META-CLASSIFIERS IN ACOUSTIC AND LINGUISTIC FEATURE FUSION-BASED AFFECT RECOGNITION

Björn Schuller, Raquel Jiménez Villar, Gerhard Rigoll, and Manfred Lang

Institute for Human-Machine Communication
Technische Universität München
(schuller | rigoll | lang)@ei.tum.de

ABSTRACT

Within this work we suggest a novel approach to affect recognition based on acoustic and linguistic analysis of spoken utterances. In order to achieve maximum discrimination power within robust integration of these information sources a fusion on the feature level is introduced. Considering classification we use meta-classifiers as StackingC and Boosting for a stabilized performance and combination of classifiers within ensembles. Extensive comparison of diverse base-classifiers comprising among others Support Vector Machines, Neural Networks, stochastic models, and Decision Trees will be fulfilled. 381 acoustic features are extracted and their relevance is calculated by a Sequential Forward Floating Search in comparison to reduction by a Principal Component Analysis. Several variants for linguistic feature calculation are described and ranked including bunch-of-words, n-grams, salience, and mutual information. Furthermore reduction by stopping and stemming or filter-based selection methods is evaluated reducing 2,334 linguistic features. Seven discrete emotions described in the MPEG-4 standard are recognized within an existing recognition engine. The presented results base on two large databases of 4,336 acted and real emotion samples from movies, chat and car interaction dialogues. A significant gain and an outstanding overall performance are observed by this novel fusion and use of ensembles.

1. INTRODUCTION

A variety of approaches towards automatic emotion recognition were presented since the research activities started in the late last decade. Today we are all aware of the great importance of the integration of emotional aspects as the next step towards more natural human-machine interaction [6]. Growing interest in the extraction out of diverse modalities among which speech is found on top-level can be observed at the time.

Large numbers of diverse acoustic hi-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. 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 [3]. 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*) [8], which is known for its high performance. Thereby the evaluation function is the classifier, in our case SVMs as described in section 5, 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. 381 static acoustic high-level features form the basis for this analysis.

However, already in early speech-based emotion recognition works estimation of the emotion by the spoken content was analyzed [1]. Nowadays it seems to be broadly considered reasonable that integration of such linguistic information improves the overall performance [3][5], while suggested methods vary strongly. Examples are uni-grams [2], calculation of emotional salience [3], rule-based decision, training of neural networks [1], or use of Bayesian Networks as in our former works [5]. Language information so far is not included on the feature level, but rather in a post-stage fusion. A drawback thereby is that information for a maximum discrimination is already lost. Additionally the evaluation of the gain considering integration of spoken content information can only be judged in total. In order to achieve an early feature fusion and enable direct relevance measurement by FS we therefore decided to include language features within the acoustic vector.

Dealing with classification methods also no unity can be found so far [6]. 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. Boosting was already successfully applied in speech emotion recognition in [7]. 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 [10] 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.

The paper is structured as follows: In section 2 we describe the databases in detail and show construction of an affective vocabulary. In sections 3 and 4 we introduce our acoustic and respectively linguistic features. In section 5 meta-classification will be discussed. Finally results of the overall features are shown, and conclusions are drawn.

2. EMOTIONAL DATABASES AND VOCABULARY

The emotions used resemble the far spread MPEG-4 set, namely joy, anger, disgust, fear, sadness, surprise and added neutrality. Within the acoustic feature selection and classifier evaluation the emotional speech corpus EMO-CAR collected in the framework of the FERMUS III project was used [4]. This allows for direct comparison of results introduced in our former works presented in [5]. It consists of 2,829 emotional samples of car-user-interface interaction dialogues. In total 13 speakers, one female, are contained within the database.

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 main corpus. The textual content was taken from movie scripts of seven U.S. American movies from the years 1977 till 1999. Namely these are *Alien, Annie Hall, Five Easy Pieces, Notting Hill, Scream, Ten things I hate about you,* and *Toy Story*. Genres include Sci-Fi, Comedy, Drama, Horror and Fantasy and have been selected in order to cover all aimed at emotions. 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.

The vocabulary for the linguistic analysis bases on 3,396 phrases with further movie excerpts, and web-chat statements included. In order to cover as many regular terms as possible enlargement of the dictionary was also fulfilled by emotional labeling of the 10,000 most frequent terms in English language [9]. Finally the balanced affective word list [11] 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. Thereby two different variants have been considered for the calculation of posteriors: Once a Laplace-estimation assuming an equal initial distribution among emotion classes was used denoted as $p(e_i|w)$ of the probability of emotion e_i given the word w, and once all posteriors were initially set to zero, denoted as $p^*(e_i|w)$ in chapter 4.

3. ACOUSTIC FEATURES

In former works [4] we compared static and dynamic feature sets for the prosodic analysis and demonstrated the higher performance of derived static features. As the optimal set of global static features is broadly discussed [6], we considered an initially large set of 381 features comprising features which cannot be described in detail here. We rather concentrate on the basic extraction of top ranked attributes as the results of a FS by SVM-SFFS show a saturation point at 33 features. The following figures 1 and 2 present results of the feature reduction and an exemplary excerpt of the reduced set. Figure 1 also shows that a true reduction of features seems no drawback compared to the reduction by PCA. The feature basis is formed by the raw contours of the signal, pitch, energy and voicing probability. As within acoustic features the target is to become utmost independent of the spoken content, which is only respected in the linguistic features, only sparse spectral features are extracted. 20 ms frames of the speech signal are analyzed every 10 ms using a Hamming window function. The values of energy resemble the logarithmic mean energy within a frame. As pitch detection algorithm we apply an average magnitude difference function. Low-pass symmetrical moving average filtering smoothes the raw contours prior to the statistical analysis. Higher level features are subsequently derived and normalized by their standard deviation and mean. Silence duration is calculated using common bi-state dynamic energy threshold segmentation. Durations of voiced sounds rely on the voicing probability.

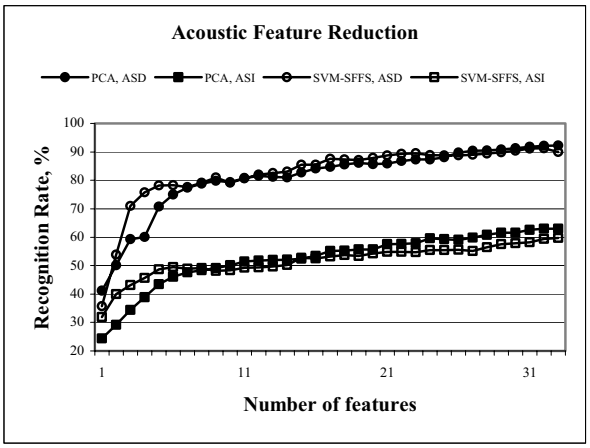

Figure (1): Acoustic feature reduction by PCA and SVM-SFFS for speaker (in-) dependent recognition ASI / ASD

Rank	GainR	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	0.201	Pitch standard deviation
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 points
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 points
18	0.050	Energy mean of fall-time
19	0.051	Energy mean dist. of reversal points
20	0.035	Energy relative maximum

Figure (2): First 20 acoustic feature ranks by a SFFS with SVM-wrapper and Gain Ratio GainR showing single feature relevance

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. In an earlier work [5] we introduced a spotting-based approach for emotional key-phrases by a Bayesian Network. The output was fused with acoustic feature processing by a Neural Net. However, 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 bunch-of-words method applied in automatic document categorization was chosen as a starting point. Thereby each word in the vocabulary adds a dimension to the linguistic vector representing the term frequency within the actual utterance. 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. Stemming clusters words of the same stem, and 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 not lost due to minor word differences as plural form verb conjunctions. A further reduction of words was obtained either by filter-based FS or reduction by a PCA. Within FS we decided for information gain ratio calculation [12] due to its low computation efforts compared to SVM-based FS as used for the lower dimensional acoustic set. While the reduced sets by these methods both clearly fell behind, an interesting side effect is that Gain Ratio shows the most emotional words in the corpus. The 20 highest ranks can be seen in figure 3.

Rank	GainR	Term	Rank	GainR	Term
1	0.553	disgusting	11	0.237	wonderful
2	0.465	throw	12	0.230	sad
3	0.465	yuck	13	0.229	cool
4	0.446	dirty	14	0.229	christ
5	0.276	face	15	0.222	bitch
6	0.272	lucky	16	0.220	beautiful
7	0.264	perfect	17	0.215	jesus
8	0.264	delighted	18	0.213	thank
9	0.259	afraid	19	0.190	glad
10	0.243	great	20	0.174	sorry

Figure (3): 20 highest ranked terms and their Gain Ratio GainR

Significantly better results were obtained by a reduction to seven dimensions. Thereby eleven variants were considered for calculation of the features directly corresponding to the emotion. The posteriors $p(e_i|w)$ described in section 2 form the basis of computation as shown in figure 4, where selected variants are shown. The table also shows the maximum performance obtained with each feature variant using SVMs as described in section 5. The solution on rank one resembles uni-grams as suggested in [2], and the rank nine version is applied in [3] besides that SVMs are used each instead of a maximum decision. The alternative ranked four corresponds to mutual information. In the table the following two abbreviations are used, where the first corresponds to the salience (*sal*) as introduced in [3]:

$$sal(w) = \sum_{i=1}^{7} p(e_i|w) \cdot ld\left(i(w,e_i)\right) \text{ and } i(w,e_i) = \frac{p(e_i|w)}{p(e_i)}$$

Rank	Rate,%	Dim.	Feature Type
1	73.9	7	$\sum_{w\in U} \log_{10}\left(i^*(w,e_i)\right)$
4	69.8	7	$\sum_{w\in U} \left(p^*(e_i\|w)\cdot\log_{10}\left(i^*(w,e_i)\right)\right)$
5	69.4	7	$\sum_{w\in U} \left(p^*(e_i\|w)\cdot sal^*(w)\right)$
6	62.0	7	$\prod_{w\in U} i(w,e_i)$
7	60.2	7	$\sum_{w\in U} \left(p(e_i\|w)\cdot sal(w)\right)$
9	39.5	7	$\prod_{w\in U} p(e_i\|w)$
10	36.4	1853	Bunch-of-words, Stop&Stem
11	35.8	2334	Bunch-of-words
12	33.2	1000	Bunch-of-words, Stop&Stem, FS

Figure (4): Selected linguistic feature set variants with mean 3-fold performance using SVMs, run on 1,507 samples

5. META-CLASSIFICATION

With relatively small training sample sizes 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 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 [10]. 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 combined. In [10] StackingC as improved variant is introduced, which includes classifier confidences e.g. by Maximum Linear Regression. It is further shown that by StackingC most ensemble learning schemes can be simulated, making it the most general and powerful ensemble learning scheme. One major question however is the choice of right base classifiers for the ensembles. In [10] 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 use a slightly changed variant of their set as seen in figure 5, which delivered better results. Results on the various tasks

applying StackingC, Bagging, Boosting and selected base-classifiers are shown. However, we can provide only a very brief introduction of the latter in the ongoing. A comprehensive description is available in [12]. The major drawback of the firstly selected well known rather simple Naïve-Bayes (*NB*) classifier is the basing assumptions that features are independent given class, and no latent features influence the result. Another rather trivial variant is a nearest distance measurement based on entropy calculation (*ND*). The further considered Neural Nets are renowned for their non-linear transfer functions, self-contained feature weighting capabilities and discriminative training. A Multi Layer Perceptron (*MLP*) with the number of input neurons equaling the number of input features, a sigmoid transfer function in the hidden layer, and 7 output neurons for each emotion was used. 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 are a simple structure where non-terminal nodes represent tests on features and terminal nodes reflect decision outcomes. The attributes are ordered by their gain ratio.

Classifier	ASI,%	ASD,%	LIN,%
NB	51.1	86.3	73.4
ND	73.8	86.9	69.5
SVM	76.1	91.0	73.9
C4.5	63.7	82.4	75.0
Bagging C4.5	75.2	86.9	76.3
Boosting C4.5	76.0	92.7	74.2
MLP	73.2	90.6	73.3
Bagging MLP	73.8	92.5	75.3
Boosting MLP	73.6	92.7	74.3
StackingC MLR NB ND SVM C4.5	**76.4**	**92.9**	**76.8**

Figure (5): Performances of single classifiers and ensembles

All tests have been carried out on the datasets described in section 2 by a three-fold stratified cross-validation. Only the mean performance is shown. The standard deviation throughout cycles never exceeded 2%. Acoustics only speaker dependent (A*SD)* and speaker independent (A*SI)* evaluations were each considered. Furthermore results for refinement of the performance on the linguistic only feature set ranked one in figure 4 by optimal classification is shown (*LIN*). Only results with optimal parameter configuration are shown.

6. FINAL RESULTS AND CONCLUSION

The final evaluation shows direct comparison for integration of linguistic features. The test ran in a 10-fold cross-validation with minor standard deviation on the described corpus. As StackingC proved most reliable in the prior runs it was chosen within this analysis. The acoustic features were analyzed speaker dependently and the mean performance was 90.3%. Evaluation based on linguistics reached 76.8% as already seen in figure 5. Within fusion 94.8% in average were reached. In an overall FS the linguistic features were ranked on places14-17, 19, 25, and 29. These results both clearly stress the importance of content analysis integration. This is especially true as language information in principal depends less of the speaker. The applied linguistic methods can also be used in text-based affect recognition. Summarized we could demonstrate the high gain achieved by a novel early feature fusion of acoustic and linguistic information in speech emotion recognition. Out of 381 acoustic features 20 most relevant could be presented. Further more 9 variants of linguistic features were shown. Selection of features showed similar performance than the often applied reduction by use of Principal Components with high eigen-values at less original extraction effort. Finally it could be shown that StackingC as classification method led to a maximum gain. The high performance achieved encourages dealing with recognition in noise in future experiments.

7. REFERENCES

[1] 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.

[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] 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.

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

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

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

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

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

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

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