first order differences
maximum of the 5th SLSC
minimum of the 1st SLSC first order differences
skewness of the 7th SLSC first order differences
interquartile range of the 6th SLSC first order differences
interquartile range of the 9th SLSC first order differences
90th percentile of the 9th SLSC first order differences
variance of the 5th SLSC second order differences
skewness of the 6th SLSC second order differences
skewness of the 7th SLSC second order differences
interquartile range of the 8th SLSC second order differences
90th percentile of the 8th SLSC second order differences
variance of the AudioSpectrumCentroid first order differences
maximum of the AudioSpectrumRolloff frequency
minimum of the AudioSpectrumSpread first order differences
maximum of the 11th MFCC
maximum of the 17th MFCC
minimum of the 6th MFCC
minimum of the 14th MFCC
variance of the 7th MFCC
mean of the 5th MFCC
median of the 5th MFCC
median of the 7th MFCC
interquartile range of the 20th MFCC
mean of the 5th MFCC first order differences
variance of the 5th MFCC second order differences
variance of the 9th MFCC second order differences
median of the 5th MFCC second order differences
skewness of the 5th MFCC second order differences
skewness of the 6th MFCC second order differences
skewness of the 7th MFCC second order differences
interquartile range of the 5th MFCC second order differences
interquartile range of the 8th MFCC second order differences
maximum of Fujisaki's $F0$ contour
variance of Fujisaki's $F0$ contour first order derivative
mean of Fujisaki's $F0$ contour second order derivative
90th percentile of Fujisaki's $F0$ contour second order derivative
minimum of the high-pass filter output contour
skewness of the high-pass filter output contour
90th percentile of the accent component first order derivative
mean of the phase component
variance of the accent commands second order derivative
mean of the accent commands second order derivative

V. Experimental procedure

A. The Applied Classifiers

The classifiers employed for discriminating negative from non-negative emotions are the KNN and the SVM. The

TABLE II

SELECTED FEATURES WHICH DISCRIMINATE BEST THE NEGATIVE EMOTIONS FROM THE NON-NEGATIVE ONES IN FEMALE UTTERANCES.

mean of the 3rd formant
skewness of the 3nd formant
interquartile range of durations within the falling slopes of pitch contours
maximum of durations in the plateaux of energy contour at minima
median of durations in the plateaux of energy contour at maxima
maximum of AudioFundamentalFrequency second order differences
mean of the 3rd SLSC
interquartile range of the 1st SLSC
minimum of the 1st SLSC first order differences
minimum of the 3rd SLSC first order differences
minimum of the 8th SLSC first order differences
skewness of the 1st SLSC first order differences
variance of the 2nd SLSC second order differences
mean of the 4th SLSC second order differences
skewness of the 5th SLSC second order differences
interquartile range of the 8th SLSC second order differences
skewness of the AudioSpectrumCentroid
maximum of the AudioSpectrumCentroid first order differences
skewness of the AudioSpectrumCentroid first order differences
interquartile range of the AudioSpectrumRolloff frequency
minimum of the AudioSpectrumRolloff frequency first order differences
variance of the AudioSpectrumRolloff frequency first order differences
90th percentile of the AudioSpectrumRolloff frequency second order differences
maximum of the 9th MFCC
minimum of the 6th MFCC
variance of the 9th MFCC
variance of the 16th MFCC
variance of the 17th MFCC
median of the 7th MFCC
skewness of the 24th MFCC
interquartile range of the 7th MFCC
variance of the 6th MFCC first order differences
mean of the 5th MFCC first order differences
interquartile range of the 5th MFCC first order differences
90th percentile of the 6th MFCC first order differences
maximum of the 7th MFCC second order differences
variance of the 6th MFCC second order differences
variance of the 7th MFCC second order differences
skewness of the 5th MFCC second order differences
skewness of the 24th MFCC second order differences
interquartile range of the 1st AudioSpectruFlatness coefficient second order differences
minimum of Fujisaki's $F0$ contour first order derivative
90th percentile of Fujisaki's $F0$ contour first order derivative
median of Fujisaki's $F0$ contour second order derivative
90th percentile of Fujisaki's logarithmic $F0$ spline
minimum of the low-pass filter output contour
skewness of the accent commands first order derivative
interquartile range of the accent commands first order derivative
median of phrase commands
interquartile range of the phrase commands

classification accuracy is tested for several parameterizations. The best parameter is indicated by experimentation, as in [2]. SVMs are ideal for the case under consideration, since they create a hyperplane that separates the data into two classes with the maximum-margin. Let $\mathbf{v}_i$ be the ith training vector. Three different kernels are used.

- Polynomial:
$$\mathcal{K}_{SVM}(\mathbf{v}_i, \mathbf{v}_j) = \left(\mathbf{v}_i^T \mathbf{v}_j\right)^M, \quad (1)$$
with M being the polynomial order;
- Multilayer perceptron:
$$\mathcal{K}_{SVM}(\mathbf{v}_i, \mathbf{v}_j) = S\left(\mathbf{v}_i^T \mathbf{v}_j - 1\right), \quad (2)$$
where $S(\cdot)$ is the sigmoid function;
- Gaussian RBF:
$$\mathcal{K}_{SVM}(\mathbf{v}_i, \mathbf{v}_j) = \exp(-\gamma ||\mathbf{v}_i - \mathbf{v}_j||^2), \quad (3)$$
where γ is a scaling factor.

The corresponding SVMs are referred to as SVMPOL, SVMMLP, and SVMRBF, respectively.

B. Experimental Protocol and Figures of Merit

With the term speaker-independent we mean that the utterances that are included in the test set come from a specific speaker, whose utterances are not included in the training set. Only a few researchers have conducted speaker-independent experiments. Speaker-independent systems are more robust and demonstrate a better generalization ability than the speaker-dependent ones. Moreover, the speaker-independent systems are suitable for real-life applications, such as call-center applications, media segmentation, public transport surveillance. Furthermore, speaker-independent systems can cope with the limited number of speakers present in the commonly used emotion speech corpora [14]. An additional advantage of the speaker-independent systems is the fact that the experimental protocol is deterministic, in the sense that the exact configuration is known. This way, result comparisons are facilitated in contrary to random speaker-dependent partitions that can not be reproduced exactly [15]. Moreover, it is reported that speaker-dependent emotion recognition leads to far better results than speaker independent modeling. For example in [14], 10 different classifiers are tested. The averaged performance equals 89.49% for the speaker-dependent case, whereas it drops to 71.29% for the speaker-independent one. In order to measure speaker-independent emotion recognition rate, leave-one-speaker-out evaluation is applied and the average rate is reported over the 5 male subjects as well as the 5 female subjects. That is, for each gender separately, the classifier is trained 5 times, each time leaving one speaker out of the training set and then testing the performance on the utterances of the speaker left out. The interest is the assessment of emotion recognition accuracy without mixing it with speaker recognition. This is guaranteed by the proposed protocol.

Let us define the figures of merit used. Let $hits$ be the number of utterances that are classified correctly and $misses$ the number of utterances that are classified incorrectly. Then, the ratio of correctly classified utterances (CCU) equals

$$CCU = \frac{hits}{hits + misses}. \quad (4)$$

For incorrectly found utterances (ICU) it holds that $ICU = 1 - CCU$, while the root mean squared error ($RMSE$) is defined as $RMSE = \sqrt{ICU}$. The standard deviation of each figure of merit across the 5 evaluations is also reported.

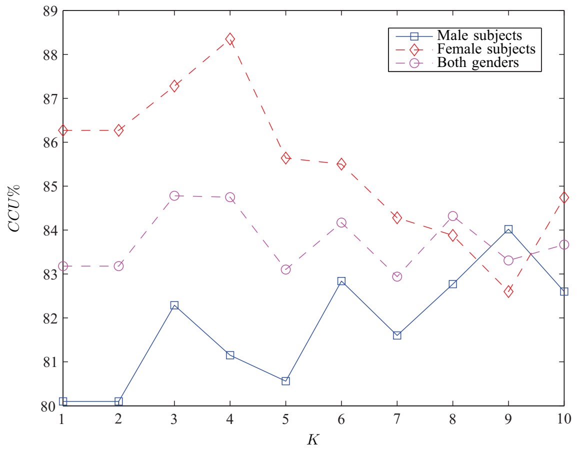

Fig. 1. Speaker-independent CCU(%) of the KNN versus various values of K for male utterances, female utterances, and utterances of both genders.

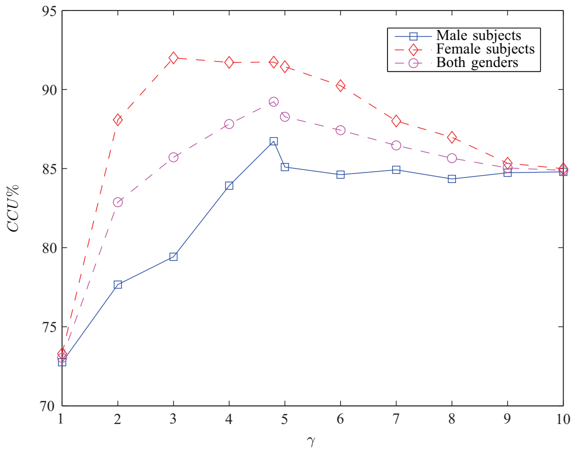

Fig. 2. Speaker-independent CCU (%) of the SVMRBF versus various values of γ for male utterances, female utterances and utterances of both genders.

The KNN classifier is used as a baseline classifier. The distance metric used within the KNN is based on the correlation. It is defined as one minus the sample correlation between the training vectors. Ten different numbers of nearest neighbors $K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ are tested, as can be seen in Figure 1. In general, CCU in female utterances tends to be higher than that in male utterances, especially for low and medium K values. The best average ratio of CCU is measured for K=3 neighbors. Exact figures of merit are listed in Table III. Concerning the SVMPOL, values of $M \in$ {3,4,5,6,7,8,9,10} are examined. The best performance is obtained when M=3 as can be seen in Table III where experimental results for the SVMMLP are also included. The SVMRBF is tested for various γ values, as demonstrated in Figure 2. γ is a real parameter whose range is $[1, 10]$. The best performance is obtained for $\gamma = 4.8$. The average ratio of the CCU for the SVMRBF is rather poor for small γ values, while it converges to 85% for large γ values. The figures of merit for the SVMRBF, when $\gamma = 4.8$, are exhibited in Table III.

For discussion, in [2] the authors also aim to negative and non-negative emotion detection. The data are derived form a call-center. Linear discriminant classifiers (LDC) and KNNs were tested for each gender utterances separately. Parameters such as the fundamental frequency, energy, duration, and formant features were extracted. Then forward feature selection was applied followed by principal component analysis. Results were reported for 10-fold cross-validation making the experimental procedure speaker-dependent. Concerning exclusively the audio channel, the lowest classification error of the LDCs was 17.85% and 12.04% in male utterances and female ones, respectively.

C. Statistical Assessment

Two sets of comparisons are carried out. The first one refers to testing the dependency between the KNN and the SVMRBF. The second one compares the performance gains for the 3 different kernels of the SVM.

1) Comparing Classifiers: The SVMRBF is considered, since the Gaussian RBF kernel yields the best CCU for both genders, as can be seen in Table III. Thus, we examine the dependency between the KNN and the SVMRBF with respect to the average ratio of CCU over both genders. Q-statistic is calculated to measure the dependency between the classifiers [16]. Q-statistic measures the pairwise symmetrical similarity. For two classifiers, Q-statistic is defined as:

$$Q = \frac{N_{11}\ N_{00} - N_{01}\ N_{10}}{N_{11}N_{00} + N_{01}N_{10}}, \tag{5}$$

where N_{11} is the number of utterances both KNN and SVMRBF classify correctly, N_{10} is the number of utterances KNN classifies correctly while the SVMRBF classifies incorrectly, N_{01} is the number of utterances the KNN classifies incorrectly while the SVMRBF classifies correctly and N_{00} is the number of utterances both the KNN and the SVMRBF classify incorrectly. The Q-statistic admits values between -1 and 1. If the classifiers are statistically independent, the Q-statistic equals 0. If the classifiers tend to recognize correctly the same utterances then positive values of Q-statistic are admitted, whereas for classifiers, which commit errors on different utterances, a negative Q-statistic is rendered. In our case, $Q = 0.803$. This can be attributed to the fact that both classifiers are trained and tested on the same data.

2) Comparing Kernels: There are 3 different SVM variants, namely the SVMPOL, the SVMMLP and the SVMRBF, due to the 3 different kernels employed. In particular, for the SVMPOL and the SVMRBF the parameters that yield the highest CCU are considered, i.e. $M = 3$ for the SVMPOL and $\gamma = 4.8$ for the SVMRBF. One-way ANOVA is applied, to test whether or not the 3 SVM variants are of equal average CCU. At 95% confidence level, the p-value equals 1.546×10^{-5}, that is less than 0.05, meaning that the average CCU differences are due to systematic variations and not due to randomness. However, no information is provided by one-way ANOVA about the pairs of the SVM variants that differentiate

TABLE III
SPEAKER-INDEPENDENT FIGURES OF MERIT ON EMODB FOR KNN, SVMPOL, SVMMLP AND SVMRBF FOR BOTH GENDERS.

	KNN (K=3)		SVMPOL (M=3)		SVMMLP		SVMRBF ($\gamma = 4.8$)	
	male	female	male	female	male	female	male	female
CCU% (mean)	81.16	88.36	85.60	89.75	74.70	69.89	86.73	91.73
CCU% (st. dev.)	4.00	4.24	8.14	6.99	4.32	11.69	3.95	4.18
RMSE (mean)	0.43	0.34	0.37	0.30	0.50	0.54	0.36	0.28
RMSE (st. dev.)	0.05	0.06	0.10	0.12	0.04	0.12	0.05	0.09

statistically. Thus, Tukey's method is applied to find these SVM variants that differentiate at the 95% confidence level. Tukey's method is optimal for one-way ANOVA, which is our scenario. The confidence intervals for all pairwise averaged CCU comparisons among the 3 SVM variants are listed in Table IV. If the confidence interval includes 0, the pairwise average CCU difference is not statistically significant.

TABLE IV
PAIRWISE COMPARISONS BETWEEN THE SVMPOL, THE SVMMLP AND THE SVMRBF.

classifier variants compared	95% confidence interval	statistically different
SVMPOL, SVMMLP	[0.07, 0.23]	yes
SVMPOL, SVMRBF	[-0.09, 0.06]	no
SVMMLP, SVMRBF	[-0.25, -0.09]	yes

VI. CONCLUSIONS

This paper has dealt with the discrimination between negative and non-negative emotions in speech utterances. 1418 features have been extracted and the best features have been selected by the bidirectional Best First algorithm with backtracking separately for male and female utterances. In both cases, the selected feature subset includes features that are first tested here for emotion recognition. The selected features have been fed as input to the KNN, the SVMPOL, the SVMMLP and the SVMRBF. Speaker-independent experiments have been conducted on the EMODB using a leave-one-speaker-out evaluation. A high ratio of correctly classified utterances has been reported for both male and female subjects by the best performing SVMRBF classifier. Statistical tests have shown that the Q-statistic equals 0.803, when comparing the KNN and the SVMRBF, whereas the SVMRBF and the SVMPOL have been shown to attain an equal performance. In the future, we plan to apply a hierarchical strategy, where the recognition of negative and non-negative emotions will be the performed first and next a more detailed classification to emotional states will be conducted. Within such an hierarchical scheme, it would be easier to accommodate different number of speakers and different number of emotional states.

ACKNOWLEDGMENT

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," in *Proc. 9th Int. Conf. Multimodal Interfaces*, November 2007, pp. 126–133.

[2] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 12, pp. 293–303, March 2005.

[3] M. Schröder, R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, and M. Sawey, "FEELTRACE: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, September 2000, pp. 1–4.

[4] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image Vision Computing*, vol. 27, no. 12, pp. 1760–1774, November 2009.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. Int. Conf. Spoken Language Processing*, September 2002, pp. 2037–2040.

[6] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting real life anger," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, April 2009, pp. 4761–4764.

[7] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Proc. 8th Int. Conf. Multimodal Interfaces*, November 2006, pp. 146–154.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th European Conf. Speech Communication and Technology*, September 2005, pp. 1517–1520, http://pascal.kgw.tu-berlin.de/emodb/index-1024.html.

[9] E. Benetos, M. Kotti, and C. Kotropoulos, "Large scale musical instrument identification," in *Proc. 4th Sound and Music Computing Conference*, July 2007, pp. 283–286, http://www.ifs.tuwien.ac.at/mir/muscle/del/audio_tools.html.

[10] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, June 2000, pp. 1281–1284.

[11] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. 9th Int. Conf. Multimodal Interfaces*, November 2007, pp. 30–37.

[12] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proc. IEEE Int. Conf. Multimedia and Expo*, July 2005, pp. 1500–1503.

[13] M. Kotti and C. Kotropoulos, "Gender classification in two emotional speech databases," in *Proc. 19th Int. Conf. Pattern Recognition*, December 2008, pp. 4898–4901.

[14] B. Schuller, R. Villar, G. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, March 2005, pp. 325–328.

[15] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. 10th Annual Int. Conf. Speech Communication Association*, September 2009, pp. 312–315.

[16] L. Kuncheva and C. Whitaker, "Measure of diversity in classifier ensembles," *Machine Learning*, vol. 51, no. 2, pp. 181–207, May 2003.