Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture

Björn Schuller, Gerhard Rigoll, Manfred Lang

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Gerhard Rigoll, and Manfred Lang. 2004. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 17-21 May 2004, Montreal, Canada*, edited by Douglas O'Shaughnessy, Li Deng, Peter Kabal, and Steven Blostein, 577–80. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2004.1326051.

SPEECH EMOTION RECOGNITION COMBINING ACOUSTIC FEATURES AND LINGUISTIC INFORMATION IN A HYBRID SUPPORT VECTOR MACHINE - BELIEF NETWORK ARCHITECTURE

Björn Schuller, Gerhard Rigoll, and Manfred Lang

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

ABSTRACT

In this contribution we introduce a novel approach to the combination of acoustic features and language information for a most robust automatic recognition of a speaker's emotion. Seven discrete emotional states are classified throughout the work. Firstly a model for the recognition of emotion by acoustic features is presented. The derived features of the signal-, pitch-, energy, and spectral contours are ranked by their quantitative contribution to the estimation of an emotion. Several different classification methods including linear classifiers, Gaussian Mixture Models, Neural Nets, and Support Vector Machines are compared by their performance within this task. Secondly an approach to emotion recognition by the spoken content is introduced applying Belief Network based spotting for emotional key-phrases. Finally the two information sources will be integrated in a soft decision fusion by using a Neural Net. The gain will be evaluated and compared to other advances. Two emotional speech corpora used for training and evaluation are described in detail and the results achieved applying the propagated novel advance to speaker emotion recognition are presented and discussed.

1. INTRODUCTION

A growing interest in the recognition and integration of users' emotions in the interaction with machines can be observed at the time. A large number of applications exists reaching from the discipline of information retrieval to medical analysis [1]. In our research emotion recognition is applied within automotive environments. An in-car board system shall be provided with information about the mental state of the driver to initiate safety strategies, initiatively provide aid or resolve errors in the communication according to the driver's emotion. Focusing on the field of man machine interaction non-invasive advances seem more popular in recent works due to a user's control of the emotion shown and a certain comfort provided by the non-invasive nature. While mimic and speech analysis seem to be most promising, we focus on speech as input channel in this work. Most of the advances to speech emotion recognition rely on acoustic characteristics of an emotional spoken utterance. However, in recent approaches more emphasis is also put on the spoken content itself [2], and the most reasonable advance seems to be the integration of acoustic and linguistic information [3]. In the work presented we therefore strive to combine these two information sources in a most robust way. Firstly we aim to show an optimal feature set and classification method in a comparison respecting a high performance and speaker independency taking only acoustic features into consideration. Secondly we concentrate on the language information. While in other works the probability of an emotion is estimated by conditional probabilities of single words in an utterance we introduce an emotional phrase spotting algorithm based on Belief Networks. The idea behind this effort is to include the context of a whole utterance as negations of feelings and allow for a speaker's indication of the emotional extent. Consider on this the following example: *"I **do not** feel **too** good **at all**."* The keyword "*good*" is neglected and furthermore *"too"* alludes the actual extent. After this discussion of acoustic- and language-based emotion recognition a novel approach to the fusion of these shall be presented. While the combination has yet been accomplished mostly in a late semantic fusion manner, we introduce a soft decision fusion saving available information for the final decision process. As still no unity about a general classification scheme for emotions in technical applications exists, and the use of discrete emotional user states is far spread among researches in the field of automatic emotion recognition, we consider the emotional states named in the MPEG4 standard: anger, joy, disgust, fear, sadness, and surprise. This set is often supplemented by a neutral state for a dissociation from a non emotional state. In view of international comparability [4][5] we decided upon this set of seven emotions in our work. The estimation of an emotion shall respect a whole spoken utterance.

2. EMOTIONAL SPEECH CORPUS

The emotional speech corpus has been collected in the framework of the FERMUS III project [6], dealing with emotion recognition in an automotive environment. A dynamic AKG-1000S MK-II microphone was used in an acoustically isolated room to record the emotional utterances. German and English sentences of 13 speakers, one female, were assembled. A first corpus consists of 2829 acted emotional samples used for the training and evaluation in the prosodic and linguistic analysis. The samples were recorded over a period of one year to avoid anticipation effects of the actors. While these acted emotions tend to form a reasonable basis for a first impression of the

obtainable performance, the use of spontaneous emotions seems to offer more realistic results, especially in view of the spoken content. A second set consists of 700 selected utterances in automotive infotainment speech interaction dialogs recorded for the evaluation of the fusion. In the project disgust and sadness were of minor interest. Therefore these have been provoked in additional usability test-setups to ensure equal distribution among the emotions in the data set. To obtain a basis for comparison the speakers had to reclassify their own samples in a random order at the end of the test series. The following table shows their average performance. A rather marginal overall standard deviation among the human classifiers of 2.11% was observed. In the following figures *ang* abbreviates anger, *dis* disgust, *fea* fear, *joy* joy, *neu* neutral, *sad* sadness, and *sur* surprise.

Emotion	**ang**	**dis**	**fea**	**joy**	**ntl**	**sad**	**sur**
Error, %	8.0	19.7	18.7	14.7	16.5	23.7	12.5

Figure (1): Human reclassification error rate, mean 16.3%

3. ACOUSTIC FEATURE SET

In former works [7] we compared static and dynamic feature sets for the prosodic analysis. Due to their higher classification performance we focus on derived static features in this work. Initially the raw contours of pitch and energy are calculated because they rely rather on broad classes of sounds. Spectral characteristics on the other hand seem to depend too strongly on phonemes and therefore on the phonetic content of an utterance, which is a drawback with respect to the premise of independency of the spoken content throughout the acoustic analysis. Therefore only spectral energy below 250 Hz and 650 Hz is used considering spectral information. 20 ms frames of the speech signal are analyzed every 10ms using a Hamming window function. The values of energy resemble the logarithmic mean energy within a frame. The pitch contour is computed by the average magnitude difference function (AMDF). Calculated in first order AMDF provides a faster alternative to the autocorrelation function due to a restriction to additions. As all pitch estimation methods this technique underlies deviations from the original pitch, which could only be measured by glottal measurement. AMDF proves robust against noise but susceptible to dominant formants. A low-pass filtering applying a symmetrical moving average filter of the filter-width three smoothens the raw contours prior to the statistical analysis. In a next step higher level features are derived out of the contours, freed of their mean value and normalized to their standard deviation. The temporal aspects of voiced sounds are approximated with respect to zero levels in the pitch contour due to the inharmonic nature of unvoiced sounds. Silence durations are calculated by an energy threshold. As the optimal set of global static features is broadly discussed [1][5], we considered an initially large set of more than 200 features. The features are ranked with aid of a Linear Discriminant Analysis, and the following table lists the elements of our final 33 dimensional feature-vector in detail. In a direct comparison a combination of all pitch related features lead to 69.81% correct recognition rate, compared to 36.58% correct recognition rate for the use of all energy related features.

Feature	**LDA,%**
Pitch maximum gradient	31.5
Pitch relative position of maximum	28.4
Pitch standard deviation	27.6
Pitch mean value gradient	26.1
Pitch mean value	25.6
Pitch relative maximum	25.2
Pitch range	24.8
Pitch relative position of minimum	24.4
Pitch relative absolute area	23.8
Pitch relative minimum	23.7
Pitch mean distance between reversal points	23.0
Pitch standard dev. of dist. between reversal points	23.0
Energy mean distance between reversal points	19.0
Energy standard dev. of dist. between reversal points	18.6
Duration mean value of voiced sounds	18.5
Spectral energy below 250 Hz	18.5
Energy standard deviation	18.1
Energy mean of fall-time	17.8
Energy median of fall-time	17.8
Energy mean value	17.7
Energy mean of rise-time	17.6
Duration of silences mean value	17.5
Rate of voiced sounds	17.0
Signal number of zero-crossings	16.9
Signal median of sample values	16.8
Energy median of rise-time	16.7
Signal mean value	16.7
Energy relative maximum	16.6
Spectral energy below 650 Hz	16.3
Energy relative position of maximum	15.9
Energy maximum gradient	15.7
Duration of silences median	15.7
Duration of voiced sounds standard deviation	15.1

Figure (2): Ranking of the acoustic features according to a Linear Discriminant Analysis

4. CLASSIFICATION OF THE ACOUSTIC SET

Various different methods have been taken into consideration for the classification on the acoustic layer. In the following the results using either linear classifiers, Gaussian Mixture Models, Neural Nets or Support Vector Machines are presented and the optimal parameter configuration will be discussed.

4.1. Linear Classifiers

As a lower end in performance a simple Euclidean Distance metric based classifier deciding for the nearest class mean vector (kMeans) was used. As a variant a k nearest neighbors classifier (kNN) was evaluated in a second advance. A decision is made according to the majority vote among the k nearest references to the input vector. The maximum performance was observed for k resembling one, respectively deciding for the best hit itself. The observed results using these classifiers clearly show the non-linear nature of the problem and demand for more sophisticated approaches.

4.2. Gaussian Mixture Models

Gaussian Mixture Models (GMM) provide a good approximation of the originally observed feature probability density functions by a mixture of weighted Gaussians. The mixture coefficients were computed by use of an Expectation Maximization algorithm. Each emotion is modeled in one GMM. The decision is made for the maximum likelihood model. A maximum recognition result was observed by use of 16 mixtures.

4.3. Neural Nets

Neural Nets are a standard procedure in pattern classification. They are renowned for their non-linear transfer functions, their self-contained feature weighting capabilities and discriminative training. Considering the sparsely available emotion training material their good performance on small training sets compared to GMMs seems advantageous. A Multi Layer Perceptron with 33 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. A performance maximum was observed using 100 neurons in the hidden layer. The outputs were normalized as posteriors by a softmax function. For the training a Backpropagation with 1000 iterations, cross-entropy as an error function, and a cross validation set were used.

4.4. Support Vector Machines

A great interest in Support Vector Machines (SVM) in classification can be observed recently. They tend to show a high generalization capability due to their 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 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 the support vectors leading to a reduction of references. A number of approaches to solving multi-class problems exist. In this evaluation we show three different solutions. Once each class is trained in its own SVM against all other classes, and the decision is made for the class with the highest distance to the other classes. In a second variant the distances are fed forward to a MLP as described in 4.3. with 7 input and output, and 400 hidden neurons. In a third advance Multi-Layer SVMs (ML-SVM) are introduced. The following figure shows the principle.

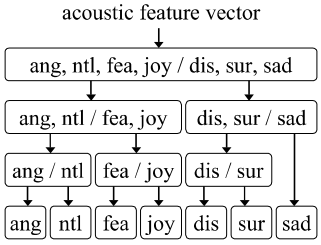

Figure (3): Optimal alignment of the emotions using ML-SVMs

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 throughout the evaluation we found that hardly separable classes should be divided at last. This can either be modeled by expert knowledge or automatically derived of the confusion matrices of the first introduced SVM approach. One disadvantage however is that no confidences can be computed for each class. This variant is therefore not used in the fusion. A radial basis kernel as mapping function showed the best results.

4.5. Classification Results

In a test-series the introduced classifiers have been tested applying the large speech corpus. Two thirds have each been used for training, one third for testing in three cycles. The mean error rates are shown in the following table. The standard deviations reached from ±0.01% to ±0.03%. A speaker dependent (*S DEP)* training with only the speaker, and speaker independent (*S IND)* evaluation were considered.

Classifier	S IND, Error, %	S DEP, Error, %
kMeans	57.05	27.38
kNN	30.41	17.39
GMM	25.17	10.88
MLP	26.85	9.36
SVM	23.88	7.05
SVM – MLP	24.55	11.3
ML–SVM	18.71	9.05

Figure (4): Comparison of the acoustic feature classifications

5. LANGUAGE INFORMATION

In general only a small amount of user utterances will consist of emotional information. Even if an utterance carries information of the actual user emotion it will in most cases be only fragments of the complete utterance. Therefore a spotting approach seems a must in search for emotional keywords or phrases in natural language. As a basis we use a standard Hidden-Markov-Model-based automatic speech recognition (ASR) engine with zero-grams as language model. It provides the n-best hypotheses including single word confidences of its estimation of the spoken content. As a mathematical background for the spotting we chose Belief Networks for their capability to handle uncertain and incomplete information. In this paper we can provide only a very brief insight in the theory of Belief Networks, which enjoy growing popularity in pattern recognition tasks. Each network consists of a set of nodes related to state variables *X*, consisting of a finite set of states. The nodes are connected by directed edges expressing quantitatively the conditional probabilities of nodes and their parent nodes. A complete representation of the network structure and conditional probabilities is provided by the joint probability distribution. Let *N* denote the total of random variables, and the distribution can be calculated as:

$$P(X_1,...,X_N) = \prod_{i=1}^{N} P\left(X_i \mid \text{parents}(X_i)\right)$$

Methods of interfering the states of some query variables based on observations regarding evidence variables are provided by the network. Similar to a standard approach to natural speech interpretation, the aim is finding the emotion hypothesis that maximizes the posterior probability of the word sequence given the acoustic observation. Each emotion is modeled in its own network. The root probabilities are distributed equally in the initialization phase and resemble the priors of each emotion. If the emotional language information interpretation is used stand-alone, a maximum likelihood decision takes place. Otherwise the root probability for each emotion is fed forward to a higher-level fusion algorithm as with the acoustic confidences. In four lower levels a clustering from words to super-words, resembling a reduction to 18.9%, phrases, super-phrases, and finally emotions takes place as can be seen in the following figure.

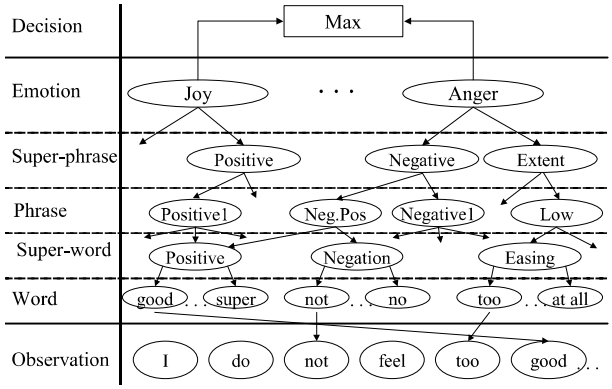

Figure (5): Principle of Belief network based phrase spotting

On the word layer the evidences are fed into the net according to the word confidences of actually observed words. As we integrate the confidences of the ASR hypotheses the traditionally certain evidences are extended as uncertain evidences. The quantitative contribution $P(e_j|w)$ of any word w to the belief in an emotion e_j is calculated in a training phase by its frequency of occurrence under the observation of the emotion on the basis of the hand-labeled large emotional speech corpus.

6. SOFT DECISION FUSION

In this chapter we aim to fuse the acoustic and linguistic information obtained. In other works the fusion is suggested as a late semantic logical "OR" combiner [3]. Since we strive to integrate information of more than two classes, a first approach might be to consider a couple-wise mean score for each emotion based on the acoustic and language information score followed by an adjacent maximum likelihood decision. As an advantage soft scores of both aspects are used in the computation prior to the final decision. However, this rather simple fusion neglects the fact that for each emotion the prior confidences in acoustical and language-based estimations differ. Further more a discriminative approach helps to integrate the knowledge of all accessible emotion confidences in one decision process. We therefore suggest the use of a MLP as introduced in 4.3 for the fusion. The 14 dimensional input feature vector consists of the 7 confidences of each the acoustic, and linguistic analysis. 7 output neurons provide the final emotion probabilities by a softmax function. A use of 100 hidden-layer neurons showed the maximum performance. The MLP was trained on a second data set disjunctive of the initial training sets. For the evaluation of the combination a third data set was used. The following table shows results achieved using the FERMUS III dialog corpus and optimal configurations. 12% of the utterances contained only acoustic information of the underlying emotion.

Model	Acoustic Information	Language Information	Fusion by means	Fusion by MLP
Error, %	25.8	40.4	16.9	8.0

Figure (6): Performance gain means-based and MLP fusion

7. CONCLUSION

We believe that this contribution shows important results considering the combination of acoustic and linguistic information in speech emotion recognition as a solid model was introduced and a significant gain was achieved reducing error rates up to 8.0%. In the emotion estimation by acoustic information a ranked set of features was presented, and different classification methods were compared. The use of SVMs predominated in robustness on this layer. Additionally a novel approach to linguistic information interpretation in view of a speaker's emotion using Belief Network based phrase-spotting could be shown. Finally the results of these analyses could be integrated in a reasonable MLP soft decision fusion, and lead to a significant improvement in overall performance.

8. REFERENCES

[1] R. Cowie, et al.: *"Emotion recognition in human-computer interaction,"* IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] L. Devillers, L. Lamel: *"Emotion Detection in Task-Oriented Dialogs,"* Proceedings of the ICME 2003, IEEE, Multimedia Human-Machine Interface and Interaction I, Vol. III, pp. 549-552, Baltimore, MD, USA, 2003.

[3] C. M. Lee, R. Pieraccini: *"Combining acoustic and language information for emotion recognition,"* Proceedings of the ICSLP 2002, Denver, CO, USA, 2002

[4] A. Nogueiras, et al.: *"Speech Emotion Recognition Using Hidden Markov Models,"* Eurospeech 2001, Poster Proceedings, pp. 2679-2682, Scandinavia, 2001.

[5] V. Hozjan, Z. Kacic: *"Improved Emotion Recognition with Large Set of Statistical Features,"* Proceedings of the Eurospeech 2003, pp. 133-136, Geneva, Switzerland, 2003.

[6] B. Schuller: *"Towards intuitive speech interaction by the integration of emotional aspects,"* IEEE Int. Conf. SMC 2002, Yasmine Hammamet, Tunisia, CD-Rom Proceedings, 2002.

[7] B. Schuller, G. Rigoll, M. Lang: *"Hidden Markov Model-Based Speech Emotion Recognition,"* Proceedings of the ICASSP 2003, IEEE, Vol. II, pp. 1-4, Hong Kong, China, 2003.