Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers

Björn Schuller, Manfred Lang, Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München, D-80333 München, Germany
Email: (Schuller | Lang | Rigoll)@tum.de

Introduction

Automatic speech recognition can fail to a certain extent when confronted with emotionally distorted speech. Great efforts have been spent so far to cope with noise conditions or speaker's characteristics. Yet, adaptation to the emotional condition of the speaker could help to further improve the overall performance. In this respect we aim at a robust and reliable recognition of the speaker's emotional state by acoustic features only prior to speech recognition itself. Thereby we can load according emotional speech models. In this work we introduce an optimal feature set for this task selected by Sequential Floating Search Methods. The set comprises high-level prosodic features resembling utterance-wise statistic analysis of low-level contours as pitch, higher-order formants, energy, and spectral development. Within classification we apply ensemble classification as Stacking, Bagging, and Boosting.

Databases

Throughout this paper we chose a database consisting of 39 speakers, three of them female. Per speaker 70 samples have been chosen resulting in 2,730 samples in total. The samples are evenly distributed among seven emotional states, namely anger, disgust, fear, joy, sadness, surprise and neutrality. The set has been chosen for comparability reasons and in view of the target application. The samples resemble short phrases of car interaction dialogs, provoked emotions in usability studies, and additionally acted ones of the same speakers. 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, as adaptation of ASR models shall be enabled for both a kind. However, we will not deal with differences between those two types within this work.

Acoustic Features

Within acoustic features the target is to become utmost independent of the spoken content. As a higher goal we also aim at independence of the speaker. We do not include semantic analysis in view of the emotion herein, as in [1], as we consider emotion recognition prior to speech recognition. In former works [2] 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 [3, 4, 5], we considered an initially large set of 276 features comprising features which cannot be described in detail here. The feature basis is formed by the raw contours of the signal, pitch, formants, energy, spectral development, and voicing probability. 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. The basis of the spectral analysis is formed by FFT computation. Low-pass symmetrical moving average filtering smoothes the raw contours prior to the statistical analysis. Higher level features are subsequently derived and normalized. Thereby duration based features rely on common bi-state dynamic energy threshold segmentation and voicing probability.

Feature Selection

Large numbers of diverse acoustic hi-level features 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 and selection of the obtained artificial features corresponding to the highest eigen-values [3, 4, 5]. As such reduction still requires calculation of the original features we aim at a real elimination of original features within the set. As search function within feature selection *(FS)* we apply a Sequential Forward Floating Search *(SFFS)* [6], which is well known for its high performance. Thereby the evaluation function is the classifier, in our case Support Vector Machines *(SVM)*. This 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 in a floating manner to an initially empty set. For single feature relevance measurement we apply fast Information Gain Ratio (IGR) calculation.

SFFS Rank	IGR	Feature
1	0.112	Pitch area
2	0.105	F0 std. dev.
3	0.102	HNR std. dev.
4	0.101	F0 rel. position max
5	0.099	Spectral Flux std. dev.
6	0.095	δF0 max
7	0.091	F0 rel. Position min
8	0.090	F0 rel. max
9	0.088	F1 min
10	0.085	Zero Crossing Rate
11	0.084	Spectral Flux max
12	0.081	HNR mean

Table 1: Excerpt of the feature ranking

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 [7]. 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 [8] 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 remains the choice of right base classifiers. In [8] an optimal set with four classifiers is introduced. We use a slightly changed variant of their set, which delivered better results in our case. Accuracy 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 assumptions that features are independent given class, and no latent features influence the result. Another rather trivial variant is a nearest distance classifier based on entropy calculation *(K*)* [9]. 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 *C4.5*. In general these are a simple structure where non-terminal nodes represent tests on features and terminal nodes reflect decision outcomes.

Classifier	Accuracy, %
Naïve Bayes	81.22
K*	76.24
SVM	90.61
C4.5	77.38
Bagged C4.5	83.64
Boosted C4.5	84.63
StackingC MLR NB ND SVM C4.5	**91.45**

Table 2: Classifier comparison, LOSO

As we intend to have a robust, but speaker independent estimation in the first place, we concentrate on discrimination between an emotion and an emotionally neutral state at this point. We later on also show performances for discrimination between further emotional states and speaker dependent setups. All tests have been carried out on the dataset described in section 2 with leave-one-speaker-out *(LOSO)* evaluation for the recognition of anger.

Conclusion

Within this work we introduced speech emotion recognition as basis for adaptation in ASR. Speaker independent discrimination between six basic emotions each and an emotionally neutral state could be realized with mean accuracy of 89.76%. Speaker dependent recognition proved much more reliable: Seven emotions could be discriminated at a time with mean accuracy of 92.72%. Feature selection techniques helped to reduce dimensionality and choice of relevant features. Thereby SFFS FS outperformed IGR FS at low feature vector dimensionality. However, extraction effort of all original features could be saved compared to PCA based FS. By construction of ensembles of classifiers the overall performance could be increased. As base classifiers we obtained the best results with SVMs within these experiments. Considering meta classifiers StackingC proved the best choice compared to Boosting and Bagging. However, an considerable increase in computation time is remains a drawback at little improvement in accuracy. In our future works we aim at detailed investigation of the effects of emotionally distorted speech and its effects on ASR.

Literature

[1] 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.

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

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

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

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

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

[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.