Automatic Emotion Recognition by the Speech Signal

Björn Schuller, Manfred Lang, Gerhard Rigoll
Institute for Human-Machine-Communication, Technical University of Munich
80290 Munich, Germany, {schuller,lang,rigoll}@ei.tum.de

ABSTRACT

This paper discusses approaches to recognize the emotional user state by analyzing spoken utterances on both, the semantic and the signal level. We classify seven emotions: joy, anger, irritation, fear, disgust, sadness and neutral inner state. The introduced methods analyze the wording, the degree of verbosity, the temporal intention rate as well as the history of user utterances. As prosodic features duration, pitch and energy contribute to a robust recognition. Further more the problem of spotting for emotional phrases in the human-computer-interaction is alluded. User profiling supports the adaptation of different cultural comprehensions of verbally expressed emotions. To legitimate the applied features results of usability studies are introduced. Finally fields of application are shown and results are discussed.

Keywords: Emotion recognition, speech processing, social competence of machines

1. INTRODUCTION

Motivation
Speech is one of the most natural communication forms between human beings. Humans also express their emotion via written and spoken language. Enabling systems to interpret user utterances for a more intuitive human machine interaction therefore suggests also understanding transmitted emotional aspects. The actual user emotion may help a system track the user's behavior by adapting to his inner mental state. Generally recognition of emotions is in the scope of research in the human-machine-interaction. Among other modalities like mimic speech is one of the most promising and established modalities for the recognition [1][2][3]. There are several emotional hints carried within the speech signal. Nowadays attempts in detecting emotional speech analyze in general signal characteristics like pitch, energy, duration or spectral distortions [4]. However, on semantically higher levels emotional clues can also be found. In literature one can even rely *almost only* on such semantic hints besides spare graphical attempts to capture prosodic elements like in bold or italic characters typed phrases. Therefore we aim to also spot emotional key-phrases, analyze the dialogue history and the degree of verbosity in the communication between man and machine. This is realized through a parallel analysis of spoken utterances in view of general system announcements, command interpretation and detection of emotional aspects. However, the semantic means introduced could as well be used for analysis of non-spoken language.

General Fields of application
In the interpersonal communication partners adapt in their acoustic parameters to show sympathy for each other. A technical system enabled to talk by speech synthesis therefore needs to know the actual user emotion and the according acoustic parameters to adapt instead of staying neutral all the time. Further more the communication channels of a speaker interact with each other. The knowledge of the implicit channel is needed to interpret the explicit channel. Irony might be a good example to demonstrate that prosodic features help understand the explicitly uttered intention. An emotion recognition system might also be called in for an objective judgment in psychiatric studies [5]. Finally there is certainly a fun-factor in automatic reaction to user emotions in many applications like video games.

Concrete fields of application
The detected emotions recognized by the methods presented in this paper are used in our man-machine-interfaces. We want to recognize errors in the man-machine-interaction by a negative user emotion. If a user seems annoyed after a system reaction error-recovery strategies are started. On the other hand a joyful user encourages a system to train user models without supervision. First or higher order user preferences can be trained to constrain the potential intention sphere for erroneously recognition instances like speech or gesture input. To do so a system online needs a reference value like a positive user reaction. Furthermore our systems initiatively provides help for a seemingly irritated user. Control or induction of user emotions is another field of application that requires the knowledge of the actual emotion. For example in high risk-tasks it seems useful to calm down a nervous person, do not distract her by shortening dialogues, or keep a tired user awake.

User studies
The basis of probabilistic pattern recognition is to find

appropriate characteristics in a signal. To recognize such features we initially sampled speech utterances in usability studies and let humans classify the appurtenant emotion that they sensed. The analysis led to features carrying emotional aspects, but a general model of feature trends and associated emotional state could not be observed for any user. Only a certain suitability to assume emotional information and style guides for the interpretation could be determined by labeling exemplary data. The lack of a general standard can be solved by training with the posterior user. In the initial study 15 users, two of them female, had to control a browser by natural speech. The average age was 31.9 years with a maximum of 65 years and a minimum of 23 years. In the 471 sampled utterances 24 emotional markers could be found what resembles 5%. This is conform to a study where children's human computer interaction in game play was observed [6]. In a questionnaire 83.3% stated that they could imagine a system reacting to users' emotional behavior. 8.3% of them judged emotion recognition as very useful, 16.6 % as useful, while only 8.3% classified it as frightening or useless. In a second study with 17 probands, one of them female, an emotional reacting system was tested. Besides the recognition results presented later in this paper a high acceptance level could be viewed.

Classified states

In a first approach we used a two-dimensional emotion sphere defined by the axes activeness and positveness [7]. In this plain different areas could be assigned to emotional states. For example a very active and positive user is meant to be joyful, while an as well passive as negative user is associated with sadness. Other approaches introduce even a third dimension [8] with an axis of control level. The basing measurement of the extent of positiveness or activeness however turned out to be over-dimensioned. In a second approach we directly distinguished between seven basic emotional states according to the MPEG-standard [9]: joy, anger, irritation, fear, disgust, sadness and neutral user state. This is also a far spread classification of emotions with more or less states [10]. However, a provided confidence level of an assumed emotion might still also be seen as a measurement of its extent.

2. SEMANTIC FEATURES

On a semantic-syntactic level the spoken words and phrases themselves can transmit clear reference. Also the extend of verbosity of a speaker as well as his intention rate and the dialog history of a machine interaction can carry information about the emotional state. The achieved suggestion how emotion may be recognized on this level is explained in detail in the following chapters.

Emotional phrases

If we want to understand emotional markers in spoken user utterances, we have to massively cope with out of vocabulary occurrences. Due to the fact that we assume only around 5% of phrases containing emotional information on the semantic level in the interaction process with a machine we have to ensure that we do not misinterpret the remaining 95% of the phrases. This leads to a spotting approach and claims for confidence measures of an emotion hypothesis to avoid over-interpretation. First we spotted only for single emotional keywords like *"fine!"* or *"perfect!"*. A disadvantage however was that we could not cope with neglected, more complex or ironic phrases. As consequence we spot for emotional phrases instead of single words. This also allows for understanding further details as the announced extent of the emotion or temporal aspects. An example is: *"Well, if this goes on like this I won't feel that good anymore!"*. Spotting only for *"good"* would ignore that the speaker seems rather irritated and the aspect of the emotional trend. The basis of our speech interpretation is a speech recognition instance providing hypotheses on a word level with a score for the whole hypothesis as well as single word confidences. The single word confidences are normalized to the hypothesis length to avoid preferential treatment of hypothesis with a higher sum of words. Each emotion possesses a network-like model in accordance to Bayesian belief networks built by sub-models of super-phrases and optional phrases invariant to permutations. The networks with a-priori probabilities for the sub-models and the single words within the phrases ensure a correct view of the word order. Nevertheless nesting of phrases is allowed. The sub-models and their phrases build a network of keywords and optional words and their a-priori probabilities. For more reduction of information and simplification words can be clustered to super-words that represent semantic units belonging to the same semantic concept. In an early solution [11] we provided a type and a value for each semantic concept. In this realization we assign only a type but optionally keep the original wording throughout the search for a latter assignment of a value. This results in even stronger clustering. Like this there are two groups of super-words: Simple-super-words that discard the original term, and parameter-super-words that keep the exact wording. Each emotion score is calculated by a maximum search of the best fitting sub-model of each emotion by evaluating each hypothesis of the recognition instance. The score resembles as a confidence measure and an n-best list is achieved for a latter fusion with other instances. It proved that the amount of hypotheses of the speech recognition engine possesses a maximum in view of recognition results at 20 hypotheses. A first saturation point is reached at 12 hypotheses. But for fast real-time calculation even 10 hypotheses deliver satisfying results. In a training phase sample-utterances are exemplary split into their super- and optional phrases with their according

words to obtain the apriori probabilities of the sub-models, phrases and words.

Verbosity level
In a series of studies confronting users with erroneous interfaces we observed three levels of error-announcements. A decreasing degree of verbosity was strongly correlated with disaffection of users. Besides the trend to shorter phrases a change of wording and an increasing intention rate could be observed. These starting points for achievement of further characteristics will be described in the next chapters. Figure 1 shows the levels of user-annoyance when coping with an uncooperative system. On the IIIrd level interacting with a system underlying 5-10% error rate, they communicated very cooperative. In the case of an occurring error they indicated the underlying error plus its description and repeated their original intention. In a second phase confronted with a higher error-rate (10-20%) they shortened their statements by only stating the error and its specification. Finally at a very high error rate (~30%) users only announced an error if anyways they still talked with the system.

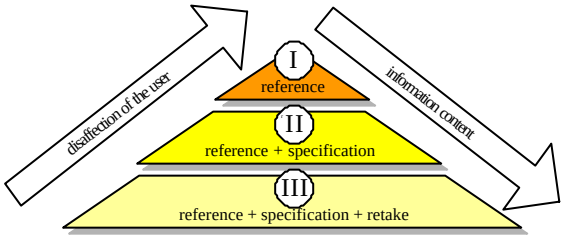


Figure (1): Levels of users' error announcement

The degree of verbosity can be measured by relating the total numbers of irrelevant semantic units and of meaningful units on a semantic-syntactic level. Therefore a garbage model of senseless units in view of the desired intention is needed.

Intention rate
It seems clear that the total amount of intentions in a set time interval portends the activeness of a user from a system's point of view. Like this the probands showed a very high activeness in stress situations according to the expectation. In combination with the dialog history negative emotion can be detected when realizing that the user cannot cope with the system.

Repetition and contradiction rates
These can allude interference in the communication between human and machine due to a distracted, tentative or tired user. To achieve the repetition or contradiction rates we have to interpret user utterances and compare them on the intention level. Furthermore we have to ensure that we take only preposterous occurrences into consideration. A matrix for repetitions and contradictions and a degree of its peculiarity allow for their decoding. A weight for the extent is needed for there are more or less reasonable repetitions or contradictions.

Rate of change in wording
If a repetition occurs we also compare the wording of the utterances. Our studies clearly showed that users tend to change their wording if they feel that the system does not understand them.

Classification
The mentioned rates for verbosity, total intentions, contradictions, repetitions, and change in wording (plus d and dd) form a 15-dimensional feature vector classified by distance metrics.

3. PROSODIC FEATURES

Besides the semantic analysis duration, pitch and energy of the speech signal contribute to the recognition.

Feature extraction
Each 10ms a frame windowed by a Hanning-function is analyzed. Pitch F0 is calculated by the average magnitude distance function. The AMDF is optimized for integer logic and bases on the auto correlation function, which is robust against surrounding noise [12]. Equation 1 shows our calculation of the AMDF-function, where MAXLAG stands for the maximum amount of AMDF-values in a frame.

$$AMDF_n(j) = \frac{1}{N}\sum_{i=1}^{N}\left|x_n(i) - x_n(i+j)\right|, \quad 1 \le j \le MAXLAG$$

Eq. (1): Calculation of the average magnitude distance function

However a weakness can be seen when the search for the maximum which stands for the fundamental frequency in the auto-correlation can result in dominant higher formants. Pitch detection therefore itself underlies errors what should be taken into consideration regarding final results. We decided to take local pitch features due to the highly speaker- and phrase-type [13] dependant global features. The energy is achieved by averaging signal energy in each frame. Finally we calculate duration by the rate of voiced sounds what proved as a reliable approximation. A voiced sound is assumed if a set threshold close to zero Hertz in pitch is exceeded.

Classification
First we analyzed a six dimensional feature vector built by pitch and derivation of energy (plus d and dd) using a DTW-algorithm with Itakura-constraints [14]. The derivation of energy was used to avoid influences of the sampled signal level. This solution proved to be satisfying besides having a too strong word or phrase dependency. To reduce the information for further generalization we analyzed derived pitch and energy

features. This approach is commonly used and seems reasonably compared with the results achieved in [15] using a 12-dimensional feature vector. In the following the elements of our 20-dimensional feature-vector that proved most important are listed in detail:

- Relative pitch maximum/minimum
- Position of maximum/minimum pitch
- Average pitch
- Standard deviation of pitch
- Mean of absolute pitch derivation
- Maximum of absolute pitch derivation
- Rate of voiced sounds
- Mean duration of voiced sounds
- Standard deviation of duration
- Mean distance between reversal points
- Standard deviation of distance between reversal points
- Relative maximum of derivation of energy
- Position of maximum of derivation of energy
- Average of derivation of energy
- Standard deviation of derivation of energy
- Maximum of second derivation of energy
- Mean distance between reversal points
- Standard deviation of distance between reversal points

Each feature can be weighted individually for the calculation in view of adaptation to a user or surrounding influences. The features are freed of their mean value and normalized to their standard deviation. We use standard distance metrics for classification. In a first trial the minimum distance for any reference vector in a class was determined. Additionally a score system that appoints a score for each feature that lies in its range of standard deviation for an assumed emotion was introduced. As final and most robust alternative we evaluated the minimum of intra-coefficient distances.

4. SEMANTIC FUSION

Single signal-analyzing instances

The introduced semantic and signal characteristics are evaluated automatically in three single signal-processing instances. The features are extracted in a preprocessing stage in cooperation with a speech-understanding unit in an open microphone manner. After pre-processing each instance calculates the score for each emotion. The instances are: Understanding of emotional phrases, semantic feature analysis, and prosodic feature analysis.

Semantic Fusion

The signal processing units can be used as isolated instances to achieve an estimation of a user emotion. However, in combination recognition results tend to be more stable. Additionally a more predicative measurement for the confidence of the emotional state by comparing coincidences is achieved. These instances can be prioritized inter-emotionally with aid of a user dependant weight matrix since it proved that users differ in the way that they express a certain emotion. Each instance possesses a vector consisting of the a-priori conditional probability. The final score list for each emotion is achieved by averaging the scores with respect to the a-posteriori probability of an instance in view of the user. An advantage of calculating a score for each emotion is the ability to backtrack the emotional development under the assumption of temporal false interpretation. Generally future modalities like mimic evaluation can be integrated easily.

5. USER ADAPTION

Analysis of early usability studies showed that a cross-cultural and user-independent reference model of the feature-trends and their cohering emotional states could not be assumed. Such a generalized model can only be seen as initialization basis for further user profiling. The training influences the a-priori expectation of the influence of the different estimation instances for the integration. The first order conditional probability $P(E_i|U_k)$ of an emotional state E_i in the context of a user U_k and the second order probability $P(E_i[n]|E_i[n-1],U_k)$ of an emotion in the context of the antecedent emotional state build the basis of the targeted user model. The index n alludes discrete temporal events. Finally a user influences recognition instance specific parameters like reference models for the phrase spotting or standard deviations and mean values of the prosodic features. User adaptation without supervision could not be realized satisfactory for an emotional model. Profiling can be done in a playful way by letting the system ask the user at a detected emotional change about his feelings in an initialization phase. After collecting first data the system can initiate more direct dialogs interrogating the user.

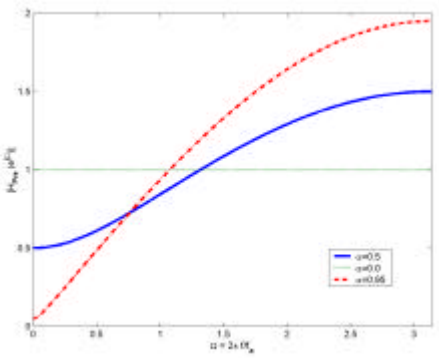
6. SPOTTING FOR EMOTIONAL PHRASES

In normal speech controlled applications users can be asked to push a button or say a keyword in front of a spoken interaction. We cannot expect a user to manually segment before speaking an emotional phrase. And we also might want to interpret the statements not directly uttered to the system. A problem of spotting for emotional phrases therefore is the temporal segmentation. Like this we decided to use an open microphone mode. In this mode every speech-like noise will be captured due to an energy-based recognition of input. To rely that we only grab utterances of the intended speaker and filter noises we use a speaker verification component.

Speaker Verification

A special requirement in this domain is the assignment of very short samples containing of single words or less.

The challenge here is that usually speaker verification proves robust with long samples free of speechless parts. Another problem occurring with the task is that speakers greatly vary in their spectral shapes when yelling or moaning emotionally. This demands for very robust speaker verification. Since the verification is only introductory to the emotional analysis real time capability is a further requirement. Our solution takes as features 256 long-term spectral coefficients transformed by a Fast-Fourier-Transform. These are achieved by band pass-filtering of the signal between 20-8000 Hz, pre-emphasizing the signal to accentuate the rather speaker dependent higher frequencies and averaging over short 10ms frames obtained by windowing the signal with a Hamming window.



$$H_{pre}(z) = 1 - a\,z^{-1}$$

Eq. (2): Pre-emphasis of the signal

Pauses are being eliminated according to the signal energy. After Mel-filtering with triangular filter shapes the remaining 20 Mel-frequency coefficients of the pseudo-long-term-spectrum are being classified by a continuous one-state Hidden-Markov-Model with Gaussian Mixture Models. Even for very short phrases compromising of only one word or syllables this solution proved very robust at 2% false rejection rate and 4% false acceptance rate tested with 16 male and 2 female speakers and 500 utterances at optimal parameters. The same principle is also used for speaker recognition at 98% recognition rate for the same 16 test persons. The instance is than trained with several speakers and decides for the maximum score among the models. This principle enables a potential system to automatically load correct user models for adaptation.

Confidence thresholds

A second way to filter noise is the introduction of two confidence thresholds. Provided the scores the maximum score must exceed a lower threshold to generally regard an emotion as recognized. If a further more restrictive threshold is exceeded the system can be also trained unsupervised. Another least restrictive level can be introduced at which the system could ask the user if its hypothesis is correct.

7. RESULTS AND CONCLUSIONS

The realization of the introduced methods has been tested with 17 speakers, one of them female. In total 595 utterances were collected. This resembles 85 samples per emotion. The emotions were acted, what surely is a disadvantage since users tend to exaggerate when acting. In an initial phase user statements were not recorded to make the probands familiar with simulating emotions naturally. For the classification of prosodic parameters the system was in advance adapted by training with ten samples for each emotion. However, these results can be seen as upper limit for achievable results.

Recognition results

In the following the results are listed in detail for the phrase-based approach, for the semantic evaluation and for the prosodic analysis. In the tables *irr* abbreviates irritation, *joy* joy, *ang* anger, *fea* fear, *dis* disgust, *sad* sadness, and *neu* neutral user emotion. Only results with optimal system parameters are presented.

	irr	joy	ang	fea	dis	sad	neu
irr	**79**	3	1	2	0	0	0
joy	2	**59**	23	0	0	0	1
ang	1	5	**62**	0	8	9	0
fea	0	4	0	**77**	0	3	1
dis	2	1	6	0	**75**	0	1
sad	0	0	16	8	0	**54**	7
neu	0	0	5	0	1	7	**72**

Figure (2): Confusion table of prosodic analysis

The table shows the distribution for prosodic feature analysis. Downwards the acted emotion will be listed, while to the right the recognized emotion can be seen. The next table shows the respective recognition rates.

	irr	joy	ang	fea	dis	sad	neu
rec. rate	**93 %**	**69 %**	**73 %**	**91 %**	**88 %**	**64 %**	**84 %**

Figure (3): Recognition results with prosodic analysis

The total recognition rate is equivalent to 80.3%. The table clearly shows that some emotions are often confused with certain others. Also some emotions seem to be recognized more easily. This may be due to the fact that the test patterns were acted emotions and test-persons have difficulties with feigning certain emotions. The results reach the abilities of a human decider of approximately 80% correct assignment rate.

In the 85 collected samples per emotion only 62.7% of direct phrases could be found. In the remaining 37.3% recordings emotion was expressed non-verbally. The following table shows the distribution of the phrases and the recognition results.

	irr	joy	ang	fea	dis	sad	neu
amount	76	62	73	81	52	68	27
rec.rate	**92 %**	**93 %**	**90 %**	**88 %**	**84 %**	**86 %**	**84 %**

Figure (4): Distribution and recognition results for emotional phrase spotting

The table shows that speakers often do not articulate their emotion in words but rather in sounds. It also demonstrates that some emotions seem to be expressed preferably by this mean. The total recognition rate for understanding emotional phrases was 88.1%. A difficulty clearly lies in the often defective articulation in emotional speech. The remaining semantic features have been tested in interaction. A recognition rate of 62.1% could be observed. The final table shows the result achieved with semantic fusion of semantic and signal characteristics as described earlier.

	irr	joy	ang	fea	dis	sad	neu
amount	76	62	73	81	52	68	27
rec.rate	**93 %**	**84 %**	**82 %**	**91 %**	**88 %**	**83 %**	**85 %**

Figure (5): Recognition results with semantic fusion

Conclusions

The introduced methods build a reasonable emotional interpretation model mostly in their combination. Nevertheless fusion can also downgrade recognition rate in the worst case as for irritation in our study. It could be shown that understanding emotional phrases seems a very promising way. However the combination with prosodic parameters is useful to capture non-verbal expressions. Further semantic features could not be used to satisfyingly detect all accosted emotions, but they also supported robust recognition in the fusion. Finally the fusion was able to resolve ironic phrases by the signal characteristics. Generally the recognition proved rather speaker dependent, but conditioning the system to a new user keeps the system applicable. The concept of integration of models allows the connection of further multimodal input data as general human expressional characteristics like mimic recognition or domain specific data like driving data in a car. The results highly motivate further investigation in this area. In a next step we aim to compare results achieved with continuous Bakis-Hidden-Markov-Models for the signal characteristics. Also integration in an early semantic fusion might improve recognition. Finally more evaluation with spontaneous data will deliver more exact results.

8. ACKNOWLEDGMENTS

The presented work has been greatly supported by the FERMUS project, which is a cooperation of the BMW Group, DaimlerChrysler, SiemensVDO and the Institute for Human Machine Communication at the Technical University of Munich. The project name FERMUS stands for error-robust multimodal speech dialogs.

9. REFERENCES

[1] N. Amir, and S. Ron, “Towards an automatic classification of emotion in speech”, in Proc. of ICSLP, Sydney, Dec. 1998, pp. 555–558.

[2] R. Cowie, and E. Douglas-Cowie, “Automatic statistical analysis of the signal and prosodic signs of emotion in speech”, in Proc. of ICSLP, Philadelphia, Dec. 1998, pp. 1989–1992.

[3] B. Heuft, T. Portele, and M. Rauth, “Emotions in time domain synthesis”, in Proc. of ICSLP, Philadelphia, Oct. 1996, pp. 1974–1977.

[4] L. Yang, “The expression of emotions through prosody”, ICSLP 2001, Beijing, China, Proc. Vol. 1, 2000, pp. 74-77.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, “Emotion recognition in human-computer interaction”, IEEE Signal Processing magazine, vol. 18, no. 1, Jan. 2001, pp. 32–80.

[6] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and Frustration Language in Child-Machine Interactions", Paper Proc. Eurospeech 2001, Proceedings, Scandinavia, 2001, pp. 2675.

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

[8] C. Pereira, “Dimensions of emotional meaning in speech”, ISCA 2000, Canada, 2000.

[9] A. Nogueiras, A. Moreno, A. Bonafonte, and J. Mariño, "Speech Emotion Recognition Using Hidden Markov Models", Eurospeech 2001, Poster Proceedings, Scandinavia, 2001, pp. 2679-2682.

[10] T. Polzin, Verbal and non-verbal cues in the communication of emotions", ICASSP 2000, Paper Proc. ID: 3485, Turkey, 2000.

[11] B. Schuller, F. Althoff, G. McGlaun, M. Lang, "Navigation in virtual worlds via natural speech", HCII 2001, 9th International Conference on HCI, New Orleans, Louisiana, USA, Poster Session Abridged Proceedings, 2001, pp. 19-21.

[12] A. de Cheveigné, H. Kawahara, “Comparative evaluation of F0 estimation algorithms”, Eurospeech 2001, Arlborg, Denmark, 2001.

[13] A. Kießling, "Exktraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung", PhD Thesis, University of Erlangen-Nuremberg, Shaker Verlag, Aachen, 1997, pp. 16.

[14] F. Itakura, Distance Measure for Speech Recognition Based on the Smoothes Group Delay Spectrum", Proc. of the ICASSP 87, 1987, pp. 1257-1260.

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