Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles

Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
{sch, mur, lg, ri}@mmk.ei.tum.de

10.21437/Interspeech.2005-379

Abstract

Herein we present a comparison of novel concepts for a robust fusion of prosodic and verbal cues in speech emotion recognition. Thereby 276 acoustic features are extracted out of a spoken phrase. For linguistic content analysis we use the Bag-of-Words text representation. This allows for integration of acoustic and linguistic features within one vector prior to a final classification. Extensive feature selection by filter- and wrapper based methods is fulfilled. Likewise optimal sets via SVM-SFFS and single feature relevance by information gain ratio calculation are presented. Overall classification is realised by diverse ensemble approaches. Among base classifiers Kernel Machines, Decision Trees, Bayesian classifiers, and memory-based learners are found. Acoustics only tests ran on a database comprising 39 speakers for speaker independent accuracy analysis. Additionally the public Berlin Emotional Speech database is used. A further database of 4,221 movie related phrases forms the basis of acoustic and linguistic information analysis evaluation. Overall remarkable performance in the discrimination of seven discrete emotions could be observed.

1. Introduction

The importance of emotion recognition for improved and natural future human computer interaction is commonly agreed [1]. Speech analysis is among the most promising information sources considering emotion recognition besides mimic, physiological or context data analysis. However, speech should be analyzed considering both: prosodic cues and the spoken content itself. A growing interest in the latter inclusion of verbal cues can be observed at the time, and a number of systems already exist that are capable of linguistic information processing in view of affect [1,2,3]. Within works that combine these two aspects improved accuracy for inclusion of both sources over each single one is reported [4,5,6]. The fusion is mostly realized within a post stage in a late semantic manner. However, inclusion prior to classification seems reasonable, as more information is saved before the final decision. We therefore aim at linguistic analysis by features that may easily be integrated with acoustic features in one vector. We chose the Bag-of-Words representation well known in document retrieval for this purpose. Considering acoustic information it is mostly reported that speaker dependent recognition leads to far better results than speaker independent modeling. 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.

Dealing with classification methods no unity can be found so far [7]. Within this work we concentrate on use of ensembles of classifiers in order to cope with biased training due to the comparably small training sets used in speech emotion recognition and the growing dimensionality by inclusion of novel features, especially linguistic content information. Boosting was already successfully applied in speech emotion recognition in [8]. While methods as Boosting or Bagging stabilize single classifiers, we introduce StackingC within speech-based affect recognition to combine the power of diverse classifiers for the final decision. In [9] it is shown that StackingC, a variant of Stacking, is usually the best choice considering maximum performance applying ensembles. The results using diverse single classifiers are also provided as a basis of comparison.

Considering the choice of the right features sparse analysis of single feature relevance by means of filter or wrapper based evaluation has been fulfilled, yet. Features are mostly reduced by means of the well known Principal Component Analysis *(PCA)* and selection of the obtained artificial features corresponding to the highest eigen-values [10]. As such reduction still requires calculation of the original features we compare it to a real elimination of original features within the set. As search function within feature selection *(FS)* we apply a Support Vector Machine *(SVM)* based Sequential Forward Floating Search *(SFFS)* [11], which is known for its high performance. Thereby the evaluation function is the target classifier which optimizes the features as a set rather than finding single features of high performance. Additionally we show single feature relevance by calculation of the Information Gain Ratio *(IGR)* of each feature.

The paper is structured as follows: Section 2 deals with databases used. In section 3 and 4 acoustic and linguistic features are introduced. Section 5 consists of feature selection results. In the following section 6 we deal with classification, especially ensemble construction. Finally overall results and conclusions are found in sections 7 and 8.

2. Databases

For tests on speaker independent recognition we chose a database consisting of 39 speakers, three of them female named *EMO-SI* in the ongoing. Per speaker 70 samples have been chosen resulting in 2,730 samples in total. The samples are evenly distributed among seven commonly used emotional states, namely anger, disgust, fear, joy, sadness, surprise and neutrality. This emotion set has been chosen for comparability reasons, as it is far spread. However, other models as an arousal valence plane exist. The samples resemble short phrases of car interaction dialogs, provoked emotions in usability studies, and additionally acted ones of the same speakers as introduced in [6,12]. Spontaneous samples have been annotated by the speakers afterwards. The intent is to obtain a high number of speakers for model construction considering speaker independent recognition. Mixing spontaneous and acted emotions seems no drawback here, as we want to recognize emotional states of both a kind. However, we will not deal with differences between those two types within this work.

In order to provide results on classifier and feature selection on a public corpus we decided for the Berlin Emotional Speech database *(EMO-DB)* [13]. It consists of 816 phrases in total. The emotion set resembles ours, besides an exchange of surprise in favor of boredom. 10 German sentences of emotionally undefined content have been acted in these emotions by 10 professional actors, 5 of them female. Throughout perception tests by 20 probands 488 phrases have been chosen that were classified as more than 60% natural and more than 80% clearly assignable.

For linguistic feature analysis we chose textual content 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. A wide bandwidth of genres, i.e. Science-Fiction, Comedy, Drama, Horror, and Fantasy could be covered in order to include all desired emotions. The utterances were annotated phrase-wise by two test persons and 1,144 phrases consisting of 7.0 words in average with identical labeling could be obtained. If a sentence was labeled for two or more emotions by both labelers it was included in the corpus with both annotations. For model construction the set was supplemented by 3,077 phrases of the movie domain labeled accordingly. In order to cover as many regular terms as possible, enlargement of the dictionary was fulfilled by emotional labelling of the 10,000 most frequent terms in the English language [14]. Finally the balanced affective word list [15] was included. The emotional vocabulary was then built by storing each new word and counting the total frequency of occurrence for each of the 2,234 disjunctive terms within the tagged emotion. For final tests of acoustic and linguistic feature inclusion the phrases were re-acted and recorded as single utterances in an anechoic chamber. The recording was fulfilled by use of a condenser microphone AKG-1000S MK-II over a long period to avoid anticipation effects of the three actors in total. We decided for re-acting as the cut-scenes of the movies tended to include too much background noise or over-laid vocals. This final database will be named *EMO-AL*.

3. Acoustic Features

In former works [2] we compared static and dynamic feature sets for the prosodic analysis and demonstrated the higher performance of derived static features. Features are therefore derived phrase-wisely. This seems also reasonable as we need to synchronize linguistic analysis and acoustic analysis for the latter inclusion within one feature vector. As an optimal set of such global features is broadly discussed [3,4,5], we considered an initially large set of 276 acoustic hi-level features which cannot all be described in detail here. However, the target is to become utmost independent of the spoken content and ideally also of the speaker, but model the underlying emotion. The feature basis is formed by the raw contours of zero crossing rate *(ZCR)*, pitch, first seven formants, energy, spectral development, and Harmonics-to-Noise-Ratio *(HNR)*. The following table shows the distribution of features among their general type. Thereby duration based features rely on common bi-state dynamic energy threshold segmentation and voicing probability.

Table 1: Distribution of the features

Type	**Pitch**	**Energy**	**Duration**	**Formant**
Number	12	11	5	105
Type	**HNR**	**MFCC**	**FFT**	**ZCR**
Number	3	120	17	3

20 ms frames of the speech signal are analyzed every 10 ms using a Hamming window function. As pitch detection algorithm we apply an average magnitude difference function. The values of energy resemble the logarithmic mean energy within a frame. For spectral development we use 15 MFCC coefficients and a FFT-spectrum. Low-pass SMA filtering smoothes the raw contours prior to the statistical analysis. The higher level features are then subsequently derived and normalized. Examples can be found in table 2.

4. Linguistic Features

Basing on the output hypothesis of a state-of-the-art HMM-based ASR-engine spoken content analysis can be included in the overall model. The aim here is to enable an integration of acoustic and linguistic features in one vector. As a consequence single linguistic features are demanded. The so called Bag-of-Words method applied in automatic document categorization is therefore chosen. Thereby each word in the vocabulary adds a dimension to the linguistic vector representing the logarithmic term frequency within the actual utterance known as *logTF*. This frequency is normalized by the phrase length. As a high dimensionality may decrease the performance of the classifier and flexions of terms reduce performance especially within small databases methods of feature reduction seem mandatory. We first consider the most natural form by use of a stop-list obtained by expert-knowledge. It consists of ignorable words due to their lack of

affective information. These have to be chosen carefully, as it may not be easily visible if a word possesses an emotional connotation. We therefore stopped mostly articles, names, etc. resulting in 93 stop-terms. Additionally by stemming words of the same stem are clustered. This also reduces dimensionality while in general directly increasing performance. This comes as hits within an utterance are crucial and their number increases significantly if none is lost due to minor word differences as plural forms or verb conjunctions. Further reduction was obtained by filter-based feature selection as described in the following section. We decided for IGR calculation here due to its low computation efforts compared to SVM-based FS.

5. Feature Selection

Selection of features is important as it saves computation time considering real-time processing. Furthermore some classifiers are susceptible to high dimensionality. Therefore search for the right features seems mandatory. We chose SVM-SFFS within acoustic feature selection for the reasons mentioned in section 1. The search is performed by forward and backward steps eliminating and adding features in a floating manner to an initially empty set. Table 2 shows the top 30 found acoustic features by SVM SFFS run on EMO-DB with their IGR. It may be surprising that MFCC based features are top ranked.

Table 2: Top 30 acoustic features by SFFS

Rank	IGR	Feature
1	0.4722	MFCC3 Mean
2	0.1529	MFCC7 Mean
3	0.2933	MFCC14 Mean
4	0.2614	MFCC13 Std. Dev.
5	0.3182	MFCC2 Mean
6	0.2788	MFCC6 Mean
7	0.2186	Spec. Flux Max
8	0.2185	δCentroid Mean
9	0.2119	δF0 Max
10	0.0851	F2 Bandwidth Mean
11	0.3732	F0 Std. Dev.
12	0.1065	Centroid Max
13	0.1670	δCentroid Std. Dev.
14	0.2292	F1 Mean
15	0.2224	Spec. Flux Std. Dev.
16	0.3635	δF0 Mean
17	0.1621	δδMFCC5 Mean
18	0.1921	δδMFCC1 Mean
19	0.1319	MFCC9 Mean
20	0.2700	MFCC15 Std. Dev.
21	0.2233	HNR Mean
22	0.2112	Spec. Flux Mean
23	0.2669	MFCC6 Std. Dev.
24	0.1950	Roll-off Point Std. Dev.
25	0.0001	F7 Min
26	0.2642	MFCC13 Mean
27	0.1199	F2 Mean
28	0.0000	δRoll-off Point Max
29	0.1385	δMFCC9 Mean
30	0.1557	MFCC12 Std. Dev.

This is may be due to the database used.

6. Ensemble Classification

Emotion samples, especially spontaneous ones, are hard to obtain. This is especially true when aiming at a high number of evenly distributed samples among emotions of diverse speakers. Having such relatively small training sample sizes compared to the dimensionality of the data, a high danger of bias due to variances in the corpus is present. In order to improve instable classifiers as neural nets 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, firstly introduced in emotion recognition in [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 AdaBoosting 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 weights, leads to the final result. A combination of these two is MultiBoosting, where Wagging, a Bagging variant that ensures use of all training samples is applied in combination with AdaBoosting. However, these methods use only instances of the same classifier type.

If we strive to combine advantages of diverse classifiers Stacking is an alternative. Hereby several outputs of diverse instances are combined. In [9] StackingC as improved variant is introduced, which includes classifier confidences e.g. by Maximum Linear Regression. One major question however is the choice of right base classifiers. In [9] an optimal set with four classifiers is introduced. We use a slightly changed variant of their set, which deliveres better results in our case. Accuracies obtained with various base-classifiers and constructed ensembles are shown in the following table. The major drawback of the firstly selected well known base classifier Naïve-Bayes *(NB)* is the basing assumption that features are independent given class, and no latent features influence the result. Another rather trivial variant is a memory-based classifier using Euclidean distance *(1NN)* [16]. Support Vector Machines *(SVM)* show a high generalization capability due to their structural risk minimization oriented training. In this evaluation we use a couple-wise decision for multi-class discrimination and a polynomial kernel. As for Decision Trees we chose *C4.5*. In general these are a simple structure where non-terminal nodes represent tests on features and terminal nodes reflect decision outcomes.

7. Recognition Results

Considering speaker independent recognition we apply leave-one-speaker-out *(LOSO)* evaluation for the discrimination of an emotional and a neutral state run on the corpus EMO-SI described in section 2. The mean accuracy over all emotions and speakers thereby reached 90.61% using SVM. With optimal configuration the accuracy dropped to 88.26% when enhancing the set by a third emotion. Having all emotions in the set the mean performance for speaker independent recognition dropped to 71.07% with a maximum of 85.71%

for one specific speaker. Compared to this 92.72% in average were observed for speaker dependant recognition.

Classifier comparison tests have been carried out on the EMO-DB database in a 10-fold stratified cross-validation. Table 3 shows the results obtained thereby.

Table 3: Classifier comparison run on EMO-DB

Accuracy [%]	All 276 features included	Top 75 by SVM SFFS
NB	73.57	73.98
1NN	63.52	75.82
SVM	**84.84**	**87.50**
C4.5	61.07	61.48
Bagged C4.5	70.70	74.80
AdaBoosted C4.5	72.34	74.59
MultiBoosted C4.5	72.54	74.59
StackingC MLR NB 1NN C4.5	75.41	79.92
StackingC MLR NB 1NN SVM C4.5	76.23	80.53

Tests on the dataset EMO-AL were fulfilled only speaker dependently. Thereby 90.30% accuracy could be reached in average having only acoustic attributes in the feature vector. Using linguistic features only, 65.07% correct assignment to the underlying emotion could be obtained. The final inclusion of both feature types improved performance by absolute 3.51%.

8. Conclusions

Within this work we introduced speech emotion recognition combining acoustic and linguistic features in one vector. Speaker independent discrimination between six basic emotions each and an emotionally neutral state could be realized with mean accuracy of 90.30%. Speaker dependent recognition proved much more reliable: Seven emotions could be discriminated at a time with mean accuracy of 92.72%. Further more we showed results on the public Berlin Emotional Speech Database. 87.50% recognition rate can be reported for the discrimination of seven emotions at a time. By feature selection techniques the most relevant features as a set with single feature relevance could be shown. Acoustic feature reduction helped to improve overall performance and save extraction effort. As for base classifiers we obtained the best results with SVMs within these experiments. By construction of ensembles of classifiers the overall performance could be increased in most cases. Multiboosting however did not further improve performance compared to AdaBoosting and Bagging. StackingC helped to improve robustness unless SVM were included in the set throughout the test-runs. However, considerable increase in computation time remains a drawback at little improvement in accuracy. While using only linguistic information fell clearly behind acoustic information only use, the overall performance could be raised by 3.51% in total by their integration.

In our future works we aim at detailed investigation of the effects of emotionally distorted vocalization on automatic speech recognition, which is crucial for the linguistic analysis. Furthermore we aim at inclusion of acoustic features on different time levels, as word and phrase based features. Finally we hope to improve by genetic feature generation and multitask learning.

9. References

[1] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., "Emotion recognition in human-computer interaction," *IEEE Signal Processing magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] Cowie, R.; Douglas-Cowie, E.; Apolloni, B.; Taylor, J.; Romano, A.; Fellenz, W.: "What a neural net needs to know about emotion words," *Computational Intelligence and Applications*, World Scientific&Engineering Society Press, pp. 109-114, 1999.

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

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

[5] Chuang, Z.; Wu, C., "Emotion Recognition using Acoustic Features and Textual Content," *Proc. ICME 2004*, Taiwan, 2004.

[6] Schuller, B.; Jimenez Villar, R.; Rigoll, G.; Lang, M., "Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition," *Proc. ICASSP 2005*, Philadelphia, PA, USA, 2005.

[7] Pantic, M; Rothkrantz, L..: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. Proccedings of the IEEE, Vol. 91, pp. 1370-1390, Sep. 2003.

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

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

[10] Ververidis, D.; Kotropoulos, C.; Pitas, I., "Automatic Emotional Speech Classification," *Proc. ICASSP 2004*, pp. 593-596, Montreal, Canada, 2004.

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

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

[13] http://pascal.kgw.tu-berlin.de/emodb, 05.05.2005.

[14] Quasthoff, U., "Tools for Automatic Lexicon Maintenance, Acquisition, Error Correction, and the Generation of Missing Values," *Proc. ELRA 1998*, pp. 853-856, 1998.

[15] Siegle, G., *The Balanced Affective Word List Project*, http://www.sci.sdsu.edu/cal/wordlist, 1995.

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