EVOLUTIONARY FEATURE GENERATION IN SPEECH EMOTION RECOGNITION

Björn Schuller, Stephan Reiter, Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
{Schuller | Reiter | Rigoll}@tum.de

ABSTRACT

Feature sets are broadly discussed within speech emotion recognition by acoustic analysis. While popular filter and wrapper based search help to retrieve relevant ones, we feel that automatic generation of such allows for more flexibility throughout search. The basis is formed by dynamic Low-Level Descriptors considering intonation, intensity, formants, spectral information and others. Next, systematic derivation of prosodic, articulatory, and voice quality high level functionals is performed by descriptive statistical analysis. From here on feature alterations are automatically fulfilled, to find an optimal representation within feature space in view of a target classifier. To avoid NP-hard exhaustive search, we suggest use of evolutionary programming. Significant overall performance improvement over former works can be reported on two public databases.

1. INTRODUCTION

In the field of multimedia retrieval there is a great interest in the capability to automatically segment media streams according to emotional. Likewise specific emotional passages within TV-broadcasts, movies or audio-plays can be easily retrieved as highlights in soccer games or exciting passages in thrillers. Also, reliable recognition of human emotion highly is expected to highly enhance next generation man machine interaction in view of naturalness [1]. More generally applications reach from detection of lies to surveillance in public transport, or intelligent customer handling in call-centers. Looking at suited modalities, speech analysis is among the most promising information sources besides facial expressions, physiological data or context analysis. As special advantage speech allows a user to control the amount of emotion shown, and provides more comfort than wiring with physiological sensors or permanent camera observation. However, reported performances are yet to be increased considering serious real-life application [1, 2] – especially in critical scenarios as emotion aware board computers in automotive environments regarded herein. We therefore strive to bridge the gap between the commercially highly interesting multiplicity of potential applications and current accuracies [1, 2, 3, 4].

Previously we compared diverse approaches to linguistic analysis of spoken utterances in view of emotion [5]. We also showed higher overall performance obtained by inclusion of such. However, within this work we want to focus on improvement considering merely the acoustic signal, to demonstrate the effects of genetic feature generation and selection in detail.

In acoustic analysis feature relevance is largely discussed [1]. Still, it seems mostly agreed that global static features lead to higher accuracies compared to dynamic classification of multivariate time-series, as shown in our explicit comparison in [6]. The feature basis is mostly formed by pitch, energy, and duration information. Some works also include spectral information or formants. Recently, large feature sets are introduced and reduced by diverse means of feature selection as floating search methods and principal component based reduction [3,4,5]. While feature selection is a reasonable starting point, we feel that a systematic generation of features helps to form a broader basis to start from. Combined with appropriate selection, a self-learning feature space optimization can be established. Deterministic generation comes to its limits, if we aim at alterations and cross-feature relations not considered, yet. In this respect we suggest an evolutionary approach to this problem: Genetic Algorithms (GA) have already proven successful in related fields [7]. In this work we therefore want to transfer this powerful tool.

In order to demonstrate the high effectiveness of the suggested approach we provide results on two popular public databases [3], namely DES and EMO-DB. However, these corpora comprise only acted samples. Having our application within an automotive infotainment system in mind, we also chose our task specific EA-CAR database which contains spontaneous emotions in the field.

The paper is structured as follows: Section 2 deals with databases used, section 3 with dynamic contour extraction, section 4 with systematic functional derivation, section 5 with support vector classification. In section 6 we discuss feature space transformation. Finally we present obtained results in section 7 and draw conclusions in section 8.

2. DATABASES

Only sparse public databases and results on these are available in the field of affective computing at the time. On two of them several accuracies are reported, which will be used in the ongoing. Firstly, the Berlin Emotional Speech database (EMO-DB) [8] will be shortly introduced. The emotion set equals the MPEG-4 standard consisting of *anger, disgust, fear, joy, sadness* and *surprise*, besides an exchange of surprise in favor of *boredom* and added *neutrality*. 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. The database is recorded in 16 bit, 16 kHz under studio noise conditions.

Secondly, we also chose the Danish Emotional Speech Corpus (DES) [9], which is proposed in a comparison of 31 corpora in [3]. In this database the five emotions *anger, joy, sadness, surprise*, and *neutrality* are contained. Four professional Danish actors, two of them female, acted the words *yes* and *no*, 9 sentences and two text passages in each emotion. By phrase-wise splitting of the text passages 414 phrases are obtained in total. The set is recorded in 16 bit, 20 kHz PCM-coding in a sound studio. 20 test-persons, 10 of them female, reclassified the samples in a perception test. Their recognition rate was 67.32% in average.

However, it is stated that acted and spontaneous samples highly differ in view of features and accuracies [4]. We therefore also use our EA-CAR database of spontaneous interaction turns within a car. These base on a user-study with 2 female and 8 male German test-subjects aged 23.4 a in average that controlled an infotainment interface by natural speech. Speech functionality was simulated by a wizard in the first half of an 80 min session. In the second half actual speech recognition and natural language interpretation engines 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 23 a to 30 a. A closed emotional set was used. Taking only phrases with full inter-labeler agreement results in 775 phrases: 225 of anger, 135 of confusion, 25 of joy, and 390 of neutrality. Due to its sparseness joy was excluded form the set.

3. DYNAMIC BASE CONTOURS

As a basis for feature generation we extract low-level contours of a whole phrase. Such global phrase-wise view is obligatory due to database annotations available. 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. We use standard frame energy to include intensity information based on physical relations. 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 a 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 usage of MFCC is highly discussed, as these tend to depend too strongly on the spoken content. This seems a drawback, as we want to recognize emotion independently of the content. However, they have been proven successful, yet, and form a very good basis for the subsequent genetic generation process, as thereby inter-band-relations will be analyzed. The further spectral features are often used in Music Retrieval, and are included to observe their relevance within this task.

Finally, for articulatory features we calculate the spectral centroid. Overall, parts of these contours are comprised within the MPEG-7 LLD standard. Likewise, the following methods may be transferred in order to recognize emotion based on MPEG-7.

4. SYSTEMATIC FUNCTIONAL DERIVATION

In former works we showed the higher performance of derived functionals instead of full-blown contour classification [6]. We therefore use systematic generation of functionals f out of multivariate time-series F by means of descriptive statistics:

$$f : F \rightarrow \mathbb{R}$$

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 base 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 represent derived contours.

Number [#]	F	$F+\partial+\partial^2$	f
Elongation	1	1	3
Intensity	1	3	11
Intonation	1	3	12
Duration	(2)	(2)	5
Formants	14	28	105
MFCC	15	45	120
HNR	1	1	3
FFT based	5	7	17
Sum	**38**	**88**	**276**

Tab. 1: Overview Derived Acoustic Functionals.

5. CLASSIFICATION

Diverse machine-learning techniques are used in the field of emotion recognition [1,2]. In [5] we made an extensive comparison including Naïve Bayes, k-Nearest Neighbor classifier, Support Vector Machines (SVM), Decision Trees, Artificial Neural Nets, and construction of ensembles as MultiBoosting or StackingC. SVM have thereby proven the optimal choice.

They show a high generalization capability due to a structural risk minimization oriented training. Non-linear problems are solved by a transformation of the input feature vectors into a generally higher dimensional feature space by a mapping function, the *Kernel*, where linear separation is possible. Maximum discrimination is obtained by an optimal placement of the separation plane between the border of two classes. The plane is spanned by *Support Vectors*.

In general, SVM can handle only two-class problems. However, a variety of strategies exist for multi-class discrimination. Among these are popular couple-wise one-against-one decision, or one-against-all classes decision. For more details refer to [11]. Herein, we use a special solution known as *SVM-Trees*. Thereby a layer-wise two class decision is repetitively made until only one class remains. The clustering of the emotions and alignment on the layers significantly influences recognition performance. As a rule hardly separable classes should be divided at last. This can either be modeled by expert knowledge or automatically derived of the confusion matrices of another multi-class SVM approach. In our case a couple-wise decision is used as starting-point. Afterwards the classes are split in the tree shape. SVM-Trees thereby always outperformed the couple-wise ones. Throughout the ongoing we use a polynomial kernel-function.

6. GENETIC FEATURE GENERATION

Having irrelevant features in the acoustic vector increases complexity for classifiers, and thereby directly decreases performance in most cases. This is especially true for sparse data, as the aimed at emotional data. It is therefore state-of-the-art to avoid this by reduction of the feature set by suited methods [3,4,5]. Thereby also computational extraction effort is spared. In general, feature reduction either considers single feature relevance, mostly done by filter-based selection as information gain calculation, or optimization of a feature set as a whole. Within the latter a classifier is used as a wrapper, ideally the target one, and a search-function obligatory in most cases to ensure computability.

However, besides reduction of the feature space, also its expansion can lead to improved accuracy. Consider hereon the Kernel-trick in SVM classification. However, while an optimal Kernel has to be selected empirically, we aim at a self-learning approach to feature space transformation based on random injection. Especially the combination of both by a suited search algorithm and the target classifier, allows for self-learning optimization of the ideal representation within feature space.

In order to expand the feature space we generate novel features based on the existing ones: Firstly, alteration of attributes by mathematical operations can be performed to lead to better representations of these. Consider hereon the standard use of logarithmic HNR representation. So far we only considered features based on single contours. By association of these we can secondly obtain a further number of new information as the named inter-band dependency. As a deterministic and systematic generation comes to its limits applying exhaustive search, we decided for GA based search through the possible feature space. The parallel selection of most relevant information and reduction is fulfilled within one pass by this GA based search.

GA, a well-known bio-analog search method, base on Darwin's *survival-of-the-fittest* principle of mutation and selection [10]. Besides single feature mutation, we also include crossing of parental DNA information - in our case feature crossing. GA are computationally expensive, but can be parallelized.

The precondition is to have a start-set of effectually different individuals that represent possible solutions to the problem. In our case these are partitions of the acoustic feature sets carrying information about the underlying emotion. The partitions are denoted in binary coding, and are called *chromosomes* in terms of GA literature. 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 forms a local maximum of the 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 subsequent *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*. Thereby the distance between parents and children should reasonably be smaller than the one between parents themselves. We therefore 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.* Depending on the mathematical operation the appropriate number of features within an individual is selected for alteration, and the operation is performed. Thereby new features can be constructed by combination of original ones. The obtained new individuals are than appended within the chromosome.

Now the *Evaluation* of the population is fulfilled, which corresponds to the fitness-test – in our case classification with the feature sub-sets. We use SVM on cross-validation set, as we want to optimize the feature space for SVM classification. At this point, one iteration is finished, and the algorithm starts over with the Selection. We decided for a maximum of 50 generations, and 40 of them without improvement.

7. RESULTS

Within this section we present results obtained by test runs on the described databases. As a general mean of evaluation we use j-fold stratified cross-validation (SCV).

In order to first demonstrate effectiveness of functional derivation as shown in section 4 compared to direct analysis of contours as shown in section 3, we exemplary pick MFCC as the most popular features in speech processing. As the base contours need dynamic modeling, we decided for common use of Hidden-

Markov-Models (HMM). However, we apply a powerful variant – a hybrid Neural-Net HMM approach having a Multilayer Perceptron (MLP) estimate state posteriors. HMM state numbers and MLP architecture have been varied to find an optimum. For static analysis we employ SVM as the most powerful variant. Tab. 2 verifies our former results [6], as the functionals clearly outperform dynamic features.

Accuracy [%]	***F*, HMM**	***f*, SVM**
MFCC	55.14	68.44
***MFCC*+∂+∂²**	59.26	**73.77**

Tab. 2: Dynamic vs. static modeling, EMO-DB, 2-fold SCV.

Next, we want to evaluate the suggested genetic feature generation. To save computation time, we firstly fulfill a Sequential Forward Floating Search (SFFS) on each database, which is known for its high performance. SFFS belongs to the group of Hill-Climbing feature selection based on a classifier as a wrapper and evaluation mean of a feature set performance. Herein the same classifier, SVM, is likewise used to optimize the basic 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 in a floating manner to avoid nesting effects.

This is the point where the novel aspect of this paper starts: After this pre-selection of base features we start genetic search and generation as described in section 6. It has to be stressed that these parts have to be executed only during the training phase. Within the recognition phase the system is a conventional speech emotion recognition engine.

Tab. 3 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 SFFS are shown. The features selected highly depend on the corpus and so does the number where maximum accuracy is observed of such: for DES the optimum was found at 99 features, for EMO-DB at 73, and for EA-CAR at 92. Finally, performance boost by application of the feature space optimization by combined genetic generation and reduction successive to SFFS (Gen. + Red.) is shown.

Accuracy [%]	**DES**	**EMO-DB**	**EA-CAR**
Initial Set	65.94	84.84	68.21
SFFS Sel.	74.15	87.50	75.08
Genetic Gen.+Red.	**76.15**	**88.82**	**77.18**

Tab. 3: Feature Space Optimization, SVM-Trees, 10-fold SCV.

In a last test we added newly generated features of the terminal evolutionary set to the one obtained by SFFS. Thereby performance could be further increased, yet not exceeding the genetic generation and reduction.

7. CONCLUSIONS

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 led to a significant further improvement. Significance bases on a paired Student-T-Test and a level of α=0.05.

The results presented outperform those shown in other works for the databases DES and EMO-DB [3,4]. Furthermore they are within the range of human performance, though under quite idealistic conditions. The accuracy on a spontaneous database, the EA-CAR corpus, reached similar level.

In future research we aim at analysis of MPEG-7 LLD, multi-task learning, and investigation on the spontaneous CEICES Emotional Speech Corpus [12].

8. REFERENCES

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

[2] 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.

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

[4] Vogt, T.; Andre, E.: *"Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition,"* Proc. ICME 2005, Amsterdam, Netherlands, 2005.

[5] Schuller, B.; Reiter, S.; Müller, R.; Al-Hames, M.; Lang, M.; Rigoll, G.: *"Speaker Independent Speech Emotion Recognition by Ensemble Classification,"* Proc. ICME 2005, Amsterdam, Netherlands, 2005.

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

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

[8] 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.

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

[10] Goldberg, D. E.: *Genetic Algorithms in Search, Optimization & Machine Learning,* Addison-Wesley Publishing Company, Inc., 1989.

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

[12] Batliner, A.; Hacker, C.; Seidl, C.; Nöth, E.; D'Arcy, S.; Russel, M.; Wong, M.: *"You stupid tin box – Children interacting with the AIBO robot: A cross-linguistic speech corpus,"* Proc. LREC 2004, Lisbon Portugal, 2004.