Recognition of Spontaneous Emotions by Speech within Automotive Environment

Björn Schuller, Manfred Lang, and Gerhard Rigoll
Technische Universität München, Arcisstraße 21, D-80333 München, Deutschland,
Email: {sch|lg|ri}@mmk.ei.tum.de

Introduction

Within the car, recognition of emotion largely helps to design communication more natural. Speech interaction is here used more broadly today, and affective cues are contained within acoustic and linguistic parameters. However, we introduce novel concepts and results considering the estimation of a driver's emotion by focusing on acoustic information herein. As a database we recorded 2k dialog turns directed to an automotive infotainment interface during extensive usability studies. Speech recognition and natural language interpretation have thereby been realized once as a Wizard-of-Oz simulation, and once by actual recognition technology. Recorded utterances have been labelled using a closed set of four emotions, namely anger, confusion, joy, and neutrality. As acoustic features we apply a high number of prosodic, speech quality, and articulatory functionals derived by descriptive statistic analysis out of base contours as intonation, intensity, and spectral information. Self-learning feature generation and selection is employed to optimize complexity for the successive classification by Support Vector Machines. Semantic information is included by vector-space representation of the spoken content within an early feature fusion. Overall, high recognition performances can be reported for this task by the suggested approach.

Data-Collection

In order to test emotion recognition in the car we carried out a user study that led to the EA-CAR database: 2 female and 8 male German test-subjects aged 23.4a in average controlled an infotainment interface by natural speech. Speech functionality was simulated by a wizard in the first half of a 80 min session. In the second half actual ASR and NLU units were used. In total 2,022 phrases were recorded by a Yoga EM 240 condenser microphone in 16 bit, 11 kHz, within a genuine car. 45 interaction goals had to be fulfilled while driving in a simulation. The collected samples were labeled by three annotators, one female, aged 23a to 30a. A closed emotional set was used. Taking only phrases with full inter-label agreement leaves 775 phrases: 225 anger, 135 confusion, 25 joy, 390 neutrality. Due to its sparseness joy was excluded form the set. Recognition of these emotions helps to actively provide help to the driver in case of confusion or start error-recovery strategies in case of anger.

In order to compare these results we also include test-runs on the renowned *Berlin Emotional Spech Database* (EMO-DB) [1] and the *Danish Emotional Speech* (DES) database [2].

Feature Extraction

As a basis for feature generation we extract low-level contours of a whole phrase. We use state-of-the-art preprocessing of the audio signal: 20 ms Hamming-windowed frames are analyzed every 10 ms. For prosodic information we extract the contours of elongation, intensity, and intonation. We furthermore estimate durations of pauses and voiced syllables. Out of the elongation we calculate the zero-crossing-rate. By standard frame energy intensity information based on physical relations is included. Intonation is respected by auto-correlation-based pitch estimation. We thereby divide the speech signal correlation function by the normalized correlation function of the window function and search for local maxima besides the origin. Dynamic programming is used to back-track the pitch contour in order to avoid inconsistencies and reduce error form a global point of view. Finally, the named durations are estimated based on intensity considering pause duration, and voiced/unvoiced parts duration for syllable length based on intonation. In order to include voice quality information we also integrate the location and bandwidth of formants one to seven, harmonics-to-noise-ratio (HNR), MFCC coefficients well known in speech processing, and a perception conform dB-corrected FFT spectrum as basis for low-band energies -250 Hz and -650 Hz, spectral roll-off-point, and spectral flux. Formant location and bandwidth estimation is based on resonance frequencies in the LPC-spectrum of the order 18. Back-tracking is used here, as well. The HNR is calculated as logHNR to better model human perception. It also bases on the auto correlation of the input signal. The further spectral features are often used in Music Retrieval, and are included to observe their relevance within this task. Finally, as articulatory features we use the spectral centroid.

In former works we showed the higher performance of derived functionals instead of full-blown contour classification [3]. We therefore use systematic generation of functionals out of multivariate time-series by means of descriptive statistics. First of all the contours are smoothed by symmetrical moving average filtering with a window size of three, to be less prone to noise. Successively, speed (∂) and acceleration (∂^2) coefficients are calculated for each basic contour. Afterwards we compute linear momentums of the first four orders, namely mean, Centroid, standard deviation, Skewness and Kurtosis, as well as extrema, turning points and ranges. In order to keep dimensionality within range we decide by expert knowledge which functionals to calculate. Tab. 1 provides a rough overview of calculated functionals. Bracketed numbers resemble derived contours.

Classification and Feature Space Optimization

Considering our extensive classifier comparison in [3] we chose Support Vector Machines (SVM) with couple-wise one-vs.-one multi-class discrimination and polynomial kernel herein. Reduction of less relevant features often leads to higher classification performance, as the classifier is confronted with less complexity. We therefore apply Sequential-Forward-Floating-Search (SFFS) - a Hill-Climbing search - to reduce feature set size having SVM as wrapper-function. Apart from mere reduction of the feature space, also a combination with supervised expansion can lead to improved accuracy. We therefore generate novel features based on the so far pre-selected ones: Firstly, alteration of attributes by mathematical operations can be performed to lead to better representations of these. Secondly, by association of attributes we can obtain a further number of new information. As a deterministic and systematic generation comes to its limits applying exhaustive search, we decided for Genetic Algorithm (GA) based search through the possible feature space. As a start-set of effectually different individuals that represent possible solutions to the problem we use partitions of the acoustic feature sets reduced to a reasonable size by now. The partitions are denoted in binary coding, called *chromosomes*. Each chromosome consists of *genes* that correspond to single features within the partition. A feature's gene consists of one bit for its activity status. The partitioning is done randomly throughout initialization and we obtain $N=\dim(\underline{x})/n$ individuals if $\underline{x}$ denotes the feature vector, and n the partition size. By an initialization probability, set to 0.5 in our case, it is randomly decided which original features are chosen for one step of genetic generation. We decided to have a *population* size of 20 individuals at a time. Next a *fitness* function is needed in order to decide which individuals survive. Thereby the aimed at classifier forms a reasonable basis in view of wrapper based set optimization. A cyclic run over multiple *generations* is afterwards executed until an optimal set is found, which resembles a local maximum of a problem: Firstly, a *Selection* takes place, based on the fitness of an *individual*. We use common *Roulette Wheel* selection within this step. Thereby the 360° of a roulette wheel are shared proportional to the fitness of an individual. Afterwards the "wheel" is turned several times, resembling N times selecting out of N individuals. Selected individuals are assembled in a *Mating Pool*. Likewise, fitter individuals are selected more probably. We also ensure mandatory selection of the best one, known as *Elitist Selection*. The oncoming *Crossing* of pairs is fulfilled by picking $N/2$ times individuals with the probability $1/N$. After selection, individuals are put aside. Opposing traditional GA, we use a variable chromosome length from hereon, as we aim at generation of features. First we have to pick to *parents* in order to cross their chromosomes and thereby obtain new *children*. We then choose simple *Single-Point-Crossing* which splits each parent chromosome close to its center and pastes the two halves cross-wise to obtain two children. The fitness thereby also limits the total number of children an individual may produce. Afterwards, *Mutation* takes place: the state of a gene, respectively of a feature within a partition, is randomly changed by a probability of 0.5. Likewise features can be excluded from a set. To generate new feature we insert a random selection of an alteration method out of *reciprocal value*, *addition*, *subtraction*, *multiplication* and *division* [4]. Depending on the mathematical operation the appropriate number of features within an individual is selected for alteration, the operation is performed, and obtained features are appended. Now the evaluation of the population is fulfilled, resembling the fitness-test respectively classification with the feature sub-sets. At this point, one iteration is finished, and the algorithm starts over with selection. We decided for a maximum of 50 generations, and 40 without improvement.

Experiments and Conclusion

As a general mean of evaluation we use j-fold stratified cross-validation (SCV). Table 1 shows results on each dataset starting with the accuracies for the complete initial set of 276 features. Next results with the optimally reduced set by SVM-SFFS are shown. The features selected highly depend on the corpus and so does the number where maximum accuracy is observed of such: for EA-CAR the optimum was found at 92 features, for EMO-DB at 98, and for DES at 75. Finally, performance boost by application of the feature space optimization by combined genetic generation and reduction successive to SFFS (Gen. + Red.) is shown.

Table 1: Accuracies on diverse databases

Accuracy [%]	**EA-CAR**	**EMO-DB**	**DES**
Initial Set	70.2	84.8	65.9
SFFS Sel.	75.1	87.5	74.2
Gen.+Red.	**77.8**	**88.8**	**76.2**

The general principle shown in this paper could be demonstrated highly effective on all three databases used. Optimization of the feature space clearly boosted performance. However, besides mere reduction of complexity, also a combination with newly generated features clearly led to a significant further improvement.

In future research we aim at analysis of MPEG-7 LLD, and multi-instance-based learning.

Literature

[1] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B.: *"A Database of German Emotional Speech,"* Proc. Interspeech 2005, ISCA, pp. 1517-1520, Lisbon, Portugal, 2005.

[2] Engberg, I. S.; Hansen, A. V.: *"Documentation of the Danish Emotional Speech Database DES,"* Aalborg, Denmark, 1996.

[3] Schuller, B. et al: "*Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles,"* Proc. Interspeech 2005, pp. 805-809, Lisbon, Portugal, 2005.

[4] Mierswa, I.: Automatic Feature Extraction from Large Time Series," *Proc. 28. Annual Conference of the GfKl 2004*, Springer, pp. 600-607, 2004.