Bioanalog acoustic emotion recognition by genetic feature generation based on low-level-descriptors

Björn Schuller, Dejan Arsic, Frank Wallhoff, Manfred Lang, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Dejan Arsic, Frank Wallhoff, Manfred Lang, and Gerhard Rigoll. 2005. "Bioanalog acoustic emotion recognition by genetic feature generation based on low-level-descriptors." In *EUROCON 2005 - The International Conference on "Computer as a Tool," 21-24 November 2005, Belgrade, Serbia*, edited by Ljiljana Milić and Đorđe Paunović, 1292–95. Piscataway, NJ: IEEE. https://doi.org/10.1109/EURCON.2005.1630194.

Bioanalog Acoustic Emotion Recognition by Genetic Feature Generation Based on Low-Level-Descriptors

Björn Schuller, Dejan Arsić, and Frank Wallhoff, *Members, IEEE*
Manfred Lang, and Gerhard Rigoll, *Senior Members, IEEE*

***Abstract* — Affective Computing has grown an important field in today's man-machine-interaction, and the acoustic speech signal is very popular as basis for an automatic classification at the moment. However, recognition performances reported today are mostly not sufficient for a real usage within working systems. Therefore we want to improve on this challenge by evolutionary programming. As a starting point we use prosodic, voice quality and articulatory feature contours. We next propose systematic derivation of functionals by means of descriptive statistics. In order to analyze cross-feature information and feature permutations we use Genetic Algorithms, as a complete coverage of possible alterations is NP-hard. The final attribute set is at the same time optimized by reduction to the most relevant information in order to reduce complexity for the classifier and ensure real-time capability during extraction process. Classification is fulfilled by diverse machine learning methods for utmost discrimination power. We decided for two public databases, namely the Berlin Emotional Speech Database, and the Danish Emotional Speech Corpus for test-runs. These clearly show the high effectiveness of the suggested approach.**

***Keywords* — Affective Computing, Emotion Recognition, Speech Processing, Genetic Feature Generation**

I. Introduction

AFFECTIVE COMPUTING has grown an important field of research in today's man machine interaction and multimedia retrieval [1]. Applications reach from surveillance in public transport to emotion aware board computers in cars. Speech analysis is among the most promising information sources considering automatic emotion recognition besides mimic, physiological or context data analysis. While performance obtained by automatic systems based on this channel are among most reliable ones, it is still not sufficient for usage in real-life scenarios [1, 2]. We therefore strive to bridge the gap between the commercially highly interesting multiplicity of potential applications and current accuracies [1, 2, 3, 4].

In our previous work we demonstrated that inclusion of linguistic analysis of the spoken content helps to improve overall performance [5]. However, within this work we want to focus on the analysis of the acoustic signal, to demonstrate genetic feature generation and selection to improve on this stream. One advantage thereby is that no speech recognition engine is necessary in the latter usage. The optimal set of acoustic features is largely discussed [1]. Still, it seems mostly agreed that static features perform better than dynamic classification of multivariate time-series, as shown in our explicit comparison [6]. The basis of most works is formed by pitch, energy, and durations. Some works also include spectral information or formants. In [3,4,5] large feature sets are introduced and reduced by diverse means of feature selection as floating search methods and principal component based reduction. 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. Deterministic generation comes to its limits, if we aim at cross-feature relations and novel permutations not considered, yet. In this respect we suggest an evolutionary approach to this problem. Genetic Algorithms (GA) have already been shown successful in the field of Music Information Retrieval [7]. In this work we therefore want to transfer this powerful tool.

Most recognition performances are reported on individual databases and emotion sets. However, there are public corpora evolving at the time. We therefore want to demonstrate the effectiveness of our suggestions on such. As the right choice of classifiers and effect of emotion set reduction is also of interest, we provide results in this respect on the chosen public databases, of which one has been proposed in a comparison of 31 corpora in [3].

The paper is structured as follows: Section II deals with basic contour extraction, section III with the systematic generation of functionals, and section IV with the genetic extension of this idea. In section V we discuss optimal classification. In sections VI and VII the databases are described in detail and results are presented. Finally we draw conclusions and refer to future research activity.

B. Schuller is with the Institute for Human-Machine Communication, Technische Universität München, Germany (phone: +49-89-289-28548; fax: +49-89-289-28535; e-mail: Schuller@tum.de).

D. Arsić is with the Institute for Human-Machine Communication, Technische Universität München, Germany.

F. Wallhoff is with the Institute for Human-Machine Communication, Technische Universität München, Germany.

M. Lang is with the Institute for Human-Machine Communication, Technische Universität München, Germany.

G. Rigoll is with the Institute for Human-Machine Communication, Technische Universität München, Germany.

II. Low-Level-Descriptors

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: a Hanning windowing is used for contours in the time domain, Hamming for spectral analysis. 20 ms 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 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 the 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 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 genetic generation, 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, as articulatory features we use the spectral centroid. It should be mentioned that part of these contours are comprised within the novel MPEG-7 LLD standard. Likewise, the following methods can be partly transferred in order to recognize emotion basing on MPEG-7.

III. Systematic Functional Generation

In former works we showed the higher performance of derived functionals instead of full-blown contour classification [4]. We therefore use systematic generation of functionals f out of time-series F by means of descriptive statistics:

$$f : F \rightarrow \mathbb{R} \qquad (1)$$

First of all the contours are smoothed by symmetrical moving average filtering with a window size of three. Likewise we are less prone to noise in the calculation. Successively, speed (∂) and acceleration (∂^2) are calculated for each basic contour described in section II. 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. The following table 1 gives an overview.

Table 1: Overview Derived Features.

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
Total	**38**	**88**	**276**

IV. Genetic Feature Generation

So far we only considered features based on single contours. By association of these we can obtain a high number of new information as the named inter-band dependency. As a deterministic and systematic generation comes to its limits if we aim at full blown search - already with a limited number of allowed operations - we decided for GA based search through the possible feature space. Thereby further more alteration of attributes by mathematical operations can be performed and may lead to better representations of these. Consider here fore the standard use of logarithm for HNR representation. Also, such feature permutation can be seen similar to the *Kernel-trick* in Support Vector classification (see section V). However, while an optimal Kernel has to be selected empirically, genetic generation is a self-learning approach to feature space transformation based on random injection.

Still, if we increase the total number of attributes most classifiers suffer under complexity problems. This is especially true for sparse data, as the aimed at emotional data. A parallel selection of most relevant information and reduction to it is therefore mandatory. Feature selection is also fulfilled by GA based search.

GA form a very powerful bioanalog method basing on Darwin's *survival-of-the-fittest* principle of mutation and selection [10]. Following Neo-Darwinists, we also include crossing of parental DNA information - in our case feature crossing. GA are computationally expensive, but they can be parallelized to a high degree. Calculation is only needed once prior to find the optimal feature set.

The precondition is to have a start-set of effectually different individuals that represent possible solutions to the problem. In our respect these are acoustic features carrying information about the underlying emotion. We have accomplished this step in section III. A cyclic run is

afterwards executed until an optimal set is found, which resembles a local maximum of a problem. By an initialization probability, set to 0.8 in our case, it is randomly decided which original features are chosen for one step of genetic generation. We here fore decided to have a *population* size of 20 features 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. As a selection algorithm we chose *Roulette Wheel*, which gives a higher probability for individuals with high fitness to be chosen. However, we always additionally ensure to keep the best feature within a cycle. From a population size of N we make a bootstrap sample of the same size N. Thereby multiple instances of the same individual may be contained in a so called *Mating Pool*. Out of this pool we randomly take $N/2$ individuals for *single-point-crossing* with a given cross-over probability set to 0.6 in our case. The mentioned fitness is also used to decide how many children may be produced by a pair. The next phase is *mutation*, again by a certain probability, 0.6 in this case. We chose *reciprocal value*, *addition*, *subtraction*, *multiplication* and *division* as mathematical operations for this purpose. After mutation an iterative jump to population generation takes place, until an abruption criterion is fulfilled. We decided for a maximum of 50 generations, and 40 of them without improvement.

Figure 1 gives an overview of the principle of iterative genetic generation and genetic selection until maximum accuracy is reached. These parts have to be executed only during the training phase. It has to be mentioned that the final sets selected are influenced by the learning set. Within the recognition phase the system resembles a conventional pattern recognition engine.

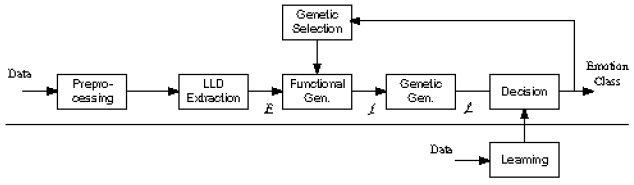

Fig. 1. Overview of the iterative generation principle

V. Automatic Classification

The optimal classifier is broadly discussed [1,2]. In [5] we made an extensive comparison including Naïve Bayes, k-Nearest Neighbor classifier, Support Vector Machines (SVM), Decision Trees, and Neural Nets. The major drawbacks of the firstly named well known statistical model-based Naïve-Bayes (NB) classifier is its 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. If a majority vote among the k nearest neighbor (kNN) reference instances is fulfilled this classifier also resembles a statistical approach. Support Vector Machines (SVM) can be seen as an analogon to electrostatics. Thereby a training sample corresponds to a charged conductor at a certain space, the decision function an electrostatic potential function, and the learning target function the Coulomb energy. SVM show a high generalization capability due to a 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. Finally, Neural Nets are a standard procedure in pattern classification. We chose them in order to show a wholly bioanalog variant combining the power of genetic algorithms and neural nets. They are renowned for their non-linear transfer functions, their self-contained feature weighting capabilities and discriminative training. We use a Multi-Layer Perceptron (MLP) having one hidden with sigmoid transfer functions, and softmax outputs. Further more we demonstrate how more powerful classifiers can be constructed by means of meta-classification as MultiBoosting or Stacking. However, we focus on base classifier herein. For more details on classifiers refer to [11].

VI. Databases

In order to provide results on a public corpus we decided for the Berlin Emotional Speech database (EMO-DB) [8], which consists of 816 phrases in total. The emotion set resembles 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.

For further results on another dataset we also chose the Danish Emotional Speech Corpus (DES) [9]. In this database the four emotions *anger, joy, sadness,* and *surprise* of the MPEG-4 set plus *neutrality* are contained. Four professional Danish actors, two of them female, simulated the word *yes* and *no*, 9 sentences and two text passages in each emotion. We split the text passages into single sentences and thereby obtained 414 phrases in total. The set was 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 between 59% and 80% with an average resembling 67.32%.

VII. Results and Conclusions

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

In order to verify our former results [6] considering direct time series-classification versus functional

approaches, we used standard Hidden-Markov-Models to classify the contours of intensity, intonation, MFCC and their derivatives. A test-run was performed on the EMO-DB corpus. 50.79% maximum accuracy could be reached compared to 74.39% by use of functionals basing on the same contours, and SVM classification.

In table 2 the performance based on single feature groups is shown. It can be clearly seen, that the compound of all systematically generated features outperforms each individual group. However, genetic generation (gen.) and selection (sel.) leads to far better results, especially in the case of DES. The final optimal feature vector consisted of 101 features for EMO-DB and 75 for DES.

TABLE 2: RECOGNITION ACCURACIES BY FEATURE GROUPS.

Accuracy [%]	EMO-DB	DES
Duration	27.46	19.08
HNR	30.33	25.60
Elongation	33.40	36.23
Intensity	48.16	39.61
Intonation	62.09	32.85
Formants	63.73	41.79
FFT Spectrum	70.70	39.86
MFCC	77.25	57.97
All Features	84.84	65.94
Genetic Gen. + Sel.	**87.70**	**75.36**

The next table 3 gives an overview of classifier comparison. We always chose the optimal classifier configuration and feature set as shown in table 2. SVM are first choice on both datasets, but MLP follow close.

TABLE 3: RECOGNITION ACCURACIES BY CLASSIFIERS.

Accuracy [%]	EMO-DB	DES
NB	73.98	52.42
1NN	75.82	31.16
kNN	78.89	50.73
SVM	**87.70**	**74.15**
C4.5	61.48	51.69
MLP	86.48	71.98

Figure 2 and 3 demonstrate the effect of emotion set reduction. A sliding window technique over classes was used to estimate the 120 permutations for the 7 emotions contained in EMO-DB by 42 runs in total. On DES we calculated all 26 variants of the 5 emotions.

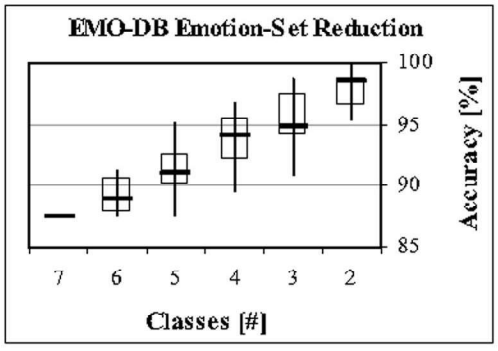

Fig. 2. Influence of the emotion-set size, EMO-DB.

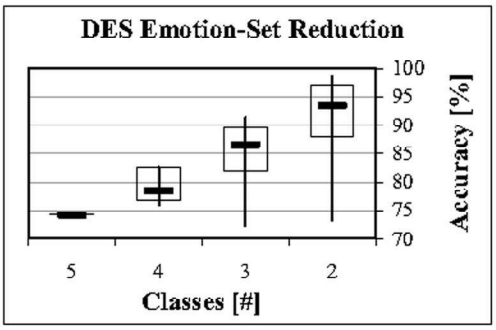

Fig. 3.Influence of the emotion set-size, DES.

The final table demonstrates the effectiveness of our suggested approach. It provides human perception accuracy as reported in [8, 9] and maximal automatic classification performance as reported in [3, 4]. Finally our best results with and without genetic generation and optimization are shown.

TABLE 4: COMPARISON RECOGNITION ACCURACIES.

Accuracy [%]	EMO-DB	DES
Human	84.25	67.32
Other works	77.4 [4]	51.6 [3]
Deterministic set	84.84	65.94
Genetic set	**87.70**	**74.15**

Likewise, we showed an effective approach to systematic functional derivation and successful cross-functional analysis and permutation considering by Genetic Feature Generation. In future research we aim at investigation on the spontaneous AEC database.

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, IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005.

[5] 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.

[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," in Weihs, C. and Gaul, W. (editors), Classification -- the Ubiquitous Challenge, Proc. of the 28. Annual Conference of the GfKl 2004, Springer, pp. 600--607, 2004.

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

[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.