Emotion recognition from speech – Tools and Challenges

Abdulbasit Al-Talabani, Harin Sellahewa, & Sabah A. Jassim
Department of Applied Computing, Buckingham University, UK

ABSTRACT

Human emotion recognition from speech is studied frequently for its importance in many applications, e.g. human-computer interaction. There is a wide diversity and non-agreement about the basic emotion or emotion-related states on one hand and about where the emotion related information lies in the speech signal on the other side. These diversities motivate our investigations into extracting Meta-features using the PCA approach, or using a non-adaptive random projection RP, which significantly reduce the large dimensional speech feature vectors that may contain a wide range of emotion related information. Subsets of Meta-features are fused to increase the performance of the recognition model that adopts the score-based LDC classifier. We shall demonstrate that our scheme outperform the state of the art results when tested on non-prompted databases or acted databases (i.e. when subjects act specific emotions while uttering a sentence). However, the huge gap between accuracy rates achieved on the different types of datasets of speech raises questions about the way emotions modulate the speech. In particular we shall argue that emotion recognition from speech should not be dealt with as a classification problem. We shall demonstrate the presence of a spectrum of different emotions in the same speech portion especially in the non-prompted data sets, which tends to be more "natural" than the acted datasets where the subjects attempt to suppress all but one emotion.

Keywords: Classification, Emotion recognition, Dimension reduction.

1. INTRODUCTION

Speech Emotion Recognition (SER) is no longer a side issue. In the last decade, research in SER has become a major endeavor in HCI, and speech processing. The large number of studies published on SER reflects the demand for its use. Automatic speech recognition system can be more effective by incorporating emotion processing/analysis [1].

Variation in voice tones as well as internal physiological changes while uttering a sentence (or even a single word) combine to generate the speaker emotional state. Perfect recognition of emotions is not easy even by human when listening to each other; sometimes the human cannot recognize his own innermost emotion. In fact some aspects of internal feeling remains hidden and not detectible from the speech, especially when the speaker need to suppress emotions. Therefor computer-based system cannot do beyond what is observed from the speech sample input [2]. Consequently, categorizing emotional speech samples is a serious challenge due to the long debate about the real meaning of "emotion" and the emotional classes that should be dealt with. That is why Batliner et al. [3] prefer the concept of emotion-related states to avoid that "fruitless debate".

Another challenge is to identify emotion relevant features that can be extracted from the raw speech signal or its frequency domain version. Recent studies observed that emotion related information in the speech is spread along different kind of features. This could be due to the acoustic variability as a consequent of the existence of different sentences, speakers, speaking styles, and speaking rates [4].

Researchers adopted large feature set (thousands), whether directly as in [5], or followed by feature selection step like [3]. The high number of features is a consequence of the non-agreement on a limited number of relevant features to emotion [6]. Studies on the relationship between global prosodic speech features and the basic emotions have shown that prosodic features provide a reliable indication of the emotion. However, other studies yield contradictory conclusions on the effect of emotions on prosodic features. For instance, while Murray and Arnott [7] indicate that a high speaking rate is associated with the emotion of anger, Oster and Risbeerg [8] conclude the opposite effect. In a correlation based feature selection study, Perez et al. [9], argued that the most relevant features to discriminate classes in the valence, arousal, and dominance dimensions are MFCC, cochlea grams, and LPC, in addition to contributions of Mel features in valence and dominance, and energy in arousal dimension discriminating. A strong relation between voice quality and perceived emotion have been demonstrated subjectively (i.e. human observers) in [10]. Voice quality is also reported to regularly reference full-blown emotions [11]. A study by Schuller et al. [12] reports that pooling together features extracted at different sites improved emotion recognition accuracy.

This non-agreement on a set of features could be due to the nature of emotional databases used [3], which are captured under different conditions and factors. Simulated versus spontaneous, number and definition of classes, cultures, and\or the circumstance of recordings could all be contributing to the need of different sets of features as well as extended feature attributes.

Normally, dimension reduction and features selection in the feature or a transformed space is applied. Feature selection is costly and impractical, due to the complexity of the optimization that target a suitable feature subset among a high number of features, especially when using the wrapper methods. The alternative filter based feature selection methods are not based on classification decision but on some data characteristics like correlation or entropy. Filters are reported to be more convenient for high dimensional data [13]. However filter-based feature selection methods are not necessarily suitable for all classifiers, and feature selection cut off points could lead to neglecting some "important" information included in the un-selected features.

Dimension reduction schemes have similar effect of feature selection and are based on transforming the data in the feature space into another subspace using a specific transformation map like Principle Component Analysis (PCA) or non-adaptive Random Projection (RP). Such transformations aim to help de-correlate the transformed data attributes that are referred to as Meta-features.

These methods exploit information included in all the available features, but they are reported to be unsuitable in feature mining [14]. PCA is a data-dependent projection that de-correlates the features of the original data space by choosing a number of eigenvectors of the covariance matrix with the largest eigenvalues. In contrast, RP is a data independent projection, i.e. the projection matrix is generated independently from the training data. Therefore the use of RP allows the model to avoid the training stage. However the success of PCA or RP in the recognition model is classifier dependent.

In this paper, we shall investigate the use of Support Vector Machine (SVM) and Linear Discriminant Classifier (LDC), when dimension of speech data are reduced at different rates by PCA and RP. We shall demonstrate that the fusion of various subsets of the Meta-features extracted by the PCA at the score level using LDC classifier outperform the state of the art scheme. However there is no significant improvement in the performance especially when using non-prompted datasets. The noticeable huge gap between on accuracy rates achieved from prompted datasets and the more natural non-prompted datasets of speech, raises questions about the way emotions modulate the speech.

The experimental work reported here is conducted on different databases, which will be presented in section 2. A description of the features extracted is given in section 3. Two models are suggested in this study and presented in section 4, while the result and discussion is shown in section 5. Finally we describe our conclusions in section 6.

2. DATABASES

This section will describe the emotional databases used in this study.

2.1 The Kurdish emotional speech database

This database [15] includes 7 emotions (Anger, Happiness, Sad, Fear, Boredom, Surprise, and Neutral) acted by 6 male and 6 female actors. Each actor utters 10 Kurdish sentences on 4 different sessions for each emotion. There is a total of 3360 recordings (i.e.,12×7×10×4 recordings). The speakers have been told to act these sentences and express them in their own style, by remembering/imagining a situation with the relevant emotion. The speakers have different acting experience (2 to 8 Years) and ages (19 to 36 years). The recording process has been done in a quiet room without restriction in the recording path. 10 listeners participated in a subjective test to determine the perceptibility of the emotions. The average of correct labeling was 41%.

2.2 Berlin emotional speech database

The Berlin emotional speech database [16] (also called Emo-DB) is an emotional speech database in German language. Ten professional native German actors (5 Male and 5 Female) were involved in recording 10 German sentences in 7 emotions. The considered emotions were Neutral, Anger, Happiness, Sad, Fear, Boredom, and Disgust. Some of the utterances were recorded in more than one session. Total of 535 utterances remained in the database after eliminating some unconvincing recordings based on a subjective test.

2.3 FAU Aibo (non-prompted) database

The Aibo database [17] [18], was designed by recording children's sound, which are colored by different emotion, when they interact with Sony's pet robot Aibo. The children were led to believe that the robot was responding to their commands, whereas it was actually controlled by a human operator in a Wizard-Of-Oz manner. Sometimes the Aibo disobeyed the child's command, to lead to different emotional reaction. The data were collected at two different schools identified by 'Ohm' and 'Mont'; the number of speakers was 26 and 25 respectively. Five experts labeled each word in the database independently, into 10 categories: angry, touchy, joyful, surprised, bored, helpless, motherese, reprimanding, emphatic, and 'other' for the remaining cases. The categories were mapped into four classes: anger, emphatic, neutral, and positive in addition to the fifth class for rest. The labels of the words were mapped to so-called 'turn' (i.e. utterances) using heuristic method described by Steidl (16), to obtain a total of 18216 sentences. The Aibo corpus formed the focus of the Interspeech 2009 emotion challenge [19].

3. FEATURE EXTRACTION AND TRANSFORMATION

3.1 Feature Extraction

Feature extraction is an essential component of any pattern recognition system that could greatly influence its performance. Here, we adopted a set of features extracted by the openEAR software. These features include five groups of Low Level Descriptors (LLDs), described in Table 1, together with several statistical parameters as described in Table 2. For each sample 39 statistical functional are computed for each one of the 33LLDs and their first and second deltas, results in 6552 features. This large 'brute force' set of features includes most of the speech signal parameters mentioned in the emotion related literature.

Table 1: (33) Low Level Descriptor (LLD) used in Acoustic analysis with OpenEAR

Feature Group	Description
Raw Signal	Zero-crossing-rate
Signal Energy Pitch	Logarithmic fundamental frequency F_0 in Hz via cep- strum and autocorrelation (ACF). Exponentially smoothed F0 envelope.
Voice Quality	Probability of voicing ($ACF(T_0)/ACF(0)$)
Spectral	Energy in bands 0-250Hz, 0-650Hz, 250- 650Hz, 1-4kHz 25%, 50 %, 75%, 90% roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table 2: (39) functionals and regressions coefficient applied to the LLD contour.

Functionals, etc.	#	Functionals, etc.	#
Respective rel. position of max./min. value	2	Quartiles and inter-quartile ranges	6
Range (max.-min.)	1	95 % and 98 % percentile	2
Max. and min. value - arithmetic mean	2	Std. deviation, variance, kurtosis, skewness	4
Arithmetic mean, quadratic mean	2	Centroid	1
Number of non-zero values	1	Zero-crossing rate	1
Geometric, and quadratic mean of non-zero values	2	# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. Mean	4
Mean of absolute values, mean of non-zero abs. values	2		
Quadratic regression coefficients and corresp. approximation error	5	Linear regression coefficients and corresp. approximation error	4

3.2 Feature Transformation

To avoid the curse of dimensionality feature reduction the PCA is a well known and widely used. However the PCA projection matrix, consisting of the eigenvectors, which depend on the training set, and therefore PCA-based recognition may not scale up well when the population grows considerably. Moreover, computing the PCA is somewhat restricted, especially when the number of samples available is relatively small in comparison to the space dimension. For example, the Emo-DB database has about 450 samples for training, while our space dimension is 6552. In such a case, PCA cannot create more eigenvectors than the number of samples, which means that the reduced dimension is of limited use. It is therefore desirable to have an efficient non-adaptive method that reduces dimensions without a training set, while distances/similarities between two projected vectors are not much different from their original vectors. In recent years, research into the mathematics of Compressive sensing has revealed random Restrictive Isometry Projection (RRIP) matrices thet can provide such dimension reduction tools. It has been shown that this type of RP can preserve the Gaussian mixture clusters without overlapping [20]. The columns of these RP matrices are required to have unit length in order to control the growth of distances after projection.

The Bernoulli and Gaussian random matrices are known to be RRIP, but these are by no means the only cases. Generally any independently and identically distributed (i.i.d.) data matrix could be used for this purpose. Below are some other suggested RRIP matrices:

1. Achlioptas suggests a matrix defined as:

$$A_c(i,j)=\sqrt{3}\begin{cases} +1\_with\_probability\_1/6 \\ 0\_with\_probability\_2/3 \\ -1\_with\_probability\_1/6 \end{cases}$$

2. Circulant Toeplitz matrix, where every row in this matrix is the right cyclic shift of the row above, as below:

$$T_n=\begin{bmatrix} a_0 & a_1 & \cdots & a_n \\ a_n & a_0 & \cdots & a_{n-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_1 & \cdots & a_n & a_0 \end{bmatrix}$$

In this study we use a special form of RP matrices, called Random Binary matrix (RB) consisting of 0s and 1s, with equal probability ratio r =0.5. The entries in each column are the coefficient of linear combinations of the sample scores in each dimension.

4. EMOTION RECOGNITION MODEL

In the design of the emotion recognition models (Model_I and Model_II), we applied Leave One Speaker Out (LOSO) for both Emo_DB and Kurdish databases. However, for the FAU Aibo database we use the same standard adopted in the interspeech09 [19] challenge, which takes the speech sample recorded in one of the schools denoted by (Ohm) as training, and the other samples, which are recorded in the second school (Mont) for testing. The two models, implemented using Matlab 7.12.0. as shown below:

Model_I:

The emotion recognition Model_I is designed using the OP features preprocessed by PCA and RB with the LDC and SVM classifiers. The LDC relies on the assumption of multivariate Gaussian distribution of the clusters, and consequently supposed to work well with a de-correlated set of data. A pooled estimate of the covariance is used to produce the multivariate model. The number of Meta-features is evaluated using LOSO cross validation.

On the other hand, this model adopts pairwise SVM with a linear kernel and SMO optimization method which needs to build n=c!/((c-2)!2) machines, where c is the number of classes. The FAU_Aibo database contained an unbalanced number of samples per class; therefore, Synthetic Minority Over-Sampling Technique (SMOTE) [21] is applied in the training stage of the FAU Aibo corpus. For the same reason, Un-weighted Average Recall (UAR) accuracy is used to measure the recognition accuracy.

Model_II: Fusion Model

Fusion of feature sets from different sites especially in biometrics is reported to improve the recognition accuracy. In this work Meta-feature subsets represented by highest 50, 150, and 250 PCs are assumed to show different shapes of clases clusters. Therefore, feature fusion at the score level is adopted using these three subsets of Meta-features followed by LDC classifier. These different Meta-feature subset galleries increase the chance of all the samples to be represented using more transformed Meta-features in the PC space. This resampling could help overcoming the issue of limited number of speaker in the datasets, which might attract the classifier to make a bias decision toward some subject(s). The fusion could be done at the score level, or at the decision level. The score is a confidence value of sample belonging to a class. In the case of LDC the score used is posterior probability of each sample to belong to each class. The score gives the most information about the similarity between a sample and an individual or a group of samples. The scores are weighted by a set of evaluated weights generated using LOSO cross validation on the training set using the same model (Figure 1).

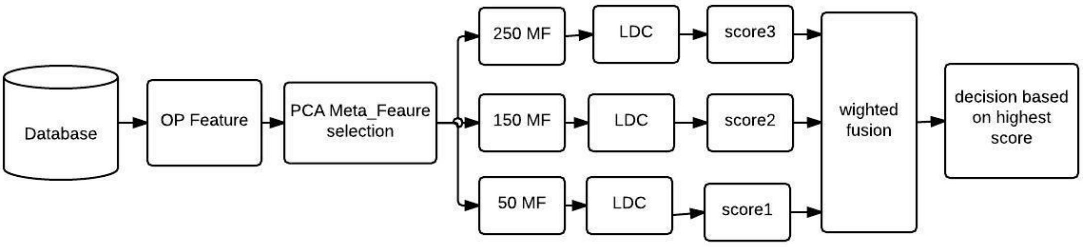


Figure 1: Meta- feature subsets using PCA fused with LDC classifier

5. RESULT AND DISCUSSION

Tables 3, below shows the accuracy of Model_I as well as the evaluated dimension of the Meta-feature selected using PCA and RB applied to the three databases. For each scheme, the results were obtained using both the LDC and the SVM classifiers. The accuracy results obtained are comparable to, and somehow better than, the state of the art schemes. For the three databases, the highest Average Recall (AR) is achieved when using PCA+LDC followed by the RB+SVM. This can be attributed to the fact that PCA de-correlates the feature vectors, and therefore it is more suitable for Gaussian distribution-based classifiers like the LDC. The LDC faces difficulties to build the multivariate Gaussian distribution model in relatively high dimensions, while the SVM is more capable in modeling high dimensional feature space. The number of available training samples limits the number of PCA meta-features, whereas the RB is not restricted. This problem is more apparent in the case of Emo_DB database. Hence SVM works better with RB than with PCA. However the time complexity and cost of the SVM is a serious disadvantage.

The first most significant PCs represent the most informative dimensions, but after a certain peak adding more PCs degrades the recognition accuracy as an indication of adding non-significant information (Figure 2). The performance of the RB, on the other hand, improves initially but never deteriorate with additional meta-features.

Model_II is an attempt to improve the recognition performance by Meta-feature fusion. A slight improvement is observed as shown in (Table 4), using validated weights by applying LOSO cross validation. Although improvement is modest, the fusion of Meta-Feature subsets looks to give different representation of the data. The interesting observation is the weights produced from the model validation. The three datasets benefited from different distribution of weights, indicating the degree of freedom of the dimensions of the transformed data.

Emotional studies are rarely involved in application of speech emotion ranking. However score based decision is convenient for ranking the detected emotions. To show the score-based emotion detection decision, a spider chart is suggested to visualize the obtained scores from the classifier. The spider chart visualizes the scores for each tested sample toward the available emotions. The produced scores from the LDC can be directly used for this purpose. For example in testing one sample which is belong to the FAU-Aibo (Figure 3 (a)), the implemented spider chart shows that the sample is more likely to belong to Anger emotion (55%), the second possible emotion is the Emphasized emotion

(37%), next emotions are Neutral, Positive, and Rest emotions for just 4%, 2%, and 2% scores respectively. The figure shows how that sample has high possibility to be ranked in advance position among the Anger and with less possibility the Emphasized emotions promotionally to its score compared with the other available sample scores.

Table 3. Model_I results for the three datasets used. UAR: is the un-weighted average recall. S. O. A.: is the State of Art. OP: refer to the use of the whole OP feature (not pre-processed) using SVM. Dim. av/std: is the average and standard deviation of evaluated number of.

Data	OP	S.O.A		Classifier	AR %	Dim. av/std
FAU-Aibo	(UAR) 39.1	(UAR) 44 [22]	PCA	LDC	46.1 (UAR)	410/45
				SVM	44 (UAR)	110/35
			RB	LDC	43.1 (UAR)	495/45
				SVM	**46.5** (UAR)	690/75
Emo_DB	85.2	89.9 [23]	PCA	LDC	**90.7**	200/20
				SVM	82.6	315/66.9
			RB	LDC	80.7	115/25
				SVM	87.2	690/78.2
Kurdish	38.4	44.6 [15]	PCA	LDC	**43.5**	190.9/30.1
				SVM	38.2	350/48
			RB	LDC	39.7	245.8/14.4
				SVM	42.2	710/69

Table 4. Model_II results for the three data sets

Datasets	Average recall before fusion %	Average recall %	Weights	S. O. A.
FAU_Aibo	46.1 (UAR)	47.6(UAR)	0.3,0.2,0.5	44 [22]
Emo_DB	90.7	91.2	0.5,0.4,0.1	89.9 [23]
Kuridsh	43.5	44.9	0.3,0.6,0.1	44.6 [15]

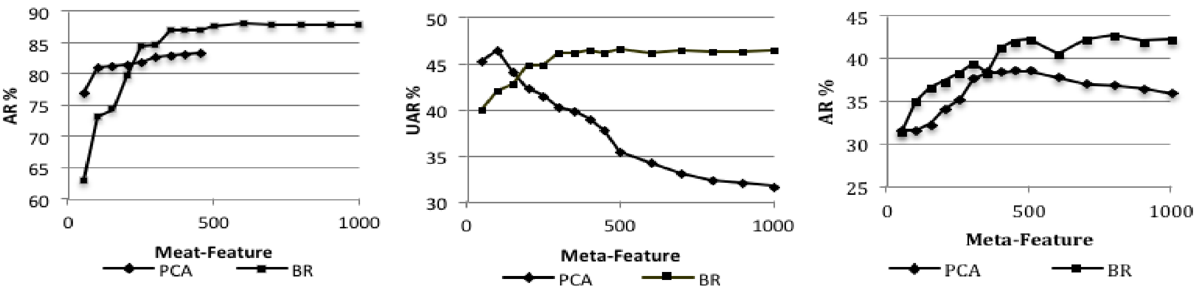
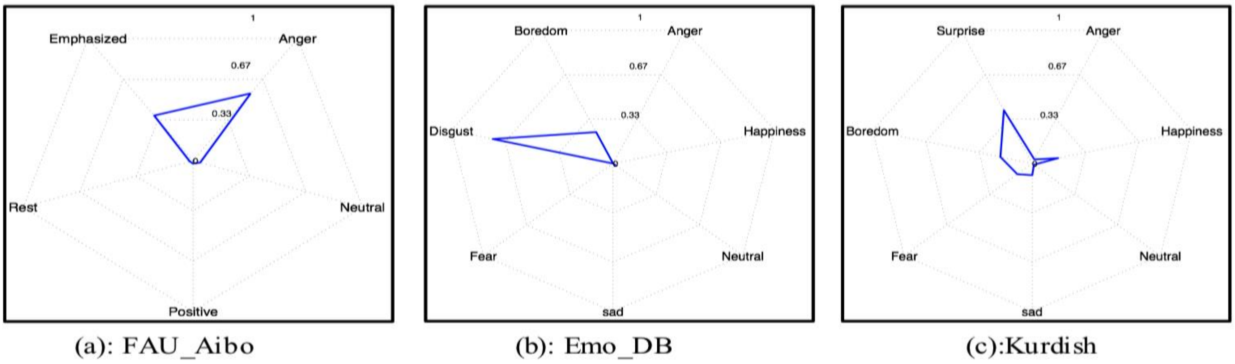


Figure 2. Emotion recognition accuracy of (Emo-DB, FAU-Aibo and Kurdish databases from left to right) pre-processed by PCA and RB, using SVM.

What is striking in these results, is the huge differences in the accuracy achieved by each of the emotion recognition schemes, including the SOA schemes, when tested on “acted” and on “non-prompted databases. First of all these extreme differences in accuracy cannot be attributed to the obvious difficulties associated with selecting a good representing sample for training, otherwise such a gap would not be so significant in the case of the non-adaptive RB scheme which does not require training.



(a): FAU_Aibo (b): Emo_DB (c):Kurdish

Figure 3. Spider chart for the classes score using Model II

We argue that this performance gap in the accuracy of emotion recognition from speech could not be dealt with as a simple classification problem. In other words speech is not sufficient for emotion recognition, and the same speech portion maybe modulated by a spectrum of different emotions, especially in the non-prompted datasets. Our argument is based on the fact that these databases are recorded under different circumstances and for different purposes. In general, determining the emotion of a person from a recorded speech, by human observers, is a very difficult task without access to facial expressions and/or bodily gestures during the uttering of the speech. This is exactly what happened when recording the non-prompted databases, which were supposed to capture emotions from the real life speech. This difficulty explains the rather low accuracy rates for the FAU Aibo and the Kurdish datasets. Although the Kurdish database is acted database; but the speaker are told to act the emotions using their own style, and all of the produced sentences have been involved in the study without removing the unconvinced ones. In the case, of the acted Emo-DB datasets the subjects obviously attempt to suppress all but one emotion that they are asked to stress, and consequently all the schemes achieve significantly higher accuracy. A sceptic may refute this last argument by pointing out that the best accuracy rate is a mere 91.2%. However, we would like to argue that achieving higher accuracy requires the speakers to be professionally actors of high standard. In fact, the Emo-DB database is designed for emotion synthesis purpose (not recognition) [24], and according to the documentation of the database about 250 samples out of original 800 recorded samples have been removed from the dataset due to variation in expert listeners judgment.

The above discussion, together with the observations depicted in the above Spider charts, motivate our hypothesis, to be investigated in the future, that emotion recognition from speech should not be dealt with as a simple classification problem. In other words speech is not sufficient for emotion recognition, but the same speech portion maybe modulated by a spectrum of different emotions, especially in the non-prompted scenarios. We should attempt associate more than one “emotional” class in the same speech sample but with weights, because the speech portion might contain information related to more than one emotion.

6. CONCLUSION

Emotion recognition performance can be improved by extracting high number of feature (brute force) followed by appropriate dimension reduction techniques. Coupling dimension reduction technique with suitable classifier is important to build capable SER model. Meta-feature subsets fusion gives different representation of emotion clusters, and can contribute in improving emotion recognition performance. However a big gap of recognition performance is observed among different emotional databases, which reflect the ambiguity and difficulty of emotion classification. We provided arguments in support of a proposed hypothesis that needs to be tested in the future: emotion recognition from speech should not be dealt with as a simple classification problem. The speech samples seems to contain a spectrum of the defined “emotions”, which is again an interpretation to the “fruitless” debate on what exactly is the emotion.

REFRENCES

[1] Koolagudi, S. G. and Rao , K. S., "Emotion recognition from speech: a review," International Journal of Speech Technology. 15, 99–117 (2012).

[2] Picard, R. W., Vyzas, E., and Healey, J., "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State ," IEEE Transactions on Pattern Analysis and Machine Intelligence. 23(10), 1175-1191 (2001).

[3] Batliner, A., et al., "Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech," Computer Speech & Language. 25(1), 4 – 28 (2011).

[4] El Ayadi, M., Kamel, M. S., and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, 44(3), 572 – 587 (2011).

[5] Schuller, B., Vlasenko, B., and Eyben, F., Rigoll, G., and Wendemuth, A.,"Acoustic Emotion Recognition: A Benchmark Comparission of Performance," Proc. ASRU, 2009.

[6] Vogt, T. and Andre, E., "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," Proc. ICME, 474 – 477(2005).

[7] Murray, I. R. and Arnott, J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion," J Acoust Soc Am, 93(2), 1097–1108(1993).

[8] Oster, A. and Risberg, A., "The identification of the mood of a speaker by hearing impaired listeners," STL-QPSR, 27(4), 79-90(1986).

[9] Pérez-Espinosa, H., Reyes-Garcia, C. A. and Villasenor-Pineda, L., "Acoustic feature selection and classification of emotions in speech using a 3D," Biomedical Signal Processing and Control, 7(1), 79-87 (2012).

[10] Gobl, C. and Chasaide, A. N., "The role of voice quality in communicating emotion, mood and attitude," Speech Communication, vol. 40(1-2), 189 – 212(2003).

[11] Cowie, R. , Douglas-Cowie, E. , Tsapatsoulis, N. and Votsis, G., "Emotion recognition in human–computer interaction," IEEE Signal Process. Mag., 18(1) 32–80 (2001).

[12] Schuller B. et al., "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," Proc. interspeech, 2253-2256 (2007).

[13] Yu L. and Liu, H. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th International conferance of Machine Learning, 856-863 (2003).

[14] Batliner, A., Steidl, S., Hacker, C. and Nöth, E. "Private Emotion vs. Social intraction a Data driven Approach toward Analysis Emotion in Speech," User Modelling and User-Adapted Intraction (umani), 18(1-2), 175-206 (2008).

[15] Al-Talabani, A., Sellahewa, H. and Jassim S., "Excitation source and low level descriptor features fusion for emotion recognition using SVM and ANN," Proc. CEEC, 156 – 161 (2013).

[16] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B., "A database of german emotional speech," Proc. Interspeech, 1517-1520 (2005).

[17] Steidl, S., [Automatic classification of emotion-related user statesin spontaneous childern's speech], PhD thesis. Depatrment of Computer Science, University of Erlangen-Nuremberg, Germany, 2009.

[18] Batliner, A., Steidl, S. and Noth, E. "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus," Proc. Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect, (2008).

[19] Schuller, B., Steidl, S., Batliner, A., Schiel, F. and Krajewski, J. "INTERSPEECH 2011 speaker state challenge. In: Interspeech'11," Proc. Interspeech11, 3201-3204 (2011).

[20] Dasgupta, S., "Learning Mixtures of Gaussians," Proc. Foundations of Computer Science, 634-644 (1999).

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 16, 321–357 (2012).

[22] Schuller, B., Batliner, A., Steidl, S and Seppi, D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Communication, 53(9-10), 1062–1087 (2011).

[23] Vlasenko, B., Schuller, B., Wendemuth, A. and Rigoll, G., "Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech," Proc. Interspeech07, 2249-2252 (2007).

[24] Bjorn W., [Schuller, Intellegent Audio Analysis], Springer, Munchen, Germany, 194-195 (2013).