Benchmarking classification models for emotion recognition in natural speech: a multi-corporal study

Alexey Tarasov and Sarah Jane Delany

***Abstract*— A significant amount of the research on automatic emotion recognition from speech focuses on acted speech that is produced by professional actors. This approach often leads to overoptimistic results as the recognition of emotion in real-life conditions is more challenging due the propensity of mixed and less intense emotions in natural speech. The paper presents an empirical study of the most widely used classifiers in the domain of emotion recognition from speech, across multiple non-acted emotional speech corpora. The results indicate that Support Vector Machines have the best performance and that they along with Multi-Layer Perceptron networks and k-nearest neighbour classifiers perform significantly better (using the appropriate statistical tests) than decision trees, Naïve Bayes classifiers and Radial Basis Function networks.**

I. INTRODUCTION

The aim of this paper is to compare a number of supervised learning algorithms on the task of emotion recognition from speech. The training of supervised learning techniques requires using training data that is representative of the classification problem in question. The focus in this paper is on non-acted emotions which is important as emotion recognition systems for real-life applications require training on instances where the emotion expressed in the speech is not acted [7]. In addition, the complexity of the task of automatic emotion recognition from speech increases with the naturalness of the assets—the recognition of natural emotions is much more challenging than that of acted ones [39].

Numerous classification techniques have been used in the current research in this area, our focus in this paper is a comprehensive evaluation of the more widely used algorithms across a number of datasets that have been generated from non-acted speech corpora.

The structure of the paper is as follows. The next section presents a brief overview of the process involved in using supervised machine learning for the recognition of emotion from speech. Section III then describes the datasets that were generated and used in this work, while section IV provides details of the classifiers used. Section V outlines the experimental methodology and section VI discusses the evaluation results. The paper concludes in section VII with directions for future work.

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253

A. Tarasov and S.J. Delany are with Digital Media Centre, Dublin Institute of Technology, Aungier St., Dublin 2, Ireland aleksejs.tarasovs@student.dit.ie, sarahjane.delany@dit.ie

II. CLASSIFYING EMOTION IN SPEECH

Automatic recognition of emotion from speech is a supervised machine learning problem that requires a training set—in this case a collection of emotional speech recordings. Each recording or sample in the dataset has to be labelled with the emotion expressed by the speaker, normally manually by experts, and represented as an n-dimensional vector of predictive characteristics or features such as pitch values or spectrum coefficients which are extracted from the sample. The training set is then presented to a learning algorithm that produces a classifier. Once trained on a representative set of recordings a good classifier will be able to predict correct labels for samples that were not present in the training set.

Speech corpora that are used in emotion recognition from speech can be separated into three groups: natural speech (using data from call-centers [10], talk-shows [18] and similar sources where people exhibit natural emotions), acted speech (where actors are asked to portray specific emotions [6]) and elicited speech (subjects are placed in a controlled environment [3] and emotions are induced by changing their environment). Acted speech should not be used in the training of a classifier for spontaneous, real-life emotions [7] and for this reason this study concentrates only on natural or elicited corpora.

There are two main approaches to labelling emotional assets—using discrete categorical labels and using emotional dimensions. The first uses a limited set of distinct emotions, for example, one of most widely used sets consists of joy, sadness, fear, anger, surprise and disgust [13] (often called "the big six"), and often used with an addition of a special category for neutral, non-emotional speech. Discrete labelling has some drawbacks—emotions in everyday speech are usually weak and mixed rather than pure [8], so a significant part of the collected assets can be categorised as neutral. With emotional dimensions the emotion expressed in the speech is not represented as a discrete label, but as a point in multi-dimensional space. The three most frequently used dimensions are activation (how active or passive the emotion is), evaluation (whether it is negative or positive) and power (does the speaker feel powerful or weak) [29]. Due to the subjectivity of emotion many state-of-the-art corpora are rated by a number of labellers and the label is determined from the various ratings given by the labellers. In some cases a measure of agreement or confidence can also be given—for example, in the case of dimensional annotation an average of the labeller's rating values can be used as a consensus label and the standard deviation of ratings can be used as an

agreement measure [18].

There is no agreed feature set for emotion recognition from speech, but most of the research uses a combination of prosodic and spectral features. It has been stated that prosodic features are especially useful in the case of acted speech, but spectral features are valuable when recognising emotion from natural or elicited speech [38]. Some researchers also state the importance of lexical and contextual features [11], but their extraction usually requires the additional effort of annotating the speech corpus.

The most widely used classification techniques for emotion recognition from speech are k-nearest-neighbour methods (k-NN), C4.5 decision trees, support vector machines (SVMs), artificial neural networks (ANNs), and Naïve Bayes (NB) classifiers. These techniques have often been compared on emotional speech assets [4], [23], [28], [33], [37], [40], but not on more than two natural/elicited datasets or for statistically significant differences. Each of these classifiers was used in 24% of 84 papers we have surveyed in average.

III. DATASETS USED

Table I presents an overview of the datasets used in this study which were derived from natural emotional speech corpora that used both categorical and dimensional labels. Natural and elicited emotional datasets are difficult to obtain and the ones used in this study are quite diverse—the recorded subjects speak in different languages, a variety of target classes was used and the quality of recordings also differs. The following categorical-rated corpora were used:

1) **FAU Aibo Emotion Corpus**—this corpus [3], [36] contains recordings of children speaking to the AIBO robot controlled by a human invisible to them. All the 18216 assets were used in the Interspeech 2009 Challenge as a five-class classification problem (angry, emphatic, neutral, positive, rest) and the datasets extracted from this corpus used this labelling scheme. Assets with a high confidence level were selected where possible (above 0.5 on a 0 to 1 scale[1]).
   Four datasets were extracted from this corpus—300 confident neutral assets and 300 angry assets were randomly selected and formed the first dataset labelled as AIBO-NA. The second and third datasets, labelled AIBO-NE and AIBO-NP, were generated in the same way but contained 300 emphatic assets and 300 positive assets respectively, in place of the angry assets. The fourth dataset, AIBO-AENPR involved selecting 200 random assets from each of the five categories. In all cases, except Rest where there were few assets with high confidence, these were confident assets.
   An additional dataset was derived from the 4-class corpus proposed by the HUMAINE Network of Excellence CEICES initiative that contained 4513 assets from the AIBO corpus. The same procedure of selecting assets was applied, creating a dataset containing 200 confident assets from each class.
2) **BabyEars**—this corpus contains recordings of parents speaking to their children [34]. All instances belong to one of three classes—approval (when the action of a child is approved), attention (when the parent attracts the attention of the child) and prohibition (when parents prohibit some actions).

Also two dimensionally-rated corpora were used in this study. Since the classification task is being considered, there is a need to discretise the numerical values of the dimensions into categories. Assets were sorted within a particular dimension and then separated into three classes in such a way that each discrete class contains approximately the same number of instances. Table II provides the number of instances in each class created in this way and the range of values assigned to each class. The following describes the datasets generated from the dimensionally-rated corpora using this approach:

1) **Vera am Mittag German Audio-Visual Emotional Speech Database** is a natural corpus of German talk-show recordings [18] which contains assets rated on three dimensional scales on a scale of -1 to +1. The dimensions were activation, valence (a synonym for evaluation) and dominance (a synonym for power). A single dataset of three classes was generated from the ratings across each of these dimensions, VAM-ACT, VAM-EVAL and VAM-POW.
2) **Utsunomiya University Spoken Dialogue Database For Paralinguistic Information Studies**[2]—a Japanese elicited corpus that contains 4840 assets labelled across six dimensions (pleasantness, arousal, dominance, credibility, interest and positivity) on a scale of 1 to 7. Each asset in this corpus has rating values supplied by three experts and the mean of these values is used as the target rating for each asset. For the purposes of dataset generation for this study a measure of confidence was assigned to each asset where confidence was measured as the difference between the minimal and maximal rating given to each asset.
   Only the ratings from the dimensions of arousal (activation), pleasantness (evaluation) and dominance (power) were used. For each dimension the assets were discretized into three classes in the manner described above. Where classes contained more than 300 confident assets, a random selection was chosen to generate the datasets labelled UUDB-ACT, UUDB-EVAL and UUDB-POW reflecting the arousal, pleasantness and dominance dimensions.

For the experiments a feature set of 384 acoustical and spectral features was extracted using openEAR software[3]. This was the feature set used in the Interspeech 2009 challenge which consists of different features based on pitch,

[1]The AIBO corpus was labelled at the word level, and these labels were used to get a label for the whole phrase. The confidence level for the utterance denotes the proportion of words from that utterance that have the same label as the utterance itself.

[2]http://uudb.speech-lab.org/

[3]http://openart.sourceforge.net/

TABLE I

DATASETS USED

Name	Language	Description of classes	Number of instances	Number of classes	Data distribution, %
AIBO-AENPR	German	Anger, empathy, neutral, positive, rest	1000	5	20/20/20/20/20
AIBO-AMEN	German	Anger, motherese, empathy, neutral	800	4	25/25/25/25
AIBO-NA	German	Neutral, anger	600	2	50/50
AIBO-NE	German	Neutral, empathic	600	2	50/50
AIBO-NP	German	Neutral, positive	600	2	50/50
BabyEars	English	Attention, approval, prohibition	509	3	42/29/29
UUDB-ACT	Japanese	Three levels of activation	900	3	33/33/33
UUDB-EVAL	Japanese	Three levels of valence	900	3	33/33/33
UUDB-POW	Japanese	Three levels of dominance	900	3	33/33/33
VAM-ACT	German	Three levels of activation	947	3	33/33/33
VAM-EVAL	German	Three levels of valence	947	3	33/33/33
VAM-POW	German	Three levels of dominance	947	3	33/33/33

TABLE II

DISCRETISATION OF THE DIMENSIONAL DATASETS

Dataset	Lower class		Middle class		Higher class	
	Number of instances	Range	Number of instances	Range	Number of instances	Range
UUDB-ACT	1496	[1; 3.6667]	1619	[4; 4.6667]	1725	[5; 7]
UUDB-EVAL	1847	[1; 3.6667]	1753	[4; 4.3333]	1240	[4.6667; 7]
UUDB-POW	1484	[1; 3]	1564	[3.3333; 4.6667]	1792	[5; 7]
VAM-ACT	317	[-1; -0.1625]	315	[-0.1586; 0.1169]	315	[0.1172; 1]
VAM-EVAL	317	[-1; -0.293]	315	[-0.2892; -0.163]	315	[-0.1623; 1]
VAM-POW	317	[-1; -0.0562]	315	[-0.0546; 0.1852]	315	[0.1853; 1]

energy, zero-crossing rate, harmonics to noise ratio as well as of 12 mel-frequency cepstral coefficients. All features were normalised to the interval [0; 1] with the only exception in the case of ANNs where the interval [-1; 1] was used as it is best-practice for the back-propagation algorithm allowing it to converge faster [22].

IV. CLASSIFIERS

The most widely used classifiers for emotion recognition from speech are k-NN, NB, decision trees, SVMs and ANNs. This section will give a brief description of each of these supervised learning techniques.

A. k-NN

The k-NN classifier compares a given target instance with the k training instances that are the most similar or closest to it [20]. There are a variety of metrics used to measure similarity and the Euclidian distance metric is frequently used. The target instance is assigned to the class to which the majority of these nearest neighbours belongs.

There is no consensus on which value of k should be used in the case of emotion recognition from speech. Different researchers proposes values from $k = 1$ [23], [30] to $k = 20$ [14]. Usually the value of k is found by trial and error, different values are taken and the performance of these classifiers is compared [16], [17].

B. NB

The NB classifier is a technique that uses Bayes theorem to predict class membership probabilities [20]. It finds the class E_j^* $(j = 1, 2, ..., l)$ to which the target instance $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is assigned in the following way:

$$E_j^* = \arg\max_j = P(E_j)P(\mathbf{x}|E_j) \tag{1}$$

where the prior probabilities $P(E_j)$ are estimated from the occurrences of instances belonging to category E_j in the training set, and the likelihoods $P(\mathbf{x}|E_j)$ in the case of a so-called flexible Naïve Bayes classifier [24] which is used in this research are calculated as a sum of Gaussians $g(x_i, \mu_j, \sigma_j)$ as follows:

$$P(\mathbf{x}|E_j) = \frac{1}{n}\sum_{i=1}^{n} g(x_i, \mu_j, \sigma_j).$$

C. Decision trees

A decision tree is a hierarchical data structure which is the result of a recursive binary partitioning of the training set [1]. The classification process involves traversal of the tree following a path to a specific leaf node according to the decision criteria at each node. The leaf nodes represent the classifications. The most commonly used algorithm for constructing decision trees in emotion recognition from speech is C4.5 [11], [23], [28], [30], [37].

D. SVMs

An SVM is a binary classifier which uses a nonlinear mapping to transform the original training data into a higher dimension and within this new dimension it searches for the linear optimal separating hyperplane [20]. The function that performs this mapping is the kernel function. The most

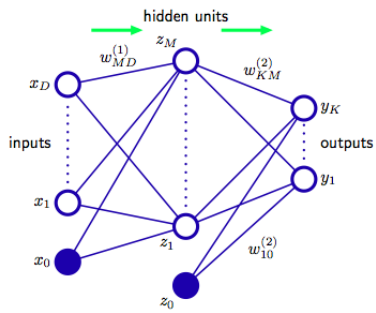

Fig. 1. The structure of an MLP [5]

frequently used SVM kernel function in the domain of emotion recognition in speech is the radial basis function (RBF) kernel [11], [28], [35], [42], although some researchers also use the polynomial kernel [23], [28], [31] or the linear kernel [2], [26], [41]. In this study we use the linear and RBF kernels, excluding the polynomial kernel because of its slow training speed.

To use SVMs for a multi-class problem an ensemble of SVM classifiers is needed. A common implementation of SVMs for multi-class problems is the round-robin ensemble which uses $l \times (l-1)/2$ base SVM classifiers, where l is the number of target classes and each base classifier is trained on the training instances from a different pair of target classes.

E. ANNs

An ANN is a model consisting of interconnected processing units known as artificial neurons, grouped into layers, that takes its inspiration from the brain [1]. The classification process involves a spreading activation through each layer of the network using an activation function to calculate the output from each node.

When authors refer to ANNs they usually mean multi-layered perceptrons (MLPs), though the same term can refer to radial basis function (RBF) networks, self-organizing maps and many other related technologies. A typical MLP is shown on Figure 1. The neurons are grouped into three layers: input, hidden and output.

An RBF network has the same structure, but there are restrictions on the activation functions that can be used. MLPs can use a broad selection of activation functions, in this study a hyperbolic tangent is used as the hidden and output layer activation functions.

V. METHODOLOGY

Each classifier was evaluated on each dataset using 5-fold cross validation where the training set was split into five folds and each fold in turn was held out as a test set with the remaining folds being used for training. At each fold iteration the parameters of each classifier (except NB which does not have any parameters) were tuned on the training set using an additional 5-fold cross validation. The ranges used to tune the parameters are detailed in Table III. The performance measure used was average class accuracy as a number of the datasets are imbalanced. This overall process was repeated three times and averaged classification accuracies for each classifier on each dataset are calculated.

The appropriate tests—Friedman [15] and Bonferonni-Dunn [12]—proposed by Demsar [9] for comparing multiple classifiers across multiple datasets were used to test for statistically significant differences in performance. Let r_i^j be the rank of the j-th algorithm ($j = 1, 2, ..., k$) on i-th dataset ($i = 1, 2, ..., N$), thus $R_j = \frac{1}{N}\sum_i r_i^j$ is the average rank of the j-th algorithm. The Friedman statistic

$$\chi_F^2 = \frac{12N}{k\,(k+1)}\left[\sum_j R_j^2 - \frac{k\,(k+1)^2}{4}\right] \qquad (2)$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom when $N > 10$ and $k > 5$ [9]. The null-hypothesis for this test is that all ranks are equal and thus the performance of the classifiers is the same. The test determines whether there are any statistically significant differences between the performance of the classifiers. If the null-hypothesis is rejected, special post-hoc tests are required to discover the classifiers with a statistically different performance; the Bonferonni-Dunn test is used for this purpose.

The MATLAB software package[4] was used to implement MLPs, the LibSVM library[5] was used for SVMs and for all other classifiers WEKA [19] was used.

VI. RESULTS & DISCUSSION

Table IV presents the average classification accuracies for each classifier on each dataset and includes the average rank of classifiers which according to Demsar [9] provides 'a fair comparison of algorithms'. Friedman's test rejects the null hypothesis which indicates that there is a statistically significant difference across the classifiers. The application of the post-hoc Bonferroni-Dunn test shows that the performance of SVM-Linear, k-NN and MLP is not significantly different from SVM-RBF which is the best classifier. The performance of RBF and NB is also found to be comparable to the performance of the worst classifier, the C4.5 decision tree, thus, there are two distinct groups of classifiers. It can therefore be concluded that SVMs, k-NN and MLP are the best classifiers for the given features and datasets, with NB, RBF and decision trees performing significantly worse.

From the results it can be seen that classifying evaluation is more difficult than classifying activation or power—the results of classification on datasets UUDB-EVAL and VAM-EVAL are considerably worse than the results on the other UUDB or VAM based datasets. This may be explained by the fact that acoustical and spectral features have been used but no correlation between these types of features and the evaluation dimension has been found [27]. In contrast, pitch has been shown to highly correlate with activation [27], which may explain why the classifiers in general have a relatively high accuracy on activation datasets.

[4] http://www.mathworks.com/products/matlab/

[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm/, implementation details are available in the manual (http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf)

TABLE III

THE RANGES OF VALUES USED IN PARAMETER TUNING.

Classifier	Parameter	Ranges Tested
C4.5	Confidence threshold for pruning	$[0.05, 0.1, ..., 0.5]$
	Minimum number of instances per leaf	$[1, 2, 3, 4]$
k-NN	Number of nearest neighbours	$[1, 3, ..., 41]$
MLP	Number of hidden neurons	$[100, 200, 300]$
RBF	Number of hidden neurons	$[2, 3, ..., 10]$
SVM-LIN	Cost parameter, C	$[2^{-5}, 2^{-4}, ..., 2^{15}]$
SVM-RBF	Cost parameter, C	$[2^{-5}, 2^{-4}, ..., 2^{15}]$
	Kernel parameter, γ	$[2^{-15}, 2^{-14}, ..., 2^{3}]$

TABLE IV

RESULTS OF THE COMPARISON OF CLASSIFIERS—ACCURACIES (A) AND RANKS (R), PERCENTAGES

Dataset	SVM-Linear		SVM-RBF		C4.5		k-NN		RBF		NB		MLP	
	A	R	A	R	A	R	A	R	A	R	A	R	A	R
AIBO-AENPR	49.00	2	50.20	1	34.40	7	45.50	4	41.10	6	42.30	5	45.89	3
AIBO-AMEN	63.98	2	66.46	1	46.67	7	61.00	3	54.67	5	54.00	6	60.50	4
AIBO-NA	81.00	1	80.50	2	71.44	6	73.83	4	72.61	5	60.67	7	78.60	3
AIBO-NE	79.33	3	80.11	1	73.17	5	77.17	4	71.28	6	64.50	7	79.51	2
AIBO-NP	76.53	2	77.54	1	68.46	7	74.61	3	72.93	6	73.38	5	74.36	4
BabyEars	70.55	2	77.56	1	55.56	7	64.27	4	56.84	5	56.80	6	68.91	3
UUDB-ACT	75.00	2	75.37	1	69.59	5	69.67	4	68.48	6	66.89	7	73.36	3
UUDB-EVAL	60.74	1	59.37	2	51.15	6	53.30	4	52.44	5	47.89	7	55.87	3
UUDB-POW	75.11	1	74.19	2	66.04	7	71.18	3	69.63	6	71.00	4	70.56	5
VAM-ACT	62.80	1	61.96	2	52.62	7	55.77	5	53.02	6	56.69	4	57.78	3
VAM-EVAL	45.64	4	49.40	1	40.98	7	46.59	2	43.14	6	46.13	3	45.07	5
VAM-POW	56.81	4	70.08	1	54.10	5	65.21	3	50.57	6	46.92	7	65.33	2
Averaged rank	2.08		1.33		6.33		3.58		5.67		5.67		3.33	

The VAM corpus is the only natural corpus used in this study and the quality of the recordings is lower than that of other corpora, which may account for the notably lower classification accuracy on the VAM-ACT, VAM-DOM and VAM-VAL datasets compared with the UUDB datasets.

The performance on the datasets based on AIBO corpus shows that the difficulty of the classification task grows appreciably with the number of target classes. The best result for binary problems typically is around 80%, dropping to 60% and 50% in the case of four and five-class problems respectively. This is to be expected as multi-class classification is a more complex problem that binary classification.

The discretisation of the continuous dimensions in the UUDB and VAM datasets results in the middle class being narrower in range than the lower or higher one (see Table II). It aligns well with the fact that elicited corpora have mostly neutral or assets that can be considered not emotive enough [25], [36], [38], but at the same time this suggests a need for more sophisticated ways of performing the discretisation.

VII. CONCLUSIONS

In this paper an empirical study of the classifiers that are most widely used in the field of emotion recognition has been carried out. The performance of SVMs, MLPs, RBFs, NB, k-NN, C4.5 decision trees (with appropriate tuning of parameters) was compared across twelve datasets extracted from four corpora containing natural or elicited speech. Using Friedman and Bonferroni-Dunn statistical tests, the performance of SVMs with both linear and RBF kernels, MLPs and k-NN classifiers was found to be superior to the other techniques with SVMs performing the best overall.

Future work will consider extending the study to include some of the state-of-the-art classification algorithms including Random Forests [32], Gaussian Mixture Models [5] and Kernel methods [21] and to include speech assets now available from the SEMAINE project. Ensemble techniques have also been widely used in the field of emotion recognition and are often compared to single classifiers [4], [23], [28], [33], [37], [40]. Our future work will consider their comparison with single classifiers and determining the most suited architecture for an ensemble that performs emotional speech recognition. Finally due to the difficulty in getting highly emotive speech assets from natural and elicited speech we will investigate alternative ways of discretising continuous emotional dimensions and investigate feature selection techniques to determine a better representation for the data.

VIII. ACKNOWLEDGMENTS

The authors would like to thank Dr Malcolm Slaney for providing the BabyEars corpus.

REFERENCES

[1] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2004.

[2] A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In *Proc. of InterSpeech-2007*, pages 2253—2256, 2007.

[3] A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private Emotions vs. Social Interaction: a Data-driven Approach towards Analysing Emotion in Speech. *User Modeling and User-Adapted Interaction*, 18(1–2):175—206, 2008.

[4] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proc. of IS-LTC 2006*, pages 240—245, 2006.

[5] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Proc. of Ninth European Conference on Speech Communication and Technology*, pages 1517—1520, 2005.

[7] Z. Callejas and R. Lopezcozar. Influence of Contextual Information in Emotion Annotation for Spoken Dialogue Systems. *Speech Communication*, 50(5):416—433, 2008.

[8] R.R. Cornelius. *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Prentice-Hall, 1996.

[9] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[10] L. Devillers and L. Vidrascu. Real-Life Emotion Recognition in Speech. *Speaker Classification II, LNCS*, 4441:34—42, 2007.

[11] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection. *Neural Networks*, 18(4):407—422, 2005.

[12] O.J. Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56:52—64, 1961.

[13] P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, K.R. Scherer, M. Tomita, and A. Tzavaras. Universals and Cultural Differences in the Judgements of Facial Expressions of Emotion. *Journal of Personality and Social Psychology*, 53(4):712—717, 1987.

[14] E. Fersini, E. Messina, G. Arosio, and F. Archetti. Audio-Based Emotion Recognition in Judicial Domain: A Multilayer Support Vector Machines Approach. *MLDM 2009: Machine Learning and Data Mining in Pattern Recognition, LNCS*, 5632:594—602, 2009.

[15] M. Friedman. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32:675—701, 1937.

[16] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-Based Evaluation and Estimation of Emotions in Speech. *Speech Communication*, 49(10–11):787—800, 2007.

[17] M. Grimm, K. Kroschel, and S. Narayanan. Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech. *Proc. of ICASSP*, pages 3—6, 2007.

[18] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, 2008.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[20] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2nd edition, 2006.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[22] S Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1994.

[23] I. Iriondo, S. Planet, J.C. Socoro, and F. Alias. Objective and Subjective Evaluation of an Expressive Speech Corpus. *Advances in Nonlinear Speech Processing, LNCS*, 4885:86, 2007.

[24] G.H. John and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *Proc. of the Eleventh Conference on Uncertainty in Artificial Intelligence*, volume 1, pages 338—345, 1995.

[25] D. Künster, R. Tato, T. Kemp, and B. Meffert. Towards Real Life Applications in Emotion Recognition. *Affective Dialogue Systems, LNAI*, 3068:25—35, 2004.

[26] C.C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion Recognition Using a Hierarchical Binary Decision Tree Approach. In *Proc. of InterSpeech-2009*, pages 320—323, 2009.

[27] I.B. Mauss and M.D. Robinson. Measures of Emotion: A Review. *Cognition & Emotion*, 23(2):209—237, 2009.

[28] D. Morrison and L.C. De Silva. Voting Ensembles for Spoken Affect Classification. *Journal of Network and Computer Applications*, 30(4):1356—1365, 2007.

[29] M. Schröder. Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-Extreme Emotions. In *Proc. of Workshop on Affective Dialogue Systems*, pages 209—220, 2004.

[30] B. Schuller, R. Muller, M. Lang, and G. Rigoll. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In *Proc. of InterSpeech-2005*, pages 805—808, 2005.

[31] B. Schuller, S. Reiter, and G. Rigoll. Evolutionary Feature Generation in Speech Emotion Recognition. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 06)*, pages 5—8, 2006.

[32] G. Seni and J.F. Elder. *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions*. Morgan & Claypool, 2010.

[33] M. Shami and W. Verhelst. Automatic Classification of Expressiveness in Speech: A Multi-Corpus Study. *Speaker Classification II, LNCS*, 4441:43—56, 2007.

[34] M. Slaney and G. McRoberts. Baby Ears: a Recognition System for Affective Vocalizations. *Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 985—988, 1998.

[35] P. Staroniewicz. Recognition of Emotional State in Polish Speech—Comparison between Human and Automatic Efficiency. *Biometric ID Management and Multimodal Communication Joint COST 2101 and 2102 International Conference, Proceedings, LNCS*, pages 33—40, 2009.

[36] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, 2009.

[37] L. Vidrascu and L. Devillers. Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center. In *Proc. of 2005 IEEE International Conference on Multimedia and Expo*, pages 944—947, 2005.

[38] T. Vogt and E. André. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *Proc. of ICME 05*, 2005.

[39] T. Vogt and J. Wagner. Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation. *Affect and Emotion in Human-Computer Interaction, LNCS*, 4868:75—91, 2008.

[40] S. Yilmazyildiz, W. Mattheyses, Y. Patsis, and W. Verhelst. Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication. *PCM 2006, LNCS*, 4261:1—8, 2006.

[41] M. You, G.Z. Li, L. Chen, and J. Tao. A Novel Classifier Based on Enhanced Lipschitz Embedding for Speech Emotion Recognition. *Proceedings of the 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, LNCS*, pages 482—490, 2008.

[42] Shiqing Zhang. Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features. *Advances in Neural Networks—ISNN 2008, LNCS*, pages 457—464, 2008.