

Speech Emotion Classification and Public Speaking Skill Assessment

Tomas Pfister and Peter Robinson

University of Cambridge
Computer Laboratory, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
{tjp35,pr10}@cam.ac.uk

Abstract. This paper presents a new classification algorithm for real-time inference of emotions from the non-verbal features of speech. It identifies simultaneously occurring emotional states by recognising correlations between emotions and features such as pitch, loudness and energy. Pairwise classifiers are constructed for nine classes from the Mind Reading emotion corpus, yielding an average cross-validation accuracy of 89% for the pairwise machines and 86% for the fused machine. The paper also shows a novel application of the classifier for assessing public speaking skills, achieving an average cross-validation accuracy of 81%. Optimisation of support vector machine coefficients is shown to improve the accuracy by up to 25%. The classifier outperforms previous research on the same emotion corpus and achieves real-time performance.

1 Introduction

Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues.

Speech emotion analysis refers to analysing vocal behaviour as a marker of affect, with focus on the non-verbal aspects of speech. Its basic assumption is that there is a set of objectively measurable features in voice that reflect the affective state of a person. This assumption is supported by the fact that most affective states involve physiological reactions which modify the process by which voice is produced. For example, anger often produces changes in respiration and increases muscle tension, influencing the vibration of the vocal folds and vocal tract shape, thus affecting the acoustic characteristics of the speech [1].

Discovering which features are indicative of emotional states and consecutively capturing them can be a difficult task. Furthermore, features indicating different states may be overlapping, and there may be multiple sets of features expressing the same emotional state. One widely used strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others. This avoids making difficult *a priori* decisions about which features may be relevant.

Previous studies indicate that several emotions can occur simultaneously [2]. Examples of co-occurring emotions include being happy at the same time as

being tired, or feeling touched, surprised and excited when hearing good news. Improving upon the inference solution for co-occurring emotions presented by Sobol Shikler [3], the new system proposed in this paper is able to achieve real-time performance and higher classification accuracy.

In this paper, we describe an approach for real-time classification of co-occurring emotions. The classification output is a set of classes rather than a single one, allowing nuances and mixtures of emotions to be detected. Moreover, rather than attempting to make difficult *a priori* decisions about which features may be relevant, our strategy is to compute as many features as possible, and then select those offering the best discrimination. Finally, we present a novel application of the classifier to virtual speech coaching for improving public speaking skills.

2 Implementation Methodology

The design of the classifier considers three main factors: (i) the choice of a training corpus, (ii) the need for real-time performance, (iii) the ability to recognise co-occurring emotions.

For emotion classification we choose the Mind Reading corpus [4] which provides a hierarchical structure between groups with a large number of emotion concepts. It was developed by psychologists at University of Cambridge Autism Research Centre, aiming to help autistic children and adults to recognise both basic and complex emotions. The corpus consists of 2927 acted sentences, covering 442 different concepts of emotions, each with 5-7 sentences. The acting was induced and the labelling was done by ten people in different age groups [5]. The labelling of each sample in the corpus required the agreement of 8 members of a panel of 10 expert assessors. Although the samples are acted, the large number of samples makes the corpus suitable for training an emotion classifier.

The main emotion groups of Mind Reading are shown in Table 1. Each of these is further divided into concepts, giving a total of 422 subgroups. For the classifier, a subset of 9 categories representing a large variety of emotions is chosen. Each category contains samples from the groups as shown in Table 1. These are chosen to minimise the overlap between categories. The categories and samples are the same as those used by Sobol Shikler [3, 6], allowing direct comparison of results.

Achieving real-time performance required a careful choice of feature extraction and classification algorithms. Recognising co-occurring emotions needed a method for ranking candidate emotions.

2.1 Support Vector Machines

Several potential classifiers were investigated. In previous work on emotion recognition from speech [3], support vector machines (SVMs) and tree algorithms such as C4.5 have been found to be effective. We also tried other methods such as the

Table 1. The 24 emotion groups in the Mind Reading corpus [5]. The superscripts indicate the main groups from which a subset of affective states is selected to allow comparison of the results to previous research [3]. These subsets are: absorbed¹, excited², interested³, joyful⁴, opposed⁵, stressed⁶, sure⁷, thinking⁸ and unsure⁹.

afraid	angry	bored	bothered ¹	disbelieving
disgusted	excited ²	fond	happy ³	hurt
interested ^{4,5}	kind	liked	romantic	sad
sneaky	sorry	sure ⁶	surprised	think ⁷
touched	unfriendly ⁸	unsure ⁹	wanting	

Naive Bayesian classifier and Perceptrons using the Weka data mining toolkit [7], but SVMs gave the most promising results.

We create the model by constructing an N -dimensional hyperplane that optimally separates data into two categories. Each data instance i is a tuple (l_i, \mathbf{f}_i) , where $l_i \in \{1, -1\}$ is a class label, with 1 and -1 indicating the class, and $\mathbf{f}_i \in \mathbb{R}^n$ is a set of feature attributes. Optimality is taken to be the maximal separation between the two classes. Any such hyperplane can be written as the set of points \mathbf{x} satisfying $\mathbf{w} \cdot \mathbf{x} - b = 0$ where $\mathbf{x} = \mathbf{f}_i$, \mathbf{w} is the normal vector perpendicular to the hyperplane, $\|\mathbf{w}\|$ is the Euclidean norm of \mathbf{w} , and $\frac{|b|}{\|\mathbf{w}\|}$ is the perpendicular distance from the hyperplane to the origin.

We use a modified version of SVMs [8] that allows for mislabelled examples by choosing a hyperplane as cleanly as possible even if there is no hyperplane that can split the two classes. We measure this degree of misclassification by the variable ξ_i and require the solution of the optimisation problem

$$\min_{\mathbf{w}, b, \xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \right\} \quad (1)$$

under constraints

$$l_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i \quad 1 \leq i \leq n \quad (2)$$

$$\xi_i \geq 0. \quad (3)$$

where $C > 0$ is the penalty for mislabelled examples and n is the number of data instances in the corpus. This can be solved using Lagrange multipliers.

We use a non-linear classifier, replacing the linear dot product $\mathbf{x}_i \cdot \mathbf{x}_j$ by a kernel function that transforms the original input space into a higher-dimensional feature space, allowing the SVM to potentially better separate the two classes. After trialling several possible kernel function candidates, the Radial Basis Function (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (4)$$

with $\gamma > 0$, was found to yield the most promising results.

To generalise SVMs to more than two classes, pairwise classification is used. A single multiclass problem is reduced into multiple binary problems by building a classifier for each pair of classes, using only instances from two classes at a time.

2.2 Training

The training system architecture is shown in Fig. 1. Its main components are discussed below.

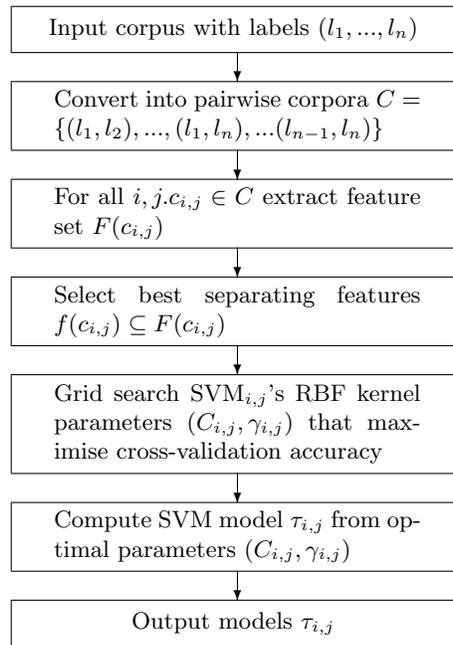


Fig. 1. The training system architecture. $SVM_{i,j}$ represents the support vector for comparing label l_i with l_j .

Feature Extraction For this work, the openSMILE [9] feature extraction algorithms are used. OpenSMILE provides sound recording and playback via the open-source PortAudio library, echo cancellation, windowing functions, fast Fourier transforms and autocorrelation. Moreover, it is capable of extracting features such as pitch, loudness, energy, mel-spectra, voice quality, mel-spectrum frequency coefficients, and can calculate various functionals such as means, extremes, peaks, percentiles and deviations with a Real-Time Factor $\ll 1$.

Feature Selection Since a large feature set will be extracted from the speech, it is expected that there are some irrelevant and redundant data that will not improve the SVM prediction performance. Classification algorithms are unable to attain high classification accuracy if there is a large number of weakly relevant and redundant features, a problem known as the *curse of dimensionality* [10]. Algorithms also suffer from computational load incurred by the high dimensional data.

Our approach is to use the predefined openSMILE set `emo_large` with 6552 features, and pick the most relevant ones using feature selection. For choosing relevant features, the Correlation-based Feature Selection (CFS) algorithm [11] is used. It uses a heuristic based on the assumption that good feature sets contain features highly correlated with the class and uncorrelated with each other.

Grid Search When using the Radial Basis Function SVM kernel, it is important to choose a suitable penalty for mislabelled examples C and the exponentiation constant γ . Because the optimal values are model-specific, a search algorithm is needed for finding a near-optimal set of values.

The goal is to identify good (C, γ) values so that the classifier can accurately predict unseen testing data, rather than choosing them to maximise prediction accuracy for the training data whose labelling is already known. In this work we use v -fold cross-validation. The training set is divided into v equal-sized subsets, with each subset sequentially tested used a classifier trained on the remaining $v - 1$ subsets.

We use a GRID SEARCH algorithm that sequentially tries pairs of (C, γ) in a given range, and picks the one with the highest cross-validation accuracy. Exponentially growing sequences worked well in practice, confirming findings in previous research [12]. The algorithm is run recursively on a shrinking area.

2.3 Classification

The real-time classification system architecture is shown in Fig. 2. Its main components are discussed below.

Segmentation Real-time analysis of speech requires segmenting the audio. Our static threshold algorithm achieves this by defining three thresholds. First, the silence threshold η defines the threshold for the energy $E = \sum_i^n |s_i|^2 > \eta$, for signals s_i in frame of size n . Second, ρ_{start} sets the number of frames with energy above η that are required until a segment start is detected. Third, ρ_{end} defines the number of frames below η until a segment end is detected. After the audio is segmented, openSMILE is used to extract the features.

Pairwise Fusion Mechanism Once the audio is segmented and the features are extracted, $n(n - 1)/2$ pairwise machines for n classes are run in parallel to predict the class for a segment. In order to determine the most probable class, the probabilities of the multiple binary classifiers are fused.

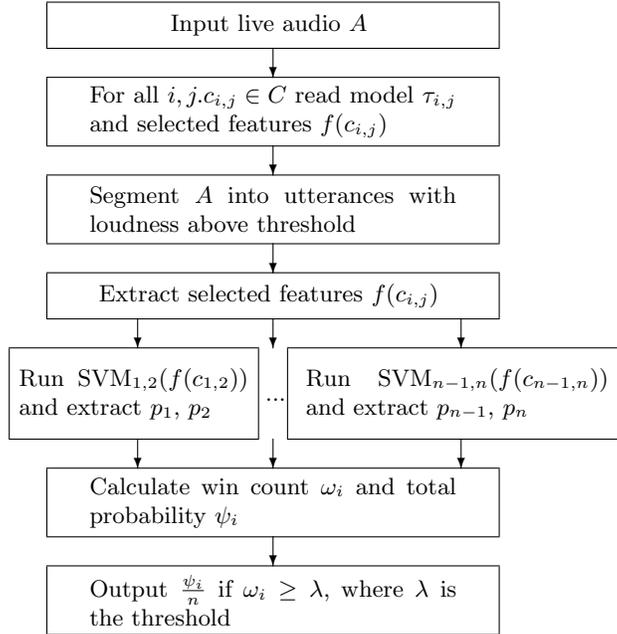


Fig. 2. The real-time classifier architecture. $SVM_{i,j}$ computes the probabilities p_i and p_j for labels i, j , using features $f(c_{i,j})$.

We propose a fusion method for determining co-occurring emotions. Whereas in traditional single-label classification a sample is associated with a single label l_i from a set of disjoint labels L , multi-label classification associates each sample with a set of labels $L' \subseteq L$. A previous study concluded that the use of complex non-linear fusion methods yielded only marginal benefits (0.3%) over linear methods when used with SVMs [13]. Therefore, three linear fusion methods are implemented:

1. Majority voting using wins from binary classifiers.
2. Maximum combined probability from binary classifiers.
3. Binary classification wins above a threshold.

In the first method we consider all $n - 1$ SVM outputs per class as votes and select the class with most votes. Assuming that the classes are mutually exclusive, the *a posteriori* probability for feature vector \mathbf{f} is $p_i = P(\mathbf{f} \in class_i)$. The classifier $SVM_{i,j}$ computes an estimate $\hat{p}_{i,j}$ of the binary decision probability

$$p_{i,j} = P(\mathbf{f} \in class_i | \mathbf{f} \in class_i \cup class_j) \quad (5)$$

between classes i and j . The final classification decision \hat{D}_{voting} is the class i for which

$$\hat{D}_{voting} = \arg \max_{1 \leq i \leq n} \sum_{j \neq i} g(\hat{p}_{i,j}) \quad (6)$$

where

$$g(p) = \begin{cases} 1 & \text{for } p \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

Ties are solved by declaring the class with higher probability to be the winner.

In the second method, the maximum probability $\psi_i = \sum_{p \in \mathcal{S}_i} p$ of the binary SVMs is determined. The winner of decision $\hat{D}_{probability}$ is i such that

$$\hat{D}_{probability} = \arg \max_{1 \leq i \leq n} \sum_{j \neq i} \hat{p}_{i,j}. \quad (8)$$

Finally, for detecting co-occurring emotions, the classes are ranked according to the number of wins. The classes with wins above a threshold λ are returned, with the classification decision $\hat{D}_{threshold}$ being the set of classes

$$\hat{D}_{threshold} = \{i \mid \sum_{j \neq i} g(\hat{p}_{i,j}) \geq \lambda\}. \quad (9)$$

We set $\lambda = \lfloor (\mu + \sigma)n \rfloor$ where μ is the mean win count, σ is the standard deviation and n is the class cardinality to allow comparison with Sobol Shikler [3].

3 Application for Public Speaking Skill Assessment

We present a novel application of the classifier for assessing the quality of public speaking skills.

In persuasive communication, the non-verbal clues a speaker conveys require special attention. Untrained speakers often come across as bland and lifeless. Precisely analysing the voice is difficult for humans and is subjective. By using a similar approach as for detecting emotions, our system enables more objective assessment of public speaking skills.

We retrain our classifier using six labels describing public speaking skills shown in Table 2. Following the requirements by Schuller et al. [14], we use non-acted, non-prompted, realistic data with many speakers, using all obtained data. An experienced speech coach was asked to label 124 one-minute-long samples of natural audio from 31 people attending speech coaching sessions. The chosen six labels are the ones that the professional is accustomed to using when assessing the public speaking skills of clients. The samples are labelled on a scale 4–10 for each class. We then divided the samples of classes into higher and lower halves according to the score. The upper half represents a positive detection of the class (e.g. *clear*), and the lower half represents a negative detection (e.g. *not clear*).

One binary SVM per class is used to derive a class-wise probability. If a pairwise approach similar to that in emotion classification had been used, the

same samples would have existed in several classes, making separating the classes intractable. As a result, unlike in emotion detection where the most prominent labels describing the speech are selected, for speech quality assessment all classes are detected, each labelled with a probability. This allows users to attempt to maximise all class probabilities, a goal which is more useful for speech coaching.

The results of public speaking skill assessment are shown in Table 2. All classes can be accurately detected. The classes *competent* and *dynamic* present slightly lower detection accuracies, perhaps due to the smaller variation in scores resulting from a small corpus size. Overall, however, the speech quality assessment accuracies are high (average 81%) and may provide useful feedback to speakers. In future work, performance using alternative evaluation metrics such as those specified by Schuller et al. [14] will be investigated.

Table 2. Detection accuracies in percentages for assessing public speaking skills.

Class	10-fold cross-validation	Training samples
clear	80	66
competent	74	49
credible	80	42
dynamic	77	45
persuasive	82	79
pleasant	93	73
Mean	81	59

4 Evaluation

In this section we evaluate the overall classification results.

The result of applying grid search is shown in Table 3. The optimisation is done on the training data, with the testing data kept unseen. A significant improvement, between 10% and 25%, was observed. As the optimisation maximises the cross-validation accuracy of the training data instead of the training data classification accuracy, the optimisation did not result in overfitting of the model.

The average latency in milliseconds of the classification stage is shown in Fig. 3. It was measured as the time between the detection of the end of a segment and the output of the result. As shown in the figure, normal sentences (1–15 s) are classified in 0.046–0.110 s, making the delay barely noticeable. Improving upon Sobol Shikler’s inference solution [3], this allows real-time classification.

The ten-fold cross-validation results for the pairwise SVMs are shown in Table 4. All accuracies are greater than the values obtained in previous research

Table 3. Detection accuracies in percentages with a 70–30% training/testing split for the three fusion methods, with and without grid search.

Type of data	Threshold	Max probability	Max wins
Grid search	86	72	70
No grid search	76	47	48

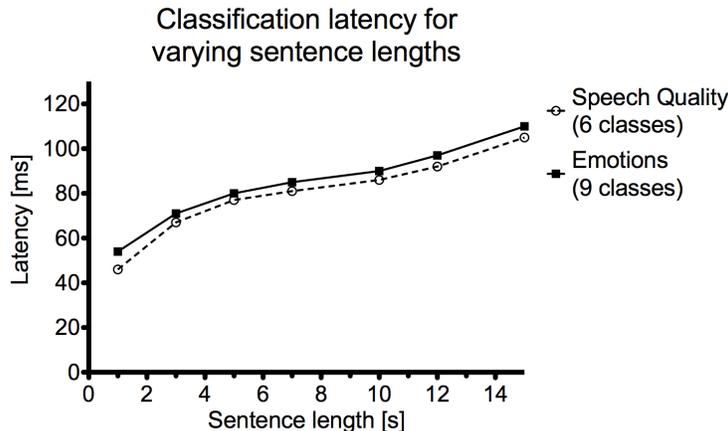


Fig. 3. Average live classification latency in milliseconds on a dual-core 2.66 GHz PC with 4 GB RAM.

using the same classes and corpus. The results are constantly above 80%, in contrast to the lower bound 60% obtained previously.

A summary of the accuracies for the three different fusion methods is shown in Table 5. The average accuracies are higher than or equal to the results achieved previously on the same corpus [3]. Notably, the average accuracy of the maximum probability fusion technique is higher than that achieved by majority voting (72% vs 70%). However, for some classes the majority voting accuracy is higher (e.g. *stressed* and *interested*). A higher average accuracy could be achieved by combining these methods. In future work, more advanced fusion methods such as the ensemble classification presented by Schuller et al. [15] and the tree-based approach by Lee et al. [16] will be investigated.

Confusion matrices for fusion using thresholding and maximum probability are shown in Tables 6 and 7 respectively. Inspection of the confusion matrices reveals that some classes are better detected than others. The classes *opposed* and *sure* present the lowest values using any method. This is reflected by the lower number of training samples (38 and 53 samples, compared to the average of 61) resulting from the categorisation choice to allow comparison to Sobol Shikler [3]. Similarly, the class with most samples (*joyful*, 94 samples) is most frequently mistaken to be the correct class. In future work classes with equal numbers of training samples could be used.

Table 4. The 10-fold cross-validation accuracy for pairwise SVMs in percentages. The average accuracy is 89%. For comparison, Sobol Shikler’s results [3] are in parentheses.

	excited	interested	joyful	opposed	stressed	sure	thinking	unsure
absorbed	93 (81)	87 (82)	96 (82)	96 (78)	89 (87)	85 (84)	82 (73)	84 (64)
excited		90 (71)	84 (60)	81 (71)	80 (61)	94 (83)	90 (72)	87 (75)
interested			92 (77)	92 (75)	91 (66)	90 (78)	90 (84)	85 (72)
joyful				86 (71)	85 (61)	99 (83)	95 (72)	92 (75)
opposed					93 (84)	91 (72)	94 (81)	92 (79)
stressed						86 (84)	88 (75)	86 (78)
sure							94 (75)	88 (78)
thinking								90 (89)

Table 5. Accuracies in percentages for the three fusion methods. Sobol Shikler’s results [3] are shown in parentheses. 2.5 classes were inferred on average with a threshold $\lambda = 6$.

Type of data	Threshold	Max probability	Voting
70–30% training/ testing split	86 (79)	72	70
Training data	99 (81)	86	88

As expected, the thresholding fusion method for co-occurring emotion classification yields highest detection accuracies since several classes can be selected at a time. This, however, also leads to much higher confusion values because of the assumption that more than one emotion can be occurring simultaneously. For example, as shown in Table 7, samples labelled *excited* are detected as *joyful* in 35% of cases, compared to a correct detection rate of 85%. It is likely that some high confusion rates are caused by the overrepresentation of certain classes.

5 Conclusion

We have presented a framework for real-time speech emotion classification whose accuracy outperforms previous work using the same corpus [3]. We have also shown that the novel application of the system for assessing public speaking skills achieves high classification accuracies.

The framework consists of $n(n-1)/2$ pairwise SVMs for n labels, each with a differing set of features selected by a correlation-based feature selection algorithm. We demonstrated a considerable improvement in classification accuracy from optimising the misclassification and exponentiation coefficients (C, γ) in

Table 6. Confusion matrix using maximum probability for pairwise fusion. The column headings show the ground truth and the rows show inferences. Average accuracy is 72%. A random choice would result in 11% accuracy.

	absorbed	excited	interested	joyful	opposed	stressed	sure	thinking	unsure
absorbed	74	0	2	0	0	1	2	1	1
excited	0	75	2	6	0	2	6	0	1
interested	4	0	69	0	0	2	2	3	1
joyful	4	10	6	79	16	11	4	3	4
opposed	0	2	0	2	62	1	2	0	0
stressed	4	8	6	3	8	67	9	1	8
sure	0	0	2	2	5	2	63	0	0
thinking	7	0	8	3	0	4	11	86	17
unsure	7	4	6	4	8	8	2	6	68

Table 7. Confusion matrix using thresholding for pairwise fusion. The column headings show the ground truth and the rows show inferences. Average accuracy is 86%. A random choice would result in 11% accuracy.

	absorbed	excited	interested	joyful	opposed	stressed	sure	thinking	unsure
absorbed	93	4	15	0	0	4	12	23	24
excited	15	85	10	29	27	46	24	6	14
interested	22	2	83	14	3	10	11	17	14
joyful	15	35	21	91	41	39	22	23	22
opposed	0	14	6	22	73	11	17	7	8
stressed	15	60	31	56	51	92	31	24	29
sure	11	19	6	4	16	9	74	11	9
thinking	48	15	42	19	24	19	28	93	56
unsure	48	8	52	24	22	31	26	56	91

(1) and (4) using a grid search algorithm. Improvements between 10% and 25% were observed.

Overall, this paper presented a high-accuracy training and classification framework for emotion detection from speech, and shows that it can be successfully applied for real-time assessment of public speaking skills.

References

1. Scherer, K.R.: Vocal affect expression: A review and a model for future research. *Psychological bulletin* **99** (1986) 143–165
2. Haynes, J.D., Rees, G.: Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* **7** (2006) 523–534
3. Sobol Shikler, T.: Analysis of affective expression in speech. PhD thesis, Cambridge University (2007)
4. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.J. In: *Mind Reading: The Interactive Guide to Emotions*. Jessica Kingsley Publishers, University of Cambridge (2004) ISBN 1 84310 214 5.
5. Golan, O., Baron-Cohen, S., Wheelwright, S., Hill, J.J.: Systemizing empathy: Teaching adults with asperger syndrome and high functioning autism to recognize complex emotions using interactive multimedia. *Development and Psychopathology* **18** (2006) 589–615
6. Sobol Shikler, T., Robinson, P.: Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **99** (2009)
7. Hall, M.A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. *SIGKDD Explorations* **11** (2009)
8. Vapnik, V.N.: *The nature of statistical learning theory*. Springer (1998)
9. Eyben, F., Wöllmer, M., Schuller, B.: openEAR – Introducing the Munich open-source emotion and affect recognition toolkit. In: *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, IEEE, Amsterdam, The Netherlands (September 2009)
10. Altun, H., Polat, G.: New frameworks to boost feature selection algorithms in emotion detection for improved human-computer interaction. In: *Advances in Brain, Vision, and Artificial Intelligence*. Volume 4729 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg (2007) 533–541
11. Hall, M.A., Smith, L.A.: Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. *Florida Artificial Intelligence Symposium* (1999) 235–239
12. Qing-kun, L., Pei-wen, Q.: Model selection for SVM using mutative scale chaos optimization algorithm. *Journal of Shanghai University* **10** (2006) 531–534
13. Pöyhönen, S., Arkkio, A., Jover, P., Hyötyniemi, H.: Coupling pairwise support vector machines for fault classification. *Control Engineering Practice* **13** (2005) 759–769
14. Schuller, B., Steidl, S., Batliner, A.: The Interspeech 2009 emotion challenge. In: *Interspeech*, Brighton, UK (2009)
15. Schuller, B., Reiter, S., Müller, R., Al-Hames, M., Lang, M., Rigoll, G.: Speaker independent speech emotion recognition by ensemble classification. In: *IEEE International Conference on Multimedia and Expo*. (2005)
16. Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. In: *Interspeech*, Brighton, UK (2009)