# SPEAKER INDEPENDENT SPEECH EMOTION RECOGNITION BY ENSEMBLE CLASSIFICATION

*Björn Schuller, Stephan Reiter, Ronald Müller, Marc Al-Hames, Manfred Lang, Gerhard Rigoll*

Institute for Human-Machine Communication
Technische Universität München
Arcisstraße 21, D-80333 München, Germany
(schuller | reiter | müller | lang | rigoll)@mmk.ei.tum.de

## ABSTRACT

Emotion recognition grows to an important factor in future media retrieval and man machine interfaces. However, even human deciders often experience problems realizing one's emotion, especially of strangers. In this work we strive to recognize emotion independent of the person concentrating on the speech channel. Single feature relevance of acoustic features is a critical point, which we address by filter-based gain ratio calculation starting at a basis of 276 features. As optimization of a minimum set as a whole in general saves more extraction effort, we furthermore apply an SVM-SFFS wrapper based search. For a more robust estimation we also integrate spoken content information by a Bayesian Net analysis of ASR outputs. Overall classification is realized in an early feature fusion by stacked ensembles of diverse base classifiers. Tests ran on a 3,947 movie and automotive interaction dialog-turns database consisting of 35 speakers. Remarkable overall performance can be reported in the discrimination of the seven discrete emotions named in the MPEG-4 standard with added neutrality.

## 1. INTRODUCTION

The importance of emotion recognition for multimedia retrieval, and human computer interaction is commonly agreed [1]. Speech information is the most promising among other sources as mimic, physiological or context data analysis. A number of systems already exist that are capable of recognition by acoustic or linguistic information. However, it is mostly reported that speaker dependent recognition leads to far better results. This is hardly surprising since first psychological assumptions could be confirmed by a survey conducted to measure human performance on this task. 12 individuals were asked to re-classify their own 70 emotional audio samples, previously recorded, on basis of an emotional category set of seven described later on. Thereby mean accuracy of 83.7% was observed. Unlike this result the recognition performance dropped to 64.7% on the task of determining the expressed emotions of unknown persons. In this contribution we therefore aim to focus on speaker independent performance, which is a must have for many applications as call centers, media segmentation, public transport observation or further scenarios, where the speaker is either unknown or no sufficient material for a model adaptation, not to mention a complete training, exists.

The paper is structured as follows: Section 2 deals with the description of the databases used, section 3 provides insight in acoustic feature selection, section 4 introduces the ensemble classification variants considered, section 5 shows linguistic content processing which is an important factor [2][3], and finally conclusions are drawn.

## 2. DATABASE DESCRIPTION

The emotions used resemble the far spread MPEG-4 set, namely joy, anger, disgust, fear, sadness, surprise and added neutrality. Within acoustic feature selection and classifier evaluation 2,440 samples of 35 speakers of automotive interaction dialog-turns are used [4]. The emotions are almost evenly distributed among the individuals, meaning that 10 samples per emotion and speaker exist in average.

In order to get a high number of samples with acoustic and linguistic content in sufficient quality considering speech recognition and extraction of acoustic emotion features we decided for acted emotions as a further corpus. The textual content was taken from movie scripts of seven U.S. American movies from the years 1977 until 1999. Namely these are *Alien, Annie Hall, Five Easy Pieces, Notting Hill, Scream, 10 things I hate about you,* and *Toy Story*. Herewith a wide bandwidth of genres, i.e. Science-Fiction, Comedy, Drama, Horror, and Fantasy could be covered and has been selected in order to include

all emotions desired. The utterances were annotated phrase-wisely by two test persons and 1,144 phrases consisting of 7.0 words in average with identical labeling could be obtained. The set was supplemented by emotions of text-based Internet conversation labeled accordingly until 1,507 utterances were collected in total. The phrases were acted and recorded as single utterances in an anechoic chamber with a condenser microphone AKG-1000S MK-II over a long period to avoid anticipation effects of the three actors in total.

## 3. ACOUSTIC FEATURE SELECTION

Large numbers of diverse acoustic high-level features based mostly on pitch, energy, and durations were discussed considering their performance. However, sparse analysis of single feature relevance by means of filter or wrapper based evaluation has been fulfilled, yet. Within here we apply a Support Vector Machine (*SVM*) based Sequential Forward Floating Search (*SFFS*) as search function within feature selection (*FS*). SFFS is known for its high performance as showed in [5]. Thereby the evaluation function is the classifier, in our case SVMs, as described in the next section, which optimizes the features as a set rather than finding single features of high performance. The search is performed by forward and backward steps eliminating and adding features to an initially empty set. 276 static acoustic high-level features form the basis for this analysis.

| Rank | GR | Feature |
|---|---|---|
| 1 | 0.279 | Pitch maximum gradient |
| 2 | 0.187 | Pitch mean value, adapted |
| 3 | 0.072 | Energy mean value, normalized |
| 4 | 0.187 | Pitch mean value gradient |
| 5 | 0.097 | Signal number of zero-crossings |
| 6 | 0.073 | Signal median of sample values |
| 7 | 0.122 | Pitch relative maximum |
| 8 | 0.046 | Duration of silences mean value |
| 9 | 0.082 | Energy maximum gradient |
| 10 | 0.140 | Pitch range |
| 11 | 0.116 | Pitch mean dist. between reversal |
| 12 | 0.057 | Duration of voiced sounds std. dev. |
| 13 | 0.069 | Energy median of rise-time |
| 14 | 0.030 | Duration of silences median |
| 15 | 0.151 | Duration mean value of voiced sounds |
| 16 | 0.066 | Spectral energy below 250 Hz |
| 17 | 0.067 | Energy std. dev. dist. of reversal |
| 18 | 0.050 | Energy mean of fall-time |
| 19 | 0.051 | Energy mean dist. of reversal points |
| 20 | 0.035 | Energy relative maximum |

Figure (1): Acoustic feature set selection with SVM-SFFS and single feature Gain Ratio *(GR)* calculation

These consist of the named pitch, energy and duration features, of higher order formants, spectral features, and further ones which cannot be described here in detail. Feature pre-processing and extraction is described in [6]. Nevertheless in the table also single feature relevance information is provided by Gain Ratio. It can be clearly seen that the set comprises features of low single feature importance which cover aspects missing in the set. By reduction of the feature set different classifiers showed diverse behavior as will be seen in the next chapter.

It shall also be mentioned that in comparable works features are reduced by means of the well known Principal Component Analysis (*PCA*) and selection of the obtained artificial features corresponding to the highest Eigen-values [7]. However, the feature selection procedure based on PCA leads to a higher extraction effort. This is due to the fact that PCA requires extraction of all original features prior to linear superposition, which endangers real-time requirements. Such reduction has therefore not been considered herein.

## 4. ENSEMBLE CLASSIFICATION

With a relatively small number of training samples compared to the dimensionality of the data, a high danger of bias due to variances in training material is present. In order to improve instable classifiers as Artificial Neural Networks or Decision Trees a solution besides regularization or noise injection is construction of many such weak classifiers and combination within so called ensembles. Two of the most popular methods are Bagging and Boosting [8]. Within the first random bootstrap replicates of the training set are built for learning with several instances of the same classifier. A simple majority vote is fulfilled in the final decision process. In Boosting the classifiers are constructed iteratively on weighted versions of the training set. Thereby erroneously classified objects achieve larger weights to concentrate on hardly separable instances. Also a majority vote, but based on the weights leads to the final result.

However, these methods both use only instances of the same classifier. If we strive to combine advantages of diverse classifiers, Stacking is an alternative. Hereby several outputs of diverse instances are fused. In [8] StackingC as improved variant is introduced, which includes classifier confidences, e.g. by Maximum Linear Regression. It is further shown that StackingC can simulate most ensemble learning schemes, making it the most general and powerful representative of its kind. One major question however is the choice of the most suitable base classifiers for the ensembles. In [8] two optimal sets built of seven and four classifiers are introduced. However, the performance with the smaller set shows similar results at less computational effort for training. We chose a slightly changed variant of their smaller proposed

set, which delivered better accuracy as seen in Figure 2. Therein results on the various tasks are presented with StackingC, Bagging, Boosting and selected base-classifiers are shown. However, due to space limitations we can provide only a very brief introduction of the latter in the ongoing. A comprehensive description is available in [9]. The major drawback of the firstly selected and well known, though rather simple, Naïve-Bayes (*NB*) classifier is the underlying assumption that feature values are uncorrelated given the class and no hidden attributes may influence the estimation. In general, it is a very trivial variant of a Bayesian Network as introduced in section 5. Another simple variant is a k-Nearest-Neighbour instance basedclassifier with Euclidean distance (*kNN*). Support Vector Machines (*SVM*) show a high generalization capability due to their structural risk minimization oriented training. In this evaluation we used a couple-wise decision for multi-class discrimination and a polynomial kernel. As Decision Tree we chose an unpruned C4.5. In general these consist of a simple structure where non-terminal nodes represent tests on one or more features and terminal nodes reflect decision outcomes.

The following tests have been carried out on the 2,440 sample acoustic dataset described in section 2. The evaluation was done in a "leave-one-speaker-out" manner. Only mean performance is shown as the standard deviation throughout cycles never exceeded 2%. The effect of feature reduction is also shown. While some performances initially increase, losses for all classifiers were observed at higher reduction rates.

| **Accuracy [%]** | **276 dim.** | **100 dim.** |
|---|---|---|
| NB | 37.84 | 44.59 |
| kNN, k=1 | 63.51 | 68.91 |
| SVM, p=2 | 50.01 | 70.27 |
| C4.5 | 50.00 | 44.59 |
| Bagging C4.5 | 64.87 | 54.05 |
| Boosting C4.5 | 63.51 | 67.56 |
| StackingC SVM NB C4.5 kNN | **63.51** | **71.62** |

Figure (2): Accuracy of single classifiers and ensembles acoustic classification

The next table deals with reduced emotion sets.

| **Emotion Set** | **Accuracy [%]** |
|---|---|
| MPEG4+Neutral | 71.62 |
| Anger, Neutral, Surprise | 84.85 |
| Surprise, Neutral | 90.91 |
| Anger, Neutral | 97.12 |

Figure (3): Mean Performances for reduced emotion sets, Leave one speaker out evaluation

The optimal classifier configuration was each used and the tests ran on the same database as in the previous table. A significant increase in performance can be observed when focusing on the discrimination of one emotional state and a neutral state. However, this may be already sufficient in many cases.

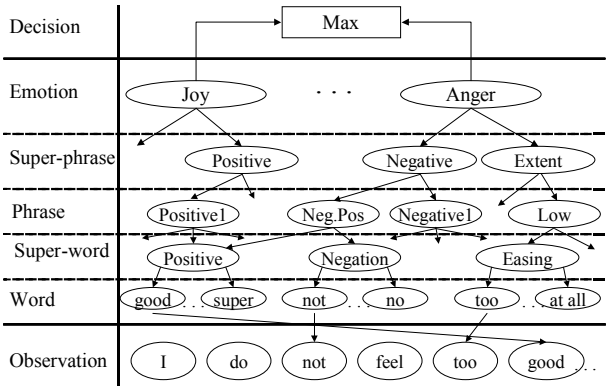
## 5. SPOKEN CONTENT ANALYSIS

As acoustic analysis shows significant speaker dependence, we strive to improve on independence by the constructive integration of content analysis. Automatic Speech Recognition *(ASR)* units nowadays show good performance for different unknown speakers. It only remains questionable whether a general learned semantic model applies sufficiently for new speakers. In general only a small amount of user utterances will contain emotional information. Even if an utterance carries information about the actual user emotion, it will in most cases be only fragments of the complete utterance. Therefore a spotting approach seems a must in search for emotional keywords or phrases in natural language. As a basis we use a standard Hidden-Markov-Model-based ASR engine with zero-grams as language model. It provides the n-best hypotheses including single word confidences of its estimation of the spoken content. As a mathematical background for the spotting we chose Bayesian Networks *(BN)* for their capability to handle uncertain and incomplete information. However, we will not deal with further approaches as vector space modeling or n-grams herein. In this paper we can provide only a very brief insight in the theory of BN, which enjoy growing popularity in pattern recognition and knowledge modeling tasks. Each network consists of a set of nodes related to state variables $X$, consisting of a finite set of states. The nodes are connected by directed edges expressing quantitatively the conditional probabilities of nodes and their parent nodes. A complete representation of the network structure and conditional probabilities is provided by the joint probability distribution. Let $N$ denote the total of random variables, and the distribution can be calculated as:

$$P(X_1,...,X_N) = \prod_{i=1}^{N} P\left(X_i \mid \text{parents}(X_i)\right)$$

Methods of interfering the states of some query variables based on observations regarding evidence variables are provided by the network. Similar to a standard approach to natural speech interpretation, the aim is finding the emotion hypothesis that maximizes the posterior probability of the word sequence given the acoustic observation. Each emotion is modeled in its own network. The root probabilities are distributed equally in the initialization phase and resemble the priors of each

emotion. If the emotional language information interpretation shall be used stand-alone, a maximum likelihood decision takes place. Otherwise the root probability for each emotion is fed forward to a feature-level fusion with the acoustic features. In four lower levels a clustering from words to super-words, phrases, super-phrases, and finally emotions takes place as can be seen in the following figure.




*"... I do not feel too good at all ..."*

Figure (4): Principle of Belief network based phrase spotting

On the word layer the evidences are fed into the net according to the word confidences of actually observed words. As we integrate the confidences of the ASR hypotheses the traditionally certain evidences are extended as uncertain evidences. The quantitative contribution $P(e_j|w)$ of any word $w$ to the belief in an emotion $e_j$ is calculated in a training phase by its frequency of occurrence under the observation of the emotion on the basis of the speech corpus described in section 2. The training and evaluation was fulfilled in a five fold stratified cross validation resulting in 302 test phrases per run. Mean performance was observed 64.8% correct assignment within the seven named emotions. Thereby, the absence of any emotionally co-notated terms or phrases, leading to indifferent root probabilities of the BN-models, has been considered as the neutral state. In a fusion within one feature vector consisting of the acoustic features and the linguistic estimates as further features the accuracy can be raised, as firstly reported in [10]. On the second described set of 1,507 sample utterances of movie and automotive interaction dialogs turns acoustic only reached 90.21 %. By inclusion of linguistic analysis this could be raised by another absolute 3.5%.

## 6. CONCLUSIONS

Within this paper we demonstrated improved emotion recognition by early fusion of acoustic and linguistic analysis. Ensemble classification helped to slightly improve overall performance. Still computational effort is considerably increased thereby. Considering speaker independent recognition high accuracy of 71.62% for the MPEG-4 set with added neutrality and up to 97.12% for the single recognition of anger can be reported. Feature reduction helped to decrease extraction time and improve recognition performance initially. However, too high reduction rate worsened the overall performance.

In our future work we focus on genetic feature generation and multi-task learning in speech emotion recognition for higher recognition performance.

## 7. REFERENCES

[1] M. Pantic, L. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," Proc. of the IEEE, Vol. 91, pp. 1370-1390, Sep. 2003.

[2] L. Devillers, L. Lamel, "Emotion Detection in Task-Oriented Dialogs," *Proc. ICME 2003*, Vol. III, USA, pp. 549-552, 2003.

[3] C. M. Lee, R. Pieraccini, "Combining acoustic and language information for emotion recognition," *Proc. ICSLP 2002*, Denver, CO, USA, 2002.

[4] B. Schuller, G. Rigoll, M. Lang, "Hidden Markov Model-Based Speech Emotion Recognition," *Proc. ICASSP 2003*, Vol. II, Hong Kong, China, pp. 1-4, 2003.

[5] P. Pudil, J. Novovičová, J.Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15/11, pp. 1119–1125, Nov. 1994.

[6] B. Schuller, et al., "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture," *Proc. ICASSP 2004*, Vol. 1, Canada, pp. 577-580, 2004.

[7] V. Petrushin, "Emotion in Speech, Recognition and Application to Call Centers," *Proc. ANNIE '99*, 1999.

[8] A. Seewald, *Towards understanding stacking – Studies of a general ensemble learning scheme*, PhD-Thesis, TU Wien, 2003.

[9] I. H. Witten, E. Frank, *Data Mining, Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, pp. 133, 2000.

[10] R. Cowie et al., "What a neural net needs to know about emotion words," *Computational Intelligence and Applications*, World Scientific&Engineering Society Press, pp. 109-114, 1999.