Speech Emotion Recognition Exploiting Acoustic and Linguistic Information Sources

Gerhard Rigoll, Ronald Müller, Björn Schuller

Institute for Human-Machine Communication
Munich University of Technology, Germany

Abstract

A variety of approaches towards Speech Emotion Recognition were presented since the research activities started in the late last decade. Today we are all aware of the great importance of emotional aspects as the next step towards more natural human-machine interaction. Yet, a growing number of isolated individual groups contribute to the general advances in the exciting landscape we see today. However, there is still a strong need for international cross-fertilization on the one hand in view of signal pre-processing, feature selection, and classification methods, and furthermore comparability with respect to emotion classes, and databases. As first commercial products arise today, the research community is challenged by more realistic out-of-the-lab conditions. This leads to the next level of Speech Emotion Recognition aiming among others at a robust recognition, independent of the speaker, background-noise, and to a certain extent even of the spoken language by reasonable integration of acoustic, linguistic, and context information, as well as spotting for emotions out of a stream. In this article we like to give an overview of the existing pieces of the puzzle, introduce novel impulses, and aim to investigate future directions in this young discipline.

1. Introduction

Now that we started substituting button pressing by more natural communication such as talking and gesturing, still human-computer communication feels somehow impersonal, insensitive, and mechanical. If we take a comparative glance at human-human communication we will realize a lack of the extra information sensed by humans concerning the affective state of the counterpart [1]. This emotional information highly influences the explicit information, as recognized by today's human-machine communication systems, and with an increasingly natural communication, respect of it will be expected [2]. Throughout the design of next generation man-machine interfaces inclusion of this implicit channel therefore seems obligatory. Automatic emotion recognition is nowadays already introduced experimentally in call centers, where an annoyed customer is handed over from a robot to a human call operator [3], [4], and in first commercial lifestyle products as Nemesysco's Handy Truster and Love Detector intended to detect lies, stress- or love level of a telephoner following the work of Libermann [5]. A large number of various application scenarios reaches from automatic ambulant supervision in psychological aid to high security operators' observation. In the end of this article we provide an exemplary use-case of our research, where emotion is utilized in a car infotainment interface to react to irritated, joyful or angry drivers [6].

2. Speech and Emotion

Human emotion is basically observable within a number of different modalities. First attempts to human emotion recognition applied invasive measurement of e.g. the skin conductivity, heart rate, or temperature [7]. While exploitation of this information source provides a reliable estimation of the underlying affect, it is often felt uncomfortable and unnatural [8], as a user needs to be wired or at least has to stay in touch with a sensor. Modern emotion recognition systems therefore focus rather on vision or audio based non-invasive approaches in the style of human emotion recognition [2]. Mehrabian [9] claims that we communicate by 55% visually, through body language, by 38% through the tone of our voice and by 7% through the actual spoken words. In this respect the most promising approach clearly seems to be a combination of vision and speech based affect estimation. However, in some systems and situations only speech may be available as interaction channel. Thereby Mehrabian's findings stress the importance of acoustic and language information. Contrary to most other modalities speech allows the user to control the amount of emotion shown, which may play an important role, if the user feels too much observed otherwise. Seen from an economical point of view a microphone as sensor is standard hardware in many HCI-systems today. Finally Speech Emotion Recognition provides reasonable results already by now, and led to first commercial products.

2.1. First steps

Basically it can be said that two main information sources are exploited considering emotion recognition from speech: the acoustic information analyzing the prosodic structure as well as the spoken content itself, namely the language information. Hereby the predominant aims besides high reliability are an independence of the speaker, the spoken language, the spoken content when considering acoustic processing, and the background noise. A number of parameters strongly influence the quality in these respects and will be addressed throughout the ongoing: the underlying emotion model, the database size and quality, the signal capturing, pre-processing, feature selection, classification method, and a reasonable integration in the interaction and application context.

2.2. The Emotion Model Discussion

Prior to recognizing emotion one needs to establish an underlying emotion model. In order to obtain a robust recognition performance it seems reasonable to limit the complexity of the model, e.g. kind and number of emotion labels used, in view of the target application. This may be one of the reasons that no consensus exists about such a model in

From: SPECOM 2005, Patras, Greece, Kokkinakis (ed.),
Moscow State Linguistic University

technical approaches, yet. Two generally different views dominate the scene: on the one hand an emotion sphere is spanned by 2 up to 3 orthogonal axes: firstly arousal or activation, respecting the readiness to take some action, secondly valence or evaluation, considering a positive or negative attitude, and finally control or power, analyzing the speaker's dominance or submission [10]. While this approach provides a good basis for emotional speech synthesis, it is often too complex for Speech Emotion Recognition purposes. On the other hand the better known way is to classify emotion by a limited set of discrete emotion tags (e.g. [11], [12], [13], [14]). A first standard set of such labels exists within the MPEG-4 standard comprising anger, disgust, fear, joy, sadness and surprise. In order to discriminate a non-emotional state it is often supplemented by a neutral state [15]. While this model opposes many psychological approaches [16], it provides a feasible basis in technical view. The MPEG-4 set with a neutral state will be used in comparative evaluations throughout this article for consistency reasons.

2.3. The Database Problem

In order to train and test automatic emotion recognition engines, a database is needed within the next step. Such an emotional speech corpus should provide spontaneous and realistic emotional behavior out of the field. The sample quality should ensure studio quality, but for analysis of robustness in the noise also samples with known background noise conditions should be included. The database has to consist of a high number of ideally equally distributed speech utterances for each emotion, both of the same speaker, and of many different speakers in total. The speakers should provide a flashy model considering genders, age groups, ethical backgrounds, among others. Respecting further variability, phrases should possess different contents, lengths, or even languages. For certain experiments it can also be interesting to have varying pronunciations of the same linguistic information in view of the emotional flavor. An unambiguous assignment of collected phrases to emotion classes is especially hard in this discipline. Finally, the database should be made publicly available in view of international comparability, which seems a problem considering the lacking consensus about emotion classes used.

A number of methods exist to create a database, with arguably different strengths. The predominant ones among these are acting or provocation of emotions in test set-ups, hidden or conscious long-term observations, and use of clips out of public media content. However, most databases use acted emotions [17], which allow for fulfillment of the named requirements besides the spontaneity, as there is doubt whether played emotions are capable of representing true characteristics of affect in spoken utterances. Still they do provide a reasonable starting point, considering that databases of real emotional speech are hard to obtain. The latter is besides other reasons given by the fact that intimate details of the speaker must be given away in order to assign them to emotions [17].

Throughout the ongoing article all results presented of our research are carried out on the same corpus and emotion set in order to provide comparable results. The corpus has been collected in the framework of the FERMUS-III project [18], 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 set 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. A second set consists of 700 selected utterances of automotive infotainment speech interaction dialogs.

2.4. Subjective Human Performance

As we know, it may be hard to rate one's emotion for sure. In this respect it seems obvious that comparatively minor recognition rates can be demanded in this discipline considering related pattern recognition tasks. However, the human performance provides a reasonable benchmark for a maximum expectation. Many tests exist respecting this problematic. However, these are generally carried out on different emotion classes, and rarely on the corpora used within evaluation of recognition engines. Therefore the actors within our database had to reclassify their own utterances at the end of the collection. While this test as a side-effect also evaluates the quality of the acting itself, it still provides a feasible basis. The following table shows the subjective performance emotion-wise with an overall error rate of 16.3% ± 0.211%. Each actor also had to classify the utterances of the other unknown speakers, which lead to a significantly higher error rate of 35.3% ± 0.96%. The latter result helps to interpret speaker independent classification performance. In the following ang abbreviates anger, dis disgust, fea fear, ntl neutral, sad sadness, and sur surprise.

Table 1: Human reclassification error rate, mean 16.3%

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

The only other well-known study on the same emotion set is found in the work of Nogueiras et al. [15] where an overall error rate of 20% is given, letting 16 probands classify 56 acted spoken utterances in four languages.

3. Acoustic Information

In order to estimate a user's emotion by acoustic information one has to carefully select suited features. Such have to carry information about the transmitted emotion, but they also need to fit the chosen modeling by means of classification algorithms. While the feature sets used in existing works differ greatly, they mostly base on the continuous measures of pitch, energy, duration and spectral information (e.g. [3], [13], [19], [20], [21]). However, first problems arise, as some of the underlying raw features as the pitch contour cannot be estimated exactly. Two in general different approaches exist considering acoustic feature processing: dynamic and static.

3.1. Dynamic vs. Static Approach

Within the dynamic approach the raw feature contours, e.g. the pitch or intensity contours, are directly analyzed frame-wise by means of dynamic programming realized by Hidden Markov Models (HMM) (e.g. [15], [18]). On the other hand derivation and classification of statistical high-level features calculated out of the raw contours over a whole utterance is named static approach in the following. In a direct comparison [22] under constant test conditions the static features outperformed the dynamic approach in our studies. This is highly due to the unsatisfactory independence of the overall contour in respect of the spoken content. Eliminating e.g. unvoiced parts from the pitch contour helps to improve in this question, but also leads to a loss of temporal information. In any case filtering of the contours leads to a gain, whether for the direct analysis or for the further processing. In our case a symmetrical moving average filter prevailed over a median filter as used by Nogueiras et al. [15]. A minimum error rate of 21.0% could be achieved using continuous HMMs. In the ongoing we will concentrate on the significantly more powerful static feature approach.

3.2. Feature Selection

As mentioned the type of statistical features used vary greatly throughout the different works. A quantitative evaluation of such features is therefore a must and provided e.g. in the works of McGilloway et al. [13]. Table 2 shows our optimal set of 33 features ranked by a Linear Discriminant Analysis performed on the FERMUS-III database as described earlier.

3.3. Classification Methods

A number of factors influence the choice of the classification method. Besides high recognition rates and efficiency, economical aspects and a reasonable integration in the target application framework play a role. In the ongoing research a broad spectrum reaching from rather basic classifiers such as Linear Discriminant Classifier (LDC, e.g. [3]), or k Nearest Neighbors (kNN) to more complex methods like Multi-Layer Perceptrons (MLP, e.g. [11]) or Support Vector Machines (SVM, e.g. [13]) is applied. In order to provide an impression of the performance on this task we compared the classifiers with the features mentioned in the previous section on the FERMUS-III database and the extended MPEG-4 emotion set. Two thirds have been used for training, one third for testing in three cycles. The mean error rates are shown in table 3. The standard deviations ranged from ±0.01% to ±0.03%. A speaker dependent (S DEP) and speaker independent (S IND) evaluation were considered.

Table 2: Ranking of the acoustic features according to a Linear Discriminant Analysis

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	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

Besides the outlined classifiers also Gaussian Mixture Models (GMM) and three alternative approaches to SVM multi-class solutions will be presented. Firstly, each class is trained against all remaining in its own SVM, choosing the minimum distance result in the recognition process. Secondly, the output of such SVMs is fed forward into a MLP (SVM - MLP), and finally Multi-Layer SVMs (ML-SVMs) with a layer-wise discrimination of emotion sets were used. The results show the maximum achieved performance under optimal parameter configuration.

Table 3: Comparison of the acoustic feature classification

Classifier	**S IND, Error, %**	**S DEP, Error, %**
LDC	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

The results obtained using ML-SVMs vary significantly with the hierarchically splitting of the emotions on the layers. Finding an ideal solution can be automated by use of the confusion matrices of a weak classifier, in order to find hardly separable emotions. These shall be discriminated on the final layer. In the following figure the optimal alignment found throughout our test-series is shown.

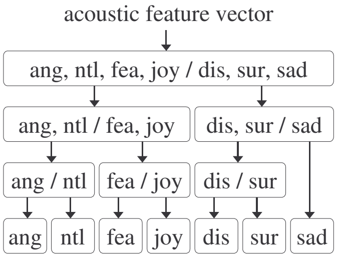


Figure 1: Optimal alignment of the emotions using ML-SVMs

4. Linguistic Information

Up to here we described a considerable amount of research effort on feature extraction and classification algorithms to the investigation of vocal properties for the purpose of inferring to probably expressed emotions from the sound. So the question after information transmitted within the implicit channel [2] "How was it said?" has been addressed with great success. Recently more attention is paid to the interpretation of the spoken content itself dealing with the related question "What was said?" in view of the underlying affect [3], [4], [23]. In psychological studies it is claimed that a connection between certain terms and the related emotion has been learned by the speaker [24]. As hereby the speaker's expression of his emotion consists in usage of certain phrases that are likely to be mixed with meaningful statements in the context of the dialog, an approach with abilities in spotting for emotional relevant information is needed. Consider for this the example "Could you please tell me much more about this awesome field of research". In the work of Arunachalam [25] observing kids during video gaming, an average of only 5% reaching from 0%-25% of emotionally relevant statements occurred. This ratio is clearly dependent of the underlying application background and the personal nature of the speaker. Nevertheless it remains questionable whether linguistic information might be sufficient applied standalone: in our database of spoken automotive interaction dialogs 12% of the emotionally interesting utterances were sounds like moaning without linguistic content. However, integration of this aspect showed clear increase in performance [3], [23], even though the conclusions drawn rely per definition on erroneous Automatic Speech Recognition (ASR) outputs. In order to reasonably handle the incomplete and uncertain data of the ASR unit, a robust approach has to take acoustic confidences into account throughout the processing. Still, none existing system for emotional language interpretation calculates an output data certainty based upon the input data certainty, except for the approach presented in the ongoing.

4.1. N-Grams and Salience

A common approach to language modeling is the use of n-grams. In language interpretation based emotion recognition only uni-grams have been applied so far [3], [4]. They provide the probability of an emotion under the condition of single observed words without modeling of neighborhood dependencies. In order to handle only emotional keywords, Lee and Pieraccini [3] suggest sorting out abstract terms that cannot carry information about the underlying emotion as names initially. Afterwards the frequency of occurrence of each observed word within each existing emotion and its derived emotional salience are computed in a training phase. Only sufficiently salient words will be considered throughout the recognition phase. Devillers and Vasilescu [4] cope with emotionally irrelevant information by a normalized log likelihood ratio between an emotion and a general task specific model. As emotion data is sparse, they introduce an interpolation coefficient by integration of expert knowledge.

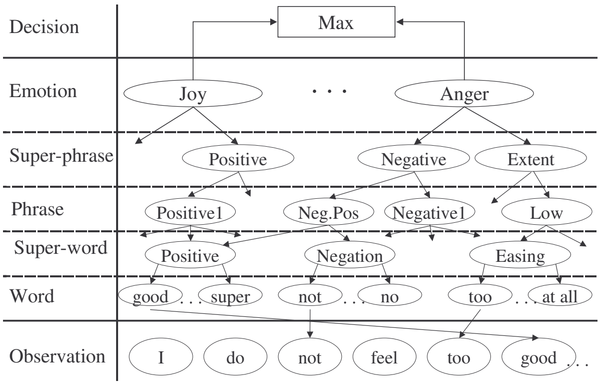
4.2. Phrase Spotting with Belief Networks

As mentioned, a key-drawback of the state-of-the-art uni-gram approach is a lacking view of the whole utterance. Consider hereby the negation in the following example: "I do not feel too good at all," where the positively perceived term "good" is negated. Therefore we chose Belief Networks (BN) as a mathematical background for the semantic analysis of spoken utterances taking advantage of their capabilities in spotting and handling uncertain and incomplete information. Within the final manuscript we intend to give a sufficient insight in the theory of Belief Networks, which enjoy growing popularity in knowledge modeling concerning artificial intelligence as well as in pattern recognition tasks. Still we would already like to give a very brief introduction to the theoretical basics here: Each Belief Network consists of a set of nodes related to state variables X, comprising 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 here is to make the net maximize the probability of the root node modeling the specific emotion expressed by the speaker via his choice of words and phrases. 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. On the input layer a standard HMM-based ASR engine with zero-grams as language model providing n-best hypotheses with single word confidences is applied. In order to deal with the acoustic certainties the traditional Belief Networks have been extended to handle soft evidences.

The approach presented here is to be based on integration and abstraction of semantically similar units to higher leveled units in several layers. This method proved to be applicable for Natural Language Interpretation in several delimited domains with remarkable results, like natural language man-machine dialogues for intuitive controlling of infotainment systems in cars. Hence, the description here is kept more general since we talk about interpretations in the top level, which may address more complex items like a number of various user commands or intentions, than a relatively small number of emotions. On the input level the n-best recognized phrases are presented to the algorithm, which maps this input on defined interpretations via its semantic model consisting in a Belief Network of the structure shown in figure 2.



"... I do not feel too good at all ..."

Figure 2: Cutout of interpretation model for emotions

At the beginning the spotting on items known to the semantic model is achieved by matching the words in the input level to word-nodes contained in the lowest layer of the BN. Within this step the knowledge about uncertainty of the recognized words represented by their confidences is completely transferred into the interpretation model by accordingly setting soft evidence in the corresponding word-nodes. While stepping forward to any superior model-layer, those units resembling each other in their semantic properties regarding the target interpretations are clustered to semantic super-units until the final layer with its root-nodes of the network is reached. Thereby the evidences assigned to word-nodes, due to corresponding appearance in the utterance, finally result in changes of probabilities in the root-nodes representing confidences of each specific emotion and their extent (Fig. 3). After all the presented approach allows for an entirely probabilistic processing of uncertain input to gain real probability afflicted output. To illustrate what is understood as semantically similar, consider for instance some words expressing positive attitude, as "good", "well", "great", etc. being integrated into the super-word "Positive". The quantitative contribution P(ej|w) of any word w to the belief in an emotion ej is calculated in a training phase by its frequency of occurrence under the observation of the emotion on basis of the speech corpus.

Given a word order within a phrase, an important modification of the classic Belief Networks has to be carried out, as BN's are in general not capable of processing sequences due to their entirely commutative evidence assignment. Further explanations will follow within the manuscript.

Training and evaluation procedures ran on the large FERMUS-III hand-labeled database comprising an equally distributed number of utterances for each emotion. The contained emotion of several samples recorded and transcribed can unambiguously be identified only via their prosodic properties, as from the spoken content alone a distinctive assignment is impossible even for a human. Hence, the error rate at 40.4% of this semantic interpretation approach left on its own seems to be rather poor compared to methods based on prosody. Nevertheless this additional information appears quite valuable, as proved within the next paragraph addressing the Late Semantic Fusion of results deriving from prosody and semantics.

5. Information Fusion

In this chapter we aim to fuse the acoustic and linguistic information obtained. In other works this integration 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. Furthermore a discriminative approach helps to integrate the knowledge of all accessible emotion confidences in one pass. In this respect we suggest the use of a Multi-Layer Perceptron (MLP) for the fusion. The 14-dimensional input feature vector consists of the 7 confidence measures of 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 to 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. As mentioned, 12% of the utterances contained only acoustic information of the underlying emotion.

Table 4: Emotion-wise performance, error rates

Error %	Acoustic Information	Language Information	Fusion by MLP
ang	4.5	15.7	2.0
dis	41.0	27.8	21.2
fea	22.5	43.4	12.1
sur	32.7	19.5	3.9
joy	27.9	39.5	3.8
ntl	23.5	44.0	2.6
sad	38.9	64.0	7.4

Table 5: Performances Means-based and MLP fusion, average error rates

Model	Acoustic Info.	Linguistic Info.	Fusion by means	Fusion by MLP
Error, %	25.8	40.4	16.9	8.0

6. Emotion Recognition in the Real Life – An Exemplary Use-Case

When using emotion recognition systems out-of-the-lab, a number of new challenges arise. This aspect has been hardly addressed within the research community, yet, as most tests were carried out under ideal conditions [8]. As these depend on requirements given by the application framework, we provide a very short insight into an exemplary use-case, show the arising problems, and denote chosen solutions:

In the FERMUS project [26] infotainment devices in the car are controlled mainly via speech, and further vision based and manual interaction modalities. Four emotional states are integrated within the interaction model of the human-machine interface: anger, joy, irritation, and a neutral user state. The knowledge about an irritated user allows for initiative assistance and tutoring. Observation of a user's anger subsequent to a system action leads to the activation of error resolving strategies and dialogs. Registration of a joyful user reaction helps as a pseudo-supervision guideline for user adaptation of further multimodal input processing modules. This idea derives from human feedback-oriented learning, like a child seeing its parents' smile, memorizes that cleaning up one's room is a good idea. While Speech Emotion Recognition produced considerable results when evaluated under ideal conditions, we will now show exemplary problems and results in the automotive environment.

In the car speech interaction is in general initiated by push-to-talk, meaning the user presses a button in order to speak. This method seems inappropriate in our case, as emotional utterances occur rather spontaneous. Therefore continuous spotting for emotional cues out of the open microphone stream seems obligatory. A simple state-of-the-art energy threshold based speech/pause-segmentation cuts out whole user utterances. As we want to focus on the driver's emotion, and avoid confusion with a passenger's, caller's or radio moderator's one, speaker verification for each emotional statement is a must-have. This proves especially challenging, as emotional utterances often tend to be short (6.7 sec in average throughout our corpus) and emotionally biased compared to standard speaker verification tasks, basing on ideally neutral voice and clips of about 30 sec. Presumed furthermore a robust speaker identification, a corresponding model can be loaded automatically for adaptation purposes. Finally we need to discriminate speech from remaining background noise as traffic, engine, or in-car music. Overall a Multi-Layer SVM discriminates between noise and voice on the first layer, and the target speaker and imposters on a second layer. For evaluation 756 sample-utterances of emotional attitude were recorded throughout interaction tests and labeled by the test-person himself afterwards. The following table shows the recognition performance and confusion achieved thereby.

Table 6: Confusion of emotions in automotive application, mean 82.8%

%	Anger	Irritation	Joy	Neutral
Anger	**68.9**	5.6	20.9	4.6
Irritation	4.8	**91.7**	0.0	3.6
Joy	18.2	0.0	**79.0**	2.8
Neutral	2.4	2.8	3.3	**91.5**

Having four drivers in the database, the correct speaker model could be loaded with an accuracy of 95.2%.

7. Conclusion

In this article we gave a deep insight in nowadays and future Speech Emotion Recognition. On the one hand an overview of state-of-the-art approaches was provided, and on the other hand we strived to lead to the next level of this young discipline by introduction of novel techniques. Throughout the paper we considered different static and dynamic features and measured the obtainable performance by a comparative evaluation. In respect of the classification method Hidden Markov Models, Linear Discriminant Classifiers, k-Means, Neural Nets, and multi-class solutions applying Support Vector Machines were discussed. While the static approach outperformed direct dynamic feature contour analysis, the Multi-Layer Support Vector Machines were the overall winners on the acoustic layer. Paying attention to the integration of the spoken content itself, we presented existing uni-gram approaches and a novel technique based on Bayesian Belief Network phrase spotting. In order to achieve a reasonable fusion of these two information sources we considered either state-of-the-art late semantic fusion, or more sophisticated soft decision fusion by means of a Multi-Layer Perceptron. It could be manifested that this integration of language information helps to improve the overall recognition performance and reduces error rates to as far as 8% resembling human performance on this task.

Now that Speech Emotion Recognition left its premature state, still a lot needs to be done within a real application framework, as shown in the exemplary use-case, where an absolute loss in performance of 8.15% was observed with a reduced set of 4 emotions.

8. References

[1] M. Schröder: "Experimental study of affect bursts," *Proceedings ISCA 2000*, Canada, 2000.

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

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

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

[5] http://www.nemesysco.com

[6] B. Schuller, M. Lang, G. Rigoll: "Multimodal Emotion Recognition in Audiovisual Communication," *Proceedings ICME 2002*, Lausanne, Switzerland, 2001.

[7] R. W. Picard: "Toward computers that recognize and respond to user emotion," *IBM Systems Journal*, vol. 39, NOS 3&4, S. 705-719, 2000.

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

[9] A. Mehrabian: “Communication without words,” *Psychology Today*, vol. 2(9), pp. 52-55, 1968.

[10] R. Cowie, E. Douglas-Cowie, S. Savvidou, E McMahon, M. Sawey and M. Schröder: “’Feeltrace’: an instrument for recording perceived emotion in real time,” *Proceedings ISCA 2000*, Canada, 2000.

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

[12] N. Amir: “Classifying emotions in speech: a comparison of methods,” *Poster Proceedings Eurospeech 2001*, pp. 127-130, Scandinavia, 2001.

[13] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve: “Approaching automatic recognition of emotion from voice: a rough benchmark,” *Proceedings ISCA workshop on Speech and Emotion*, pp. 207-212, Newcastle, 2000.

[14] B. Schuller, M. Lang, G. Rigoll: “Automatic Emotion Recognition by the Speech Signal,” *Proceedings SCI 2002*, IIIS, vol. IX, "Image, Acoustic, Speech and Signal Processing II", pp. 367-37, Orlando, Florida, USA, 2002.

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

[16] R. Cornelius: „Theoretical approaches to emotion,” *Proceedings ISCA Workshop on Speech and Emotion*, Belfast 2000.

[17] N. Campbell, N: “Databases of emotional speech,” *Proceedings ISCA Workshop on Speech and Emotion*, Northern Ireland, pp. 34-38, 2000.

[18] B. Schuller: “Towards intuitive speech interaction by the integration of emotional aspects,” *Proceedings IEEE Int. Conf. SMC 2002*, Yasmine Hammamet, Tunisia, 2002.

[19] M. Kienast, W. F. Sendlmeier: “Acoustical analysis of spectral and temporal changes in emotional speech,” *Proceedings ISCA 2000*, Canada, 2000.

[20] T. S. Polzin: “Verbal and non-verbal cues in the communication of emotions,” *Proceedings ICASSP 2000*, ID: 3485, Turkey, 2000.

[21] R. Cowie, E. Douglas-Cowie, B. Appoloni, J. Taylor, A. Romano and W. Fellenz: “What a neural net needs to know about emotion words,” *CSCC Proceedings*, pp. 5311-5316, 1999.

[22] B. Schuller, G. Rigoll, M. Lang: “Hidden Markov Model-Based Speech Emotion Recognition,” *Proceedings ICASSP 2003*, vol. II, pp. 1-4, Hong Kong, China, 2003.

[23] B. Schuller, G. Rigoll, M. Lang: “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture,” *Proceedings ICASSP 2004*, IEEE Int. Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, 17.-21.05.2004.

[24] R. Plutchik: “*The Pschology and Biology of Emotion*,” HarperCollins College, New York, 1991.

[25] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan: “Politeness and Frustration Language in Child-Machine Interactions,” *Proceedings Eurospeech 2001*, pp. 2675, Scandinavia, 2001.

[26] http://www.fermus.de

[27] B. Schuller, M. Zobl, G. Rigoll, M. Lang: “A Hybrid Music Retrieval System using Belief Networks to Integrate Queries and Contextual Knowledge,” *Proceedings ICME 2003*, IEEE, Multimedia Human-Machine Interface and Interaction I, Baltimore, MD, USA, Vol. I, pp. 57-60, 2003.

[28] B. Schuller, G. Rigoll, M. Lang: “Emotion Recognition in the Manual Interaction with Graphical User Interfaces,” *Proceedings ICME 2004*, IEEE, Taipei, Taiwan, 27.-30.06.2004.