ANALYSIS OF ROBUSTNESS OF ATTRIBUTES SELECTION APPLIED TO SPEECH EMOTION RECOGNITION

S. Casale and A. Russo

Dipartimento di Ingegneria Informatica e delle Telecomunicazioni, Universita' di Catania
Viale A. Doria, 6, 95125, Catania, Italy
phone: + (39) 095 738 2368, fax: + (39) 095 738 2397
email: (scasale,arusso)@diit.unict.it

S. Serrano

Dipartimento di Fisica della Materia e Ingegneria Elettronica, Universita' di Messina
Contrada Di Dio (S. Agata), 98166, Messina, Italy
phone: + (39) 090 397 7522, fax: + (39) 090 391 382
email: sserrano@ingegneria.unime.it

ABSTRACT

The paper presents the analysis of the robustness of an attributes selection method applied to speech emotion recognition. The features used were extracted by the front-end ETSI Aurora eXtended of a mobile terminal in compliance with the ETSI ES 202 211 V1.1.1 standard. On the basis of the time trend of these parameters, over 3700 statistical attributes were extracted to characterize semantic units of varying length (sentences, words and generic chunks). Using the WEKA (Waikato Environment for Knowledge Analysis) software the most significant attributes for the classification of two emotional states were selected using the *CFSSubsetEval-BestFirst* method. The results of classification, obtained using *NaiveBayes* models, were obtained using intra-corpus and inter-corpora experiments on four different speech corpora performing 4000 trainings and tests. On the basis of these results we can study the robustness of the attributes selection method.

1. INTRODUCTION

One important research challenge in the last few years has been automatic recognition of the emotional state of a speaker through speech. This could be especially important in situations where speech is the primary communication tool with the machine. There are various applications for a system capable of recognizing emotional states via speech in a range of fields including psychiatric diagnosis, the toy industry, Customer Relationship Management (CRM), home jukeboxes, Automatic Speech Recognition (ASR), Speech Synthesis, automatic learning, alarms and voicemail systems. Murray [17] summarized the relationship between emotion and acoustic features like pitch, intensity, rate and voice quality. Later researchers added formants, LPC and MFCC to combine phonetic and prosodic features together in emotion recognition. For a fixed-length feature vector, researchers computed derived features and statistics like range, mean and standard deviation [20][18]. In this work the features were extracted according to the specifications of the speech recognitions front-end algorithm of the ETSI ES 202 211 V1.1.1 standard [1]. On the basis of the acoustic parameters, over 3700 statistical attributes were extracted for each semantic unit, as described in section 2. But such large number of features is not suitable for classification. Because the accuracy rate will not increase along with the feature number and the generalization of the classifier will decrease while in high dimension space, feature selection is necessary to achieve high recognition performance [20, 14, 12, 6, 7, 5]. In this paper, of the attribute selection techniques provided by WEKA [13], we used *CFSSubsetEval* and to determine the best subset we used the *BestFirst* search strategy.

The goal of an automatic emotion recognizer is to assign category labels that identify emotional states. Numerous studies have been seen in the last years trying to improve on features and classifiers [15, 4, 16, 9]. Unfortunately there is a lack of a definitive description and agreement on a set of basic emotions. In this study we analyzed the robustness of the attributes selected to automatic classification of emotion in speech. So we have limited the classification only into two classes: EMO (grouping all the non neutral emotion states) and IDLe (consisting of neutral or neutral-like emotion states). Moreover, in order to evaluate the robustness over different languages, length of semantic unit and spontaneous or simulated emotion condition, four databases were used: the FAU AIBO Emotion Corpus (AIBO), the Berlin Emotional Speech (EMO-DB), the Speech Under Simulated and Actual Stress (SUSAS) and the Vera-Am-Mittag (VAM) Database. The paper is organized as follows. Section 2 provides the details on the feature extraction method used. Section 3 gives an overview of the used speech corpora. Section 4 presents the attributes selection and the performed experiments. Section 5 gives the results in terms of intra speech corpus and inter speech corpora robustness. Finally, Section 6 gives the concluding remarks and our future research directions.

2. FEATURES EXTRACTION

The features are extracted according to the specifications of the speech recognition front-end algorithm of the ETSI ES 202 211 V1.1.1 standard [1]. The specification covers the computation of feature vectors from speech waveforms sampled at a rate of 16 kHz or 8kHz. The offset-free input signal is divided into overlapping frames of $25ms$. The frame shift interval (difference between the starting points of consecutive frames) is $10ms$. The final feature vector extracted for every frame consists of 15 coefficients: the log-energy coefficient, the 12 cepstral coefficients $C_1 - C_{12}$, the pitch period, and the voicing class. The first and second time derivates are computed for the log-energy coefficient and the 12 MFCCs. In all, excluding the classification of the frame, there are $13 \cdot 3 + 1 = 40$ time series per segment. Prior to subsequent processing all null elements (unvoiced frames for which the front-end is unable to compute the pitch) are eliminated from the pitch sequence, and the initial and final frames classified as containing silence are eliminated from the MFCC sequence (thus removing any initial and final silence frames from the segment). The value of a single pa-

© EURASIP, 2010 ISSN 2076-1465

rameter extracted from a frame lasting a few milliseconds is of little significance to determine an emotional state. It is, on the contrary, of interest to investigate the trend taken by the parameter over time. Certain statistics such as the average, minimum and maximum values are extracted from time segments of speech signals. On the basis of the trend followed by each of the features extracted, the following sequences are computed: local maxima; local minima; distances between local maxima; distances between local minima; distances between local minima and maxima; distances between local maxima and minima; slopes between local minima and maxima; slopes between local maxima and minima; differences between minima and maxima; differences between maxima and minima. For all these sequences the following statistical informations are estimated: mean; variance; maximum; minimum; difference between maximum and minimum; 1st quartile; 2nd quartile (median); 3rd quartile; interquartile range (3rd quartile - 1st quartile) [20, 18]. Two further statistical features are obtained by evaluating the ratio between the number of relative minima and the number of frames and that between the number of relative maxima and the number of frames. According to the classification of the frames, the following sequences are also extracted: length of silence segments; length of unvoiced segments; length of mixed segments; length of voiced segments. For these new sequences all the statistical informations mentioned above are estimated. A final statistical attribute estimated is the ratio between the number of transitions between various states and the number of frames in the segment. We therefore had $40 \cdot 10 \cdot 9 + 40 \cdot 2 + 4 \cdot 9 + 1 = 3717$ attributes for each segment to be classified.

3. SPEECH CORPORA

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech recognition. It is evident that research into emotional speech recognition is limited to certain emotions, because the majority of emotional speech data collections encompass 5 or 6 emotions, although there are many more emotion categories in real life. Four speech corpora were used in this research: the first, in German, is called FAU AIBO Emotion Corpus (AIBO) [19, 2] and contains semantic units made up of chunks; the second, in German, is called the Berlin Database of Emotional Speech (EMO-DB) [3] and contains semantic units made up of sentences; the third, in English, is called Speech Under Simulated and Actual Stress (SUSAS) [11] and comprises semantic units made up of single words; the fourth, in German, is the audio-only part of the Vera-Am-Mittag (VAM) [10] Database and comprises semantic units made up of sentences.

3.1 AIBO Database

This corpus consists of speech data of 51 children at the age 10-13 years interacting with Sony's pet robot Aibo. The data was collected at two different schools, MONT (8 male and 17 female) and OHM (13 male and 13 female). The audio recordings of the children have been segmented manually into small, syntactically meaningful 'chunks' using syntactic-prosodic criteria. The data is annotated with 11 emotion categories by five human labelers on the word level: joyful, surprised, emphatic, helpless, touchy (i.e. irritated), angry, motherese, bored, reprimanding, rest (i.e. non-neutral, but not belonging to the other categories) and neutral. We use only the portion of the database which were recorded at OHM school. The chunks in the database were grouped to work with a two-class problem. It consists of the classes EMO (subsuming joyful, surprised, emphatic, helpless, touchy, angry, motherese, bored, reprimanding, rest) and IDL (consisting of all neutral states). The class EMO consists of 6601 chunks and the class IDL consists of 3358 chunks (respectively the 66.3% and 33.7% of the entire corpus).

3.2 Berlin Database of Emotional Speech

This database comprises 6 basic emotions (anger, boredom, disgust, anxiety, happiness and sadness) as well as neutral speech. Ten professional native German actors (5 female and 5 male) simulated these emotions, producing 10 utterances (5 short and 5 longer sentences), which could be used in everyday communication and are interpretable in all applied emotions. The recorded speech material of about 800 sentences (7 emotions · 10 actors · 10 sentences + some second versions) was evaluated with respect to recognizability and naturalness in a forced-choice automated listening test by 20-30 judges. After selection, the database contained a total of 494 sentences (286 uttered by women and 208 by men).
The sentences were not equally distributed between the various emotional states: 55 frightened; 38 disgusted; 64 happy; 79 bored; 78 neutral; 53 sad; 127 angry. The sentences in the database were grouped accordingly to the two-class problem in this manner: EMO class (subsuming anger, boredom, disgust, fear, happiness and sadness emotions) and IDL class (consisting of neutral state). The class EMO consists of 451 sentences and the class IDL consists of 78 sentences (respectively the 85.2% and 15.8% of the entire corpus).

3.3 Speech Under Simulated and Actual Stress Database

The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. The five stress domains include: talking styles (slow, fast, soft, loud, angry, clear, question); single tracking task or speech produced in noise (Lombard effect); dual tracking computer response task; actual subject motion-fear tasks (G-force, Lombard effect, noise, fear); psychiatric analysis data (speech in states of depression, fear, anxiety). The database contains both simulated speech under stress (*Simulated Domain*) and actual speech under stress (*Actual Domain*). A common highly confusable vocabulary set of 35 aircraft communication words makes up the SUSAS database. The words are uttered by 9 male speakers representing the three main USA dialects (General American, Boston, New York). Each style contains 2 recordings of the same word by each speaker. The audio is sampled at 8kHz with a resolution of 16 bits per sample. In this research only 7 of the 11 available states were used: *Angry, Fast, Lombard, Slow, Soft and Training*. Due to the short duration of the recordings which did not allow the front-end to extract the features correctly, 100 ms of silence were added at the start of each recording. The sentences in the database were grouped accordingly to the two-class problem in this manner: EMO class (subsuming Angry, Fast, Lombard, Slow and Soft states) and IDL class (consisting of Training state). The class EMO consists of 3150 words and the class IDL consists of 3780 words (respectively the 45.5% and 54.5%

of the entire corpus).

3.4 Vera-Am-Mittag Database

The database is the audio-only part of the Vera-Am-Mittag (VAM) Database, called VAM-Audio. This data was collected from a talk-show on German free-TV channel Sat1. It consists of segmented utterances of talk-show guests' speech in the show "Vera am Mittag" (Vera at noon), recorded in 2005. Each talk-show consists of several dialogs, and each dialog consists of spontaneous, unscripted discussions between 2 or 3 guests, moderated by the anchorwoman Vera. The language is German. The corpus contains 47 speakers (11 male and 36 female), and a total of 947 sentences. This database was evaluated by 17 human listeners. They assessed the emotional content in terms of the emotion primitives valence, activation, and dominance in each sentence. The merged evaluation results using the method described in [8] was used to group the database accordingly to the two-class problem in this manner: EMO class (if the merged valence is less than -0.25 or greater than 0.25) and IDL class (if the merged valence is greater than -0.25 and less then 0.25). The class EMO consists of 843 sentences and the class IDL consists of 947 sentences (respectively the 47.1% and 52.9% of the entire corpus).

4. ATTRIBUTES SELECTION

Whereas it would appear, intuitively, that a large number of features would improve the discrimination capabilities of a classification system, in reality various studies have shown that this is not always true. By reducing the size of the classification vector, the system is provided with a more compact and more easily interpretable set of data, the performance of the learning algorithm is improved and the speed of the system increased [20, 14, 12, 6, 7, 5]. In this work we try to verify the robustness of one of the most used attributes selection method applied to the emotional speech classification. An attributes selection criteria is composed of two parts: the attributes evaluator and the search method. In this paper, of the *attributes selection* techniques provided by WEKA, we used *CFSSubsetEval*. This algorithm uses as a feature evaluator *Correlation-based Feature Selection*, which tries to identify and discard components that are closely correlated with one another. As search method we used the *best-first* search strategy. To evaluate the robustness of the attribute selection criteria we split up each corpus in 2 non overlapping different sets. We then use the first set for training and the second set for testing. Then we split both the train and the test part into 10 different non overlapping subsets. Using the WEKA *CFSSubsetEval-BestFirst* method we evaluate the best subset of attributes for each split using the training part of the database. So we have 10 different subsets of attributes for each corpus. The rating of each attributes can't be evaluated regardless of the other attributes belonging to the subset of selected attributes. This because the rating of a vector of selected attributes is related to all the attributes that compose it. In other words the contribute of an attribute in a vector is not absolute but related to the other attributes composing the vector. For this reason we chose to evaluate each vector of selected attributes a predefined classification method and the performance of the classification. To evaluate the independence of the selected attributes from the split used to obtain the vector, the attributes selected for split x where used to train and test splits $\overline{x}$ (i.e., all others split except x) for each corpus. So, for example, we use the attributes selected using the split 1 of the AIBO corpus to train and test all the remaining splits $2-10$ of the same corpus. Then we use the attributes selected using the split 2 to train and test all the remaining splits $1, 3-10$, rotating until we have exploited all the possible combinations (90 for each corpus). If the attribute selection method is robust over this intra-corpus analysis we should obtain similar classification performance for each test performed. In a subsequent experiment we analyse the robustness of the attribute selection method over the different speech corpus. In this experiment we should prove if there is dependence between the selected attributes and the corpus. So we use the selected attributes from each split of a predefined corpus to train and test every other splits of the remaining corpora. For example, we use the attributes selected using the split 1 of the AIBO corpus to train and test the splits $1-10$ of the EmoDB, SUSAS and VAM corpora. For each split we execute 30 trainings and testings, so for each corpus we perform 300 trainings and testings and, globally in this experiment, we perform 1200 trainings and testings. If the attribute selection method is robust over this inter-corpus analysis we should obtain similar classification performance for each test performed, or we could point out correlations with the language, the length of the semantic unit (words or sentences), the spontaneous or simulated type of the emotions. To evaluate the performance we used always the WEKA *NaiveBayes* method to build a model with the train part of the split and to execute the test with the test part of the split.

5. RESULTS

In Figure 1 we report the performance of the bimodal classification using the same split both for attributes selection and for training and testing. In this and all others results the performance were evaluated as the ratio between the rights classification and the total number of elements to classify (correctness). Due to limitations on space we can present only results in term of correctness but we have also the results in terms of true/false positive and true/false negative for IDL and EMO classes. The results obtained with EmoDB corpus are the best. We can justify this results due to the presence of marked simulated emotions and to the use of sentences as semantic unit. Performance obtained using the SUSAS corpus are lightly better than performance obtained using the AIBO and VAM corpora. In Figures 2a-2d you can see the results of the Intra-Corpus experiment (i.e., when we use the attributes selected for a split and perform the training and test using all others splits of the same corpus). The columns of each graph outline the results, in term of correctness, varying the split used to select the attributes. Each result is averaged over all the 9 possible combinations. The graphs are respectively for AIBO, EmoDB, SUSAS and VAM corpus. As you can see the results are aligned to that obtained using the same split during the selection, training and testing phase. So we can conclude that there is absolutely no dependence caused by the split used to select the attributes. Figures 3a-3d show the results of the Inter-Corpus experiment (i.e., when we use the attributes selected for a split of a corpus and perform the training and test using all other splits of the other corpora). A group of columns of each graph outline the results, in term of correctness, varying the split used to select the attributes.

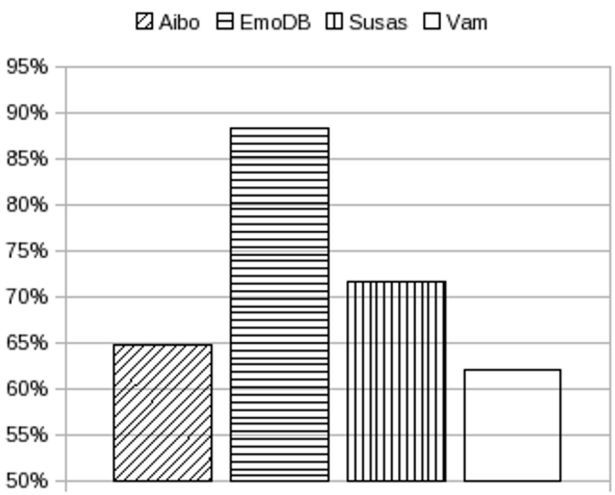


Figure 1: Performance in terms of correctness when we use the same split for attribute selection and for training

Each column in the group is related to the corpus used to the training and testing phase and the results are averaged over all the 10 possible combinations. As you can see from the Figure 3a the performance remains fairly unchanged when we use the attributes selected with the AIBO corpus. We have only a very slight degradation of about 1% for AIBO corpus, 4% for SUSAS corpus and greater then 5% for VAM corpus. However a slight performance degradation is present when we use the attributes selected with the EmoDB corpus (Figure 3b). This degradation is in the order of 5% for all the corpora. Using the attributes selected with the SUSAS corpus we obtain a degradation in the performance in the order of $2-3\%$ for the AIBO corpus, $5-10\%$ for the EmoDB corpus and $3-4\%$ for the VAM corpus (Figure 3c). As you can see from the Figure 3d, a more marked performance degradation is present when we use the VAM corpus to select the attributes and the EmoDB to perform the classification (in the order of $5-15\%$), while a slight performance degradation is present when we use AIBO and SUSAS corpora (in the order of $2-5\%$). Summarizing we can conclude that there is only a slight relation between the performance obtained in the classification and the split or corpus used to perform the attribute selection. In particular the intra-corpus experiment outline that this relation is very slightly. The inter-corpora experiment outline that there is a slight degradation in the performance when we use the attribute selected with one corpus to classify the emotions of another corpus (regardless of language, length of semantic unit and type of emotions: spontaneous or simulated). This result can be considered very important when we use the selected attributes obtained from a recorded speech corpus in a real application where the test environment consist of semantic unit not collected in a laboratory. Furthermore, the partial independence from the language and the length of the semantic unit, permit to apply the subset selection regardless of the language and or semantic unit to use in the real application.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have addressed the robustness of the attributes selection applied to speech emotion recognition. Using features extracted from an audio signal by the ETSI ES 202 211 v.1.1.1 standard front-end, we were carried out two types of experiment: intra-corpus and inter-corpus to prove the robustness of the attribute selected both for different semantic unit of the same corpus and for different semantic unit of other corpora. To outperform the experiment we use 4 different speech corpora: AIBO, EmoDB, SUSAS and VAM. We use the WEKA *CFSSubsetEval-BestFirst* method to evaluate the best subset of attributes for different splits of each corpus and the *NaiveBayes* classification method to evaluate the performance. The results obtained show only a slightly relation between the corpus used to select the attributes and the performance obtained. So we can conclude that the *CFSSubsetEval-BestFirst* is a robust method to perform the attribute selection in the speech emotion recognition field. Possible feature works include the use of other subset selection method (both different search method, like Genetic Search, Random Search and different attribute evaluation criteria) and the evaluations of the performance using the models trained with the split of different corpora. Moreover we intend to investigate the performance using other Bayes classification methods and function classification method (based on Support Vector Machines and Multilayer Perceptron).

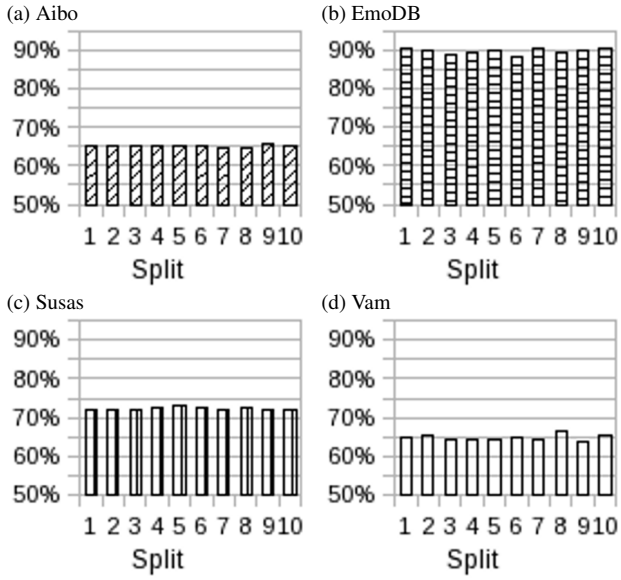


Figure 2: Performance in terms of correctness when we use a different split of the same corpus for training ad testing than that used for attribute selection

REFERENCES

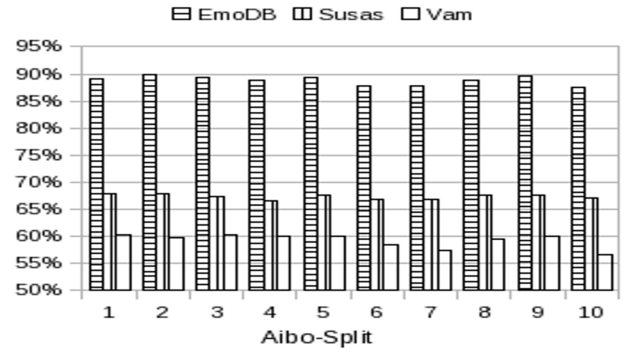
[1] ETSI ES 202 211 V1.1.1. Technical report.

[2] A. Batliner, S. Steidl, C. Hacker, and E. Noth. Private emotions vs. social interaction - a data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction*, 1-2:175–206, 2008.

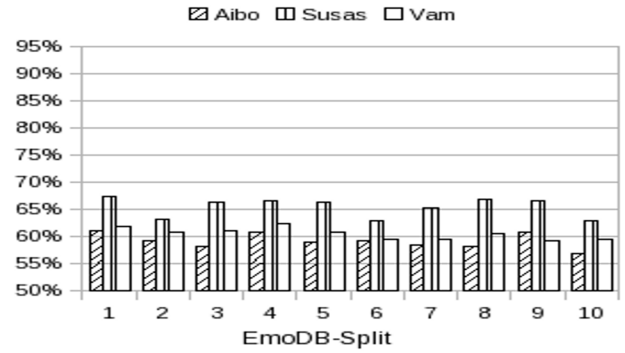
[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *InterSpeech*, pages 1517–1520, 2005.

[4] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for

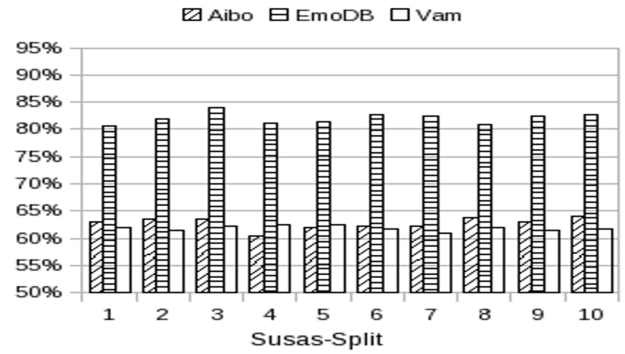
(a) Aibo



(b) EmoDB



(c) Susas



(d) Vam

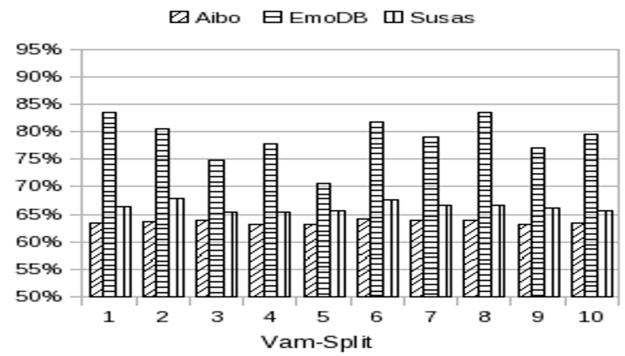


Figure 3: Performance in terms of correctness when we use a different corpus for training ad testing than that used for attributes selection

emotion detection. *IEEE Trans. on Audio, Speech and Language Processing*, 17:582–596, 2009.

[5] S. Casale, A. Russo, G. Scebba, and S. Serrano. Speech emotion classification using machine learning algorithms. In *Proc. IEEE ICSC*, pages 158–165, August, 4-7 2008. Santa Clara, CA, USA.

[6] S. Casale, A. Russo, and S. Serrano. Classification of speech under stress using features selected by genetic algorithms. In *14th European Signal Processing Conf.* Sept 4-8, 2006, Florence, Italy.

[7] S. Casale, A. Russo, and S. Serrano. Multistyle classification of speech under stress using features subset selection based on genetic algorithms. *Speech Communication*, 49(10-11):801–810, Oct-Nov 2007.

[8] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. IEEE ASRU*, pages 381–385, 2005.

[9] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49:787–800, 2007.

[10] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *ICASSP*, 2008. Las Vegas NV, USA.

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

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

[13] I. H.Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. The M. Kaufmann Series in Data Management Systems, J. Gray, Series Editor, Oct 1999.

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

[15] C. M. Lee and S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Processing*, 13(2):293–303, Mar 2005.

[16] W. V. M. Shami. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49:201–212, 2007.

[17] I. Murray and J. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Acoust. Soc. Amer.*, 93:1097–1108, 1993.

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

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

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