Multimodal emotion recognition in audiovisual communication

Björn Schuller, Manfred Lang, Gerhard Rigoll

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Manfred Lang, and Gerhard Rigoll. 2002. "Multimodal emotion recognition in audiovisual communication." In *Proceedings: 2002 IEEE International Conference on Multimedia and Expo (ICME 2002), 26-29 August 2002, Lausanne, Switzerland*, 745–48. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICME.2002.1035889.

MULTIMODAL EMOTION RECOGNITION IN AUDIOVISUAL COMMUNICATION

Björn Schuller, Manfred Lang, and Gerhard Rigoll

Institute for Human-Machine-Interaction
Technical University of Munich, Germany
(schuller | lang | rigoll)@ei.tum.de

ABSTRACT

This paper discusses innovative techniques to automatically estimate a user's emotional state analyzing the speech signal and haptical interaction on a touch-screen or via mouse. The knowledge of a user's emotion permits adaptive strategies striving for a more natural and robust interaction. We classify seven emotional states: surprise, joy, anger, fear, disgust, sadness, and neutral user state. The user's emotion is extracted by a parallel stochastic analysis of his spoken and haptical machine interactions while understanding the desired intention. The introduced methods base on the common prosodic speech features pitch and energy, but rely also on the semantic and intention based features wording, degree of verbosity, temporal intention and word rate, and finally the history of user utterances. As further modality even touch-screen or mouse interaction is analyzed. The estimates based on these features are integrated in a multimodal way. The introduced methods base on results of user studies. A realization proved to be reliable compared with subjective probands' impressions.

1. INTRODUCTION

1.1 Potential

Automatic emotion recognition is in the focus of nowadays science. Fields of applications are psychiatric diagnosis, intelligent toys or even detection of lies [1]. Particularly in risky environments as automotive or aerospace a growing interest in the knowledge of the mental user state can be observed. Also the use of nowadays erroneously input modalities like speech or gestures claims for a more natural interaction understanding user emotions. If a system e.g. is provided with the knowledge of a content user it can assume a correct antecedent reaction to the user input. This encourages it to learn without supervision. A dissatisfied user on the other hand initiates implicitly automatic error-recovery programs. If a system realizes that a user needs a maximum of concentration for a side task to the interaction, it should - just like a smart assistant or copilot - not interrupt without a given priority. To give another example, the awareness of an effete or depressed user might lead to the automatic activation of safety routines in high-risk tasks. Spoken language on the other hand is one of the most natural communication forms between human beings. Humans also express their emotion via speech. Enabling systems to interpret spoken utterances for a more intuitive human machine interaction therefore suggests also understanding transmitted emotional aspects. But also in gesturing, emotional aspects can be found.

1.2 Emotional data acquisition

We sampled speech utterances and classified the appurtenant emotion that a human senses when hearing the speech sample and analyzed them. Data was obtained in three different manners: initially we made several usability studies with a minimum of 16 probands. In these studies the test-persons had to control an internet-browser introduced in the oncoming by speech and pen-like gestures. A Wizard of Oz simulated the browser in a more or less cooperative way. The aim was to provoke diverging emotions induced by the system. To obtain more data we also analyzed sample phrases from radio plays. The emotional state of the speaker was classified according to the plot. Finally also acted emotions were used for training and testing of the system.

1.3 Classified emotional states

Nowadays attempts to recognize user emotion mostly classify from four [3], up to seven [2], or eight [4] different emotions. We classify the states defined in the MPEG 4 standard plus a neutral state: surprise, joy, anger, fear, disgust, sadness, and neutral user state.

2. MULTIMODAL ENVIRONMENT

The environment of the emotional studies is an internet-browser controlled via natural speech and mouse- or touch-screen interaction. It was chosen due to its multimodal nature and its far known usage. The haptic channel consists of conventional selection via buttons and additional dynamic gestures on the screen as partially user defined shortcuts. The gestures can either be performed by directly touching the screen with the finger, or by using the mouse. A common graphical user interface and pre-recorded speech were used as output.

2.1 Natural speech component

Natural speech is analyzed in a top-down one-pass stochastic decoder topped by a rule-based intention layer as introduced in [5]. The recognition rate of spoken user utterances exceeded 90% in the limited domain. The correct assignment of semantic

concepts even exceeded 98%. The users were allowed to speak freely and in whole sentences, with one basic restriction: no usage of content specific wording was allowed. The selection of links had to be performed via local or numerical reference.

2.2 Touch- or mouse gesture recognition

The gesture recognition is realized by a DTW algorithm with Itakura [6] local constraints and Euclidean distance metric. The features are the x and y coordinates normalized by a bounding box and their first and second order derivatives. For continuous control of browser functions as scrolling, also the z-axis can be used. To obtain a z-value, a surface acoustic wave touch-controller was chosen. By the amount of wave-absorption the strength of the touch is estimated. The engine is capable to distinguish between 40 pen-like gestures at 98.21% recognition rate. As only constraint a gesture has to be made in one stroke. Nevertheless a temporal segmentation basing on the z-value exists to avoid loss of the gesture path at light finger-lifts.

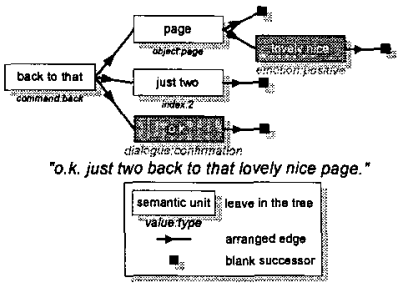
3. INTENTION BASED SPEECH FEATURES

There are a limited number of possible characteristics that transmit emotional aspects in speech. While nowadays attempts in general concentrate on the signal near prosodic characteristics [2,3,4,7,8,9] we also consider the evaluation of semantic and prosodic aspects. Basing ideas of this principle can also be found in [10]. On the semantic-syntactic level the spoken words and phrases themselves are clear reference of the mental user state, if the user is of talkative nature. We believe that it is an obvious first step to let a system understand emotional hints given by the user himself. The level of verbosity, the dialog history and the rate of intentions also carry information about the inner user state. The features used on the higher semantic levels are explained in detail in the following sub-chapters.

3.1 Emotional Phrases

When automatically understanding a user's utterance, emotional aspects can be interpreted on the semantic-syntactic layer, too. Recent studies [10] clearly show, that around 4% of machine interaction utterances are extraneous, i.e. out of domain, carrying emotional clues for e.g. excitement, disappointment, frustration or happiness. But also in modal questions or other formulations emotional clues can be found. In our research emotional markers proved most promising. We execute a parallel analysis of two different aspects of spoken utterances in one recognition process: domain specific and emotional. The idea is to spot for emotional phrases on the basis of a stochastic speech-understanding unit as introduced in [1]. A user utterance is thereby split into semantic units to which semantic concepts are assigned. These concepts are sorted hierarchically in a semantic tree structure. Especially added emotion concepts are isolated in the latter interpretation process and are in general capable to allude all introduced emotional states. It is obvious that the key-danger lies in irony. A misinterpretation of ironic statements shall be avoided by the integration of further features as introduced in the following. The following example was logged in a user study where users had to control the described browser. It shows a semantic-syntactic tree with highlighted emotional concepts.

Figure 1: Example of a semantic tree



3.2 Verbosity level

Each intention a user utters to a system possesses a minimum of semantic units needed to cover all parameters of the desired system action depending on its complexity. By building the relationship with the actual total number of semantic units in the utterance we measure the degree of verbosity at a given time instant. Users tended to talk very verbosely in a good mood or relaxed situation.

3.3 Intention rate

The total amount of intentions in a set time interval is the definition of the intention rate. This feature bases on the results of a study where users had to cope with a very non-cooperative system in stress situations. After measuring their control condition for a collaterally task their level of cognitive workload was adapted to individual skills. The probands showed a very high intention rate in displeased situations.

3.4 Dialog history

The dialog history indicates on the semantic-syntactic level the number of repetitions and contradictions over a given interval leading to repetition and contradiction rates as further features. The results found in [10] are as well legitimating the use of these features. They can allude interference in the communication between human and machine due to a distracted user. This in general induces a negative mental user condition. A high number of repetitions can also be an evidence for a tired user.

3.5 Word rate

The total amount of semantic units in a given time period defines the word rate. To estimate the word rate, the knowledge of the correct uttered statement of the user is needed. This is in general not guaranteed due to recognition errors in speech understanding. At least a very good estimate can be assumed. Depressed users tended to speak very slowly, while they talked relatively fast in stress situations. Clues for joy could also be found in this feature in our studies.

3.6 Classification

Due to the different temporal nature of intra- and inter-phrase specific aspects timing had to be regarded carefully. The classification of the intention-based features is done by Euclidean

distance metrics. A feature vector consists of the number of emotional phrase occurrences per emotion, the intention-, contradiction-, repetition and word-rate and the degree of verbosity. In general the classification is phrase-wise. History-based features are included in the feature vector by the amount of forgoing occurrences.

4. SIGNAL BASED SPEECH FEATURES

Most actual research uses prosodic features [2,3,4] to classify emotions in speech. These rather signal near characteristics are among others energy and pitch. Different classifications as distance metrics, LDA, HMMs and neuronal networks are used for the classification. Our first results on these layers are achieved by use of a DTW with local shifts and global Itakura constraints. Euclidean and City-Block metrics have been used for the evaluation. The use of Distance Metrics however seems reasonable regarding the results of [3]. In a second approach we used continuous Bakis-Hidden-Markov-Models with Gaussian Mixture Models [11]. The models were trained using Baum-Welch algorithm, evaluation with forward-algorithm delivered best results.

4.1 Pitch

As a window-function we use a Hanning window. We analyze frames every 10 ms. F_0 is estimated by maximum-search in the average magnitude distance function, which bases on auto-correlation, but is optimized for integer-logic. This method proved robust against noise but less suited to distinguish between further formants. As far as pitch is concerned we decided for local pitch features due to the highly speaker- and phrase-type [12] dependant global features. We use instantaneous pitch values and their first and second order derivatives.

4.2 Energy

We calculate the instantaneous values of energy by the logarithmic mean energy in a frame. To eliminate influences of the absolute value we use the first and second order derivatives.

4.3 Classification

The features are smoothened by moving average filtering. Further on they are freed of their mean value and normalized to their overall standard deviation. Classification of the five-dimensional feature vector is done by evaluation of continuous Bakis-Hidden-Markov-Models with Gaussian-Mixture-Models. The results presented later in the paper were achieved using 64 states and five mixtures. No remarkable increase in recognition results could be observed using more mixtures or states.

5. HAPTICAL FEATURES

While most research efforts are spent on emotion recognition by the speech signal, we also consider evaluation of haptical user behavior. Similar to the parallel recognition of speech while analyzing emotional aspects, an analysis of the emotional aspects in the pen-style touch-gestures is performed. The features for the classification are the absolute value of a pointer, the z-value itself and the tangent of the movement-phase in the screen plain in polar coordinates, each achieved by a transform of the Cartesian coordinates x, y and z. While the amplitude can be interpreted as the energy of a gesture-stroke, the phase information can be used as a clue for the jitter. The integration of the z-coordinate seems important considering the touch energy. Further on the first and second order derivatives are used. When mouse interaction is regarded, the feature complexity is reduced by the z-component.

5.1 Amplitude

The amplitude and the first and second order derivatives serve as energy features. According to the emotional state user seemed to push the screen harder or less hard. But also jitter in the energy curve alluded different typical behaviors.

5.2 Phase

The phase seems an appropriate indicator for a trembling hand. The derivatives of the phase in first and second order contributed to a more robust interpretation.

5.3 Classification

The nine dimensional feature vectors were classified using DTW as described in chapter 2.2.

6. MULTIMODAL INTERGRATION

6.1 Single signal-analyzing instances

The introduced semantic and signal based speech and the gesture features are evaluated each in a single stochastic signal-processing instance. They can be used on their own as isolated engines 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.

6.2 Late Semantic Fusion

The single instances are integrated in a late semantic fusion. By allowing each instance to impart a score for each emotion early decisions are avoided. A hypothesis is calculated for each emotion by averaging the single scores with their corresponding confidences and a-priori probabilities. A maximum likelihood decision takes part after the integration. No direct exchange between semantic and prosodic features takes place, which is still a disadvantage. This shall be ensured by an upcoming one-pass solution, e.g. early semantic fusion.

6.3 User profiling

User conditioning allows for generalization and applicability of the applied models for different cultural comprehension of emotional aspects. User profiling can at present only be achieved with supervision. The training influences an inter recognition engine matrix of individual weights for the integration. Additional user training adapts intra instance specific model parameters like reference models for the phrase spotting or control conditions for the touch-interaction. Profiling can be done in a playful way by letting the system ask the user at a

detected emotional change about his feeling in an initialization phase. After first data collection the system can initiate more direct dialogs by further interrogation.

6.4 Integration of contextual knowledge

Also the users' intentions themselves can contribute to a safe thesis of her mental state. In the application introduced the content of visited websites can indicate a certain mood. The goal of future studies is to illuminate the effect of further contextual knowledge integration.

7. RESULTS AND CONCLUSIONS

The following table shows recognition results of acted and captured emotional input of one speaker. 520 phrases were used for the evaluation with a minimum of 60 phrases per emotion. The touch-interaction data was achieved mainly of two users. Further investigation for more detailed recognition results of this technique is in progress.

Figure 2: Table of recognition results

	Speech signal features	Speech intention features	Gesture interaction	Multimodal fusion
Surprise	0.83	0.81	0.44	**0.85**
Joy	0.85	0.84	0.85	**0.89**
Anger	0.81	0.85	0.71	**0.86**
Fear	0.73	0.78	0.49	**0.81**
Disgust	0.81	0.83	0.72	**0.82**
Sadness	0.72	0.81	0.54	**0.76**
Neutral	0.95	0.84	0.87	**0.95**

Acceptance tests with 15 users showed a classification potential of more than 80% recognition rate, which is similar to a human decider. 8,3% felt that emotion recognition is a must have, 16,6% judged it good, 16,6% more thought that it was a good thing to have, but were afraid of data security, 33,3% judged it interesting but unimportant, 8,3% believed it frightening, 16,9% were not sure how they would judge it. The ideas how a system as the introduced browser should react due to user emotions were mainly emotion suiting content loading in the web, automatic help after a detected surprise, or error-recovery dialogs after anger or frustration. Combined, the introduced methods build the most reasonable emotional interpretation model. Conditioning the system to a new user is not necessary but leads to an improved recognition. Only speech can be used to detect all accosted emotional states as standalone solution. The concept of multimodal integration allows the connection of further input data as general human expressional characteristics like mimic recognition or domain specific data like the driving data in a car. Compared with nowadays focus on prosodic features our results show that the integration of semantic aspects in speech emotion analysis is most promising. It could also be shown that the innovative analysis of touch- or mouse interaction allows for estimation of a users' emotion. The final recognition results highly motivate further investigation in the interaction of further modalities.

8. ACKNOWLEDGEMENTS

The contents discussed in this paper largely benefits from the collaboration with the student assistant Stephan Reiter.

9. REFERENCES

[1] 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, pp. 32–80, Jan. 2001.

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

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

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

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

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

[7] N. Amir, and S. Ron: *"Towards an automatic classification of emotion in speech,"* in Proc. of ICSLP, Sydney, pp. 555–558, Dec. 1998.

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

[9] B. Heuft, T. Portele, and M. Rauth: *"Emotions in time domain synthesis,"* in Proc. of ICSLP, Philadelphia, pp. 1974–1977, Oct. 1996.

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

[11] L. Rabiner: *"A tutorial on Hidden Markov Models and selected applications in Speech Recognition"*, Proceedings IEEE, pp. 257-284, February 1989.

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