The Recognition of Emotion

Anton Batliner[1], Richard Huber[1], Heinrich Niemann[1], Elmar Nöth[1], Jörg Spilker[2], and Kerstin Fischer[3]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2] Lehrstuhl für Künstliche Intelligenz, Universität Erlangen-Nürnberg, Germany
[3] Institut für Informatik, AB NatS, Universität Hamburg, Germany

Abstract. To detect emotional user behavior, particularly anger, can be very useful for successful automatic dialog processing. We present databases and prosodic classifiers implemented for the recognition of emotion in Verbmobil. Using a prosodic feature vector alone is, however, not sufficient for the modelling of emotional user behavior. Therefore, a module is described that combines several knowledge sources within an integrated classification of trouble in communication.

1 Introduction

Research on the automatic processing of speech and language in the last ten years showed the beginning integration of prosody and a broadening of the view towards more sophisticated dialog systems. It might well be that for the next decade to come, the integration of emotion into such systems will be one of the pivotal topics. Applications that can be imagined are: automatic dialog systems, especially within call centers, operation of machines, especially car navigation, monitoring of pilots, interactive computer games, and the whole range of topics which is further connected with the term “affective computing”, cf. Picard (1997). In the Verbmobil system, which aims at automatic translation in a machine-mediated human-to-human communication emotion does, however, not yet play a crucial role because normally, speakers in such a dual communication understand each other and do not blame the partner if the system does not translate correctly. Moreover, they do not display their dissatisfaction with the system, if it is not functioning satisfactorily, because they do not want their emotional behavior to be translated. This can at least be observed during the demonstration of the system—it might be different if Verbmobil really was running as a real life application. The recognition of emotion is therefore not fully integrated in the Verbmobil system but can be switched on for demonstration purposes as an add-on-feature: If this special “emotion mode” is on, the prosody module, cf. Batliner et al., in this volume, does not only classify boundaries, accents, and sentence mood, but emotion as well. If the user utterance is classified as “emotional”, i.e., as “angry”, the system switches into a special mode: a big smile is displayed on the screen, the utterance is not processed any longer and the user can continue with another utterance. The research on emotion will be continued in the SmartKom project (1999–2003) where mimic will be recorded as well and used as a further knowledge source. In this paper, we will present on the one hand experiments which are tailored to the demonstration of such a restricted recognition of emotion in Verbmobil, on the other hand, we will describe briefly a module that can be used for real life applications in the future.

2 Prosodic Classification of Emotion

2.1 Databases

In a first step, data were collected from a single, experienced acting person (ACTOR data). These data comprise 1232 "neutral" turns produced within the Verbmobil scenario that were collected for reasons independent of the aims of this study, and 96 turns in which the speaker was asked to imagine situations in which the Verbmobil system was malfunctioning and in which he was getting angry, for instance: *Das ist doch unglaublich!* (That's really unbelievable!) In a second step, data were elicited from 19 more or less "naive" subjects who read 50 neutral and 50 emotional sentences each (the emotional sentences were a subset of the emotional utterances produced in the ACTOR scenario). These READ data will be treated together with the ACTOR data in this paper as PROMPTED data; for a separate treatment, cf. Batliner et al. (2000).

In a third, more elaborate, step, a Wizard-of-Oz (WOZ) scenario, cf. Fraser and Gilbert (1991), Pirker and Loderer (1999), was designed to provoke reactions to probable system malfunctions and to control the speakers' changes in attitude towards the system, i.e. their emotional behavior, over time; controllability is achieved by a fixed schema according to which the simulated system's output is produced; thus, recurrent phases are defined which are completely independent of the speakers' utterances and which are repeated several times throughout the dialogs such that the speakers' reactions to the same system output can be compared over time. The speakers are thus confronted with a fixed pattern of messages of failed understanding, misunderstanding, generation errors, and rejections of proposals, which recur in a fixed order. The impression the users have during the interaction is that of communicating with a malfunctioning automatic speech processing system. The changes in linguistic behavior, supported by results from a questionnaire speakers fill out after the recording, are interpreted as changes in speakers' attitude towards the system, i.e. as increasing anger.

Data used for the experiments reported in this paper are 20 dialogs (2254 turns); recording, transcription, and annotation are still going on. The goal is to record about 70 dialogs of approximately 25 minutes length each. All of the dialogs involved have been or will be annotated according to lexical, conversational, and prosodic peculiarities in the same way, cf. Fischer (1999). The following examples from a dialog show how the speakers' linguistic behavior differs in reaction to the same system utterance which is in both cases completely irrelevant regarding the speakers' previous utterance. While in the first occurrence the speaker reacts cooperatively and reformulates his proposal, he insults the system the second time after some interaction with the system and simply repeats his previous proposal. Furthermore, in the first reaction, no lexical and prosodic peculiarities are found and the conversational behavior can be classified as using metalanguage, i.e. fairly cooperative conversational behavior. This has been annotated as *@030@* at the beginning of the turn (first digit: lexical, second digit: conversational, and third digit: prosodic marking; no special marking is labelled as "0"). In contrast, in the later reaction to the systems utterance, the speaker uses a swear word, which is marked as lexical peculiarity, he insults the system, which is marked as a conversational irregularity, and by means of several prosodic peculiarities, such as very clear articulation (*2) and pauses between the words (*4); the annotation at the beginning of the turn thus shows *@590@* where the zero holds for all of those words in the turn which are not prosodically marked otherwise:

WoZ: *ein Termin um vier Uhr morgens ist nicht möglich.* (an appointment at four am in the morning is not possible)

user: *@030@ brauchen wir auch nicht, weil wir haben Zeit von acht bis vierzehn Uhr.* (that's not necessary since we have time from eight am to two pm)

...

WoZ: *ein Termin um vier Uhr morgens ist nicht möglich.* (an appointment at four am in the morning is not possible)

user: *@590@ deshalb machen wir ihn ja auch um acht, du Schnarchsack *2. fünfter *4 Januar *4, acht *2 bis *2 zehn *2.* (that's why we make it at eight, you snore-bag. fifth of January, eight to ten.)

2.2 Methods

For our classifiers, we use basically the same features as in the prosody module, cf. Batliner et al., in this volume, which model pausing, fundamental frequency F0 and-normalized as for their mean values across a large database—energy, speaking rate, and duration. As **word-based** features, we use 91 prosodic features per word, modelling the word itself and a context of two words to the left and to the right. For the **global** features, the time window considered is the whole utterance: There is only *one* time interval which starts at the *first voiced* frame of the whole utterance and ends with the *last voiced* frame of the utterance. F0 onset and offset position do not make any sense for such a computation and are thus omitted; in addition, we calculate the standard deviation of the F0 values. Two different kinds of global features are computed: first, features that are normalized as for their mean values, and second, features that are computed without any segmental/word-based information whatsoever. These features are described in more detail in Batliner et al. (2000), the normalization is described in Batliner et al., in this volume.

In our experiments, we classify utterances as "emotional" (class E3) and as "neutral" (class E0); these indices are chosen in analogy to our boundary and accent labels, cf. Batliner et al., in this volume. "Emotional", that is, "angry", turns are given trivially in the PROMPTED scenarios. For the WOZ data, we label all those turns as "emotional" that are annotated with one or more prosodic peculiarities; for more details, cf. Batliner et al. (2000). For the word-based features, we labelled and classified every utterance on the word level, cf. Huber et al. (1998). Each word in the emotional utterances is labelled as belonging to the class E3 and each word of the neutral utterances is labelled as belonging to the class E0. We transformed the spoken word chains into Word Hypotheses Graphs (WHG); the output of the emotion detector is a prosodically scored WHG, cf. Kompe et al. (1995). Special neural networks, i.e. Multi-Layer-Perceptrons (MLP) were trained with different topologies using r-prop as training algorithm. For every word of the WHG we calculate a feature vector that is used as input vector of the MLPs, so that the number of nodes in the input layer is exactly the number of vector coefficients. For the test every word of the utterance was assigned a probability $P(E_{3,i})$ and $P(E_{0,i})$ for the classes E3 and E0 by the MLP. According to Huber et al. (1998) we calculate the costs $C(Y_1, Y_2, \ldots, Y_n)$ of an utterance with n words $Y_1, Y_2, \ldots, Y_n$ with eqn. (1).

$$C(Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^{n} -\log(P(Y_i)) \qquad (1)$$

With eqn. (1) we can get two costs $C(E_3) = C(E_{3,1}, E_{3,2}, \ldots, E_{3,n})$ and $C(E_0) = C(E_{0,1}, E_{0,2}, \ldots, E_{0,n})$ for every utterance belonging to E_3 and E_0, respectively. If $C(E_3) \leq C(E_0)$ is true, we classify the utterance as emotional, otherwise as neutral. Experiments were run without and with Part of Speech (POS) features which are described in Batliner et al. (1999). For the PROMPTED data, six main POS features for a context of two words to the left and to the right yielding 30 POS features were used, for the WOZ data, we used 13 more detailed POS features modelling only one word because here, a larger context yielded worse results due to sparse data. For the global features, for every utterance we get only one prosodic feature vector and use it as input vector for the MLPs. Two different feature sets were used: for one set, no word-based, segmental information was used for normalization (-segInfo), for the other set, such information was included (+ segInfo). Different MLPs with different topologies and r-prop as training algorithm were trained. For the test every utterance was assigned a probability $P(E3)$ and $P(E0)$ belonging to the class E3 and E0 respectively. If $P(E3) \leq P(E0)$ is true, the utterance is classified as emotional, otherwise as neutral.

2.3 Results

We divided the 3228 turns of the PROMPTED data set into the four data sets *train*, *validation*, *test-seen*, and *test-unseen*. The data set *test-unseen* contains all 50 emotional and 50 neutral turns of three speakers of the READ scenario, altogether 300 turns; these three speaker (2m/1f) were not used for training or validation. Five emotional and five neutral turns from each of the other 16 speakers (12m/4f) constitute *test-seen*. Furthermore we added 48 emotional and 612 neutral turns of the ACTOR scenario to *test-seen* yielding altogether 820 turns (128 emotional and 692 neutral). The data set *validation* contains 10 emotional and 124 neutral turns from the ACTOR scenario and three emotional and three neutral turns of *each* of the other 16 speakers of the READ scenario, altogether 230 turns. The *validation* set is used after the training of the different MLPs for the selection of the optimal network topology.

The *train* set contains 38 emotional and 496 neutral turns of the ACTOR scenario and 42 emotional and 42 neutral turns of every of the remaining 16 speaker of the READ scenario, altogether 1878 turns. These data are used for training of the different MLPs.

For the WOZ scenario we also generated four data sets *training*, *validation*, *test-seen*, and *test-unseen*. *test-unseen* contains the utterances of five speakers (4m/1f) with 245 emotional and 345 neutral utterances, altogether 590 utterances. *test-seen* contains 170 utterances of the other 15 speakers (7m/8f), 85 neutral and 85 emotional. For the training of the MLPs we used 1184 utterances, 857 emotional and 327 neutral; for validation, we used 310 utterances, 82 neutral and 228 emotional. All results are given in Table 5.

Table 1. Recall in percent for the different constellations

PROMPTED **scenario**								
	word-based features				27 global features			
	91 [– POS]		121 [+ POS]		[– segInfo]		[+ segInfo]	
	E3	E0	E3	E0	E3	E0	E3	E0
test-seen	74	95	69	94	57	68	79	94
test-unseen	85	82	84	81	72	31	86	69

WOZ **scenario**								
	word-based features				27 global features			
	91 [– POS]		104 [+ POS]		[– segInfo]		[+ segInfo]	
	E3	E0	E3	E0	E3	E0	E3	E0
test-seen	72	71	68	68	69	71	81	55
test-unseen	69	62	56	83	52	59	77	50

For the PROMPTED data, POS information does not contribute to the word-based features, probably because the sentences were uniform (fixed set) in the READ scenario; for the global features, segmental information contributes to a large extent. One would expect that the known speakers of *test-seen* can throughout be better classified than the unknown speakers of *test-unseen*; however, this holds only for E0 whereas the opposite can be observed for E3. For the WOZ data, *test-seen* can most of the time be better classified than *test-unseen*; this meets our expectations. With POS information, only E0 for *test-unseen* can be classified better. With segmental information, E3 but not E0 can be better classified for both *test-seen* and *test-unseen*. All in all, the classification rate is in the range of 80% for the PROMPTED data and some ten percent less for the WOZ data.

Reasons for the observation that speakers use prosody less in the WOZ data may be firstly that actors display emotions overtly because they have been asked to do so (ACTOR scenario). This needs not be the case for normal speakers. A second reason for the different results may be that in read speech (READ scenario), to use prosody is the only strategy available, i.e. the only cue that can be varied. In the WOZ scenario speakers are not restricted to the use of prosody alone but can choose among a number of different strategies available. Thus, speakers in the communication with artificial communication partners, unlike in the ACTOR and READ situations, may use different communicative strategies besides the use of prosody. Thus we distinguish between two classes of strategies: on the one hand those which are rather **context-independent**, such as the use of prosody, mimic, or lexical features, in particular swear words; on the other those which are **context-dependent**, that is, which are constituted only within a sequence of turns, such as the use of repetitions. The context-dependency of these strategies is already indicated by the prefix *re–* in *reformulation* and *repetition*. Thus, our search for (prosodic) indicators of emotions has to be replaced by a search for any indicator of TROUBLE IN COMMUNICATION. This means that we have to combine the prosodic classifier, and, if available, a classifier of mimic, with other knowledge sources, such as the modelling of dialog act sequences, the recognition of repetitions, key word spotting (swear words), and the recognition of out-of-domain sequences (meta-communication, speaking aside), cf. Oviatt et al. (1998a), Oviatt et al. (1998b).

3 Broadening the View

In this section, we sketch our module **M**onitoring **o**f **U**ser **S**tate [especially of] **E**motion MOUSE; Figure 1 gives a rough outline: in the communication of the system with the user, the user behavior is supposed to mirror the state of the communication. If there are no problems (felicitous communication) or if there are only minor problems (slight misunderstandings) which can be solved, the user behaves neutral and is not engaged emotionally. If, however, there are severe recurrent misunderstandings, i.e. error "spirals", cf. Levow (1999), that is, if there is TROUBLE IN COMMUNICATION, then the user behavior changes accordingly; it is marked: overt signalling of emotion—changes in prosody, mimic, etc.—and particular, context-dependent strategies, i.e. different strategies to find ways out of these error spirals, can be observed. If there is such trouble, our module MOUSE should trigger an action, for instance, by initiating a clarification dialog, cf. Figure 2. In such a case, the communication will recover gracefully. If, however, no action is taken, chances are that the user becomes more and more frustrated, and sooner or later he will break off the communication (dead end, point of no return).

In Figure 2, the architecture of MOUSE is sketched in more detail. The components that are already implemented are highlighted. Starting point has to be a user independent training based on data that are as close to the intended application as possible. For the training of the "normal" modules other than MOUSE in an automatic dialog system, such as word recognition, "neutral" and "emotional" data are processed together; for the training of the classifier of emotionality, two separate classes have to be trained. For the actual use of this module, it is advantageous to use a clearly defined neutral phase for adaptation of the system. For each of the pertaining phenomena that can be found, a separate classifier is used whose output is a

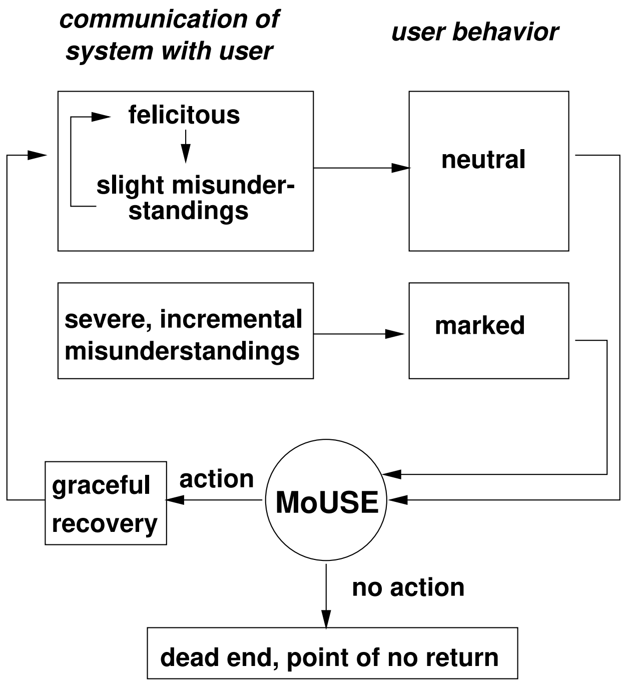

Fig. 1. MoUSE: General outline

probability rating. All probabilities are weighted and result in one single probability that triggers an action if it is above a certain value. This value has to be adjusted to the special needs of the application, for instance, whether one wants to get a high recall or a high precision, or whether both should be balanced. (If the costs of failing to recognize emotions are high—for instance, if important costumers will be lost—recall should be high, even if there are many false alarms and by that, precision is low.) Retraining and a different weighting of classifier results may also be necessary for adaptation to different scenarios. The action invoked can at least be one of the three possibilities: Easiest is probably to return to a very **restricted, system-guided dialog**; a **clarification dialog** needs more sophistication; to **hand over to a human operator** means to cut off automatic processing but, of course, it is the most secure strategy to yield graceful recovery of the communication and thus a neutral behavior of a content user.

With a combination of the prosodic classifier and a classifier that takes reformulations and repetitions into account, in preliminary experiments, the error rate for E3 was reduced by 38%, and for precision, by 32%; for details, cf. Batliner et al. (2000).

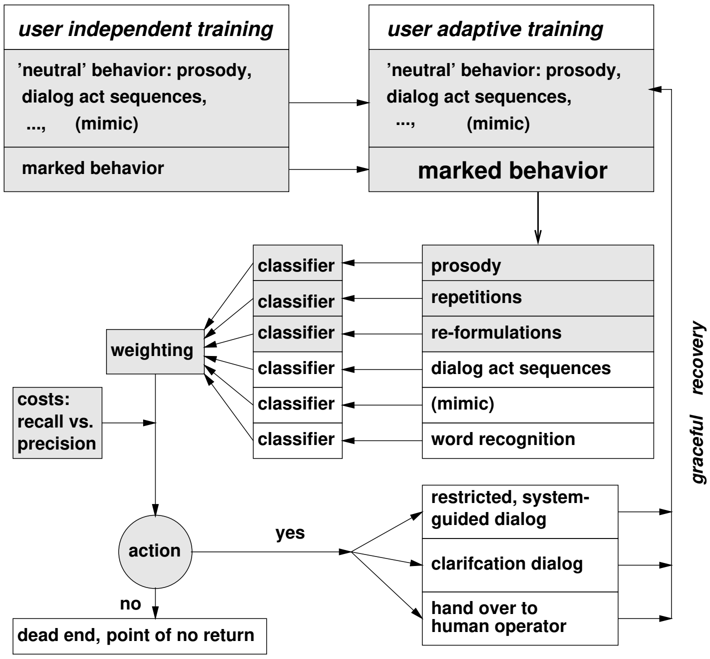

Fig. 2. MoUSE: A sketch of the architecture

4 Conclusion

It turned out that prosody alone is a fairly good indicator of emotion (here: anger) as long as subjects do not have other possibilities to express it: recognition results for the PROMPTED data were significantly better than those for the WOZ data. We are thus faced with a well-known problem: the closer we get to the constellation we want to model (dialog between automatic speech understanding systems and "naive" users), the worse our recognition rates will be. The solution is to look for all kinds of indicators of trouble in communication. The model resulting is already implemented in parts in the module Monitoring of User State.

References

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., and Niemann, H. (1999). Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, 519–522.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2000). Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In *Proceedings of the ISCA Workshop on Speech and Emotion*, (to appear).

Batliner, A., Buckow, J., Niemann, H., Nöth, E., and Warnke, V. The Prosody Module. In *this volume*.

Fischer, K. (1999). Annotating Emotional Language Data. Verbmobil Report 236.

Fraser, N., and Gilbert, G. (1991). Simulating Speech Systems. *Computer Speech & Language* 5(1):81–99.

Huber, R., Nöth, E., Batliner, A., Buckow, J., Warnke, V., and Niemann, H. (1998). You Beep Machine–Emotion in Automatic Speech Understanding Systems. In *Proceedings of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, 223–228. Brno: Masaryk University.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E., Zottmann, A., and Batliner, A. (1995). Prosodic Scoring of Word Hypotheses Graphs. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, 1333–1336.

Levow, G. A. (1999). Understanding Recognition Failures in Spoken Corrections in Human-Computer Dialog. In *Proceedings of the ESCA Workshop on Dialogue and Prosody, September 1rst - 3rd, 1999, De Koningshof, Veldhoven, The Netherlands*, 193–198.

Oviatt, S., Bernard, J., and Levow, G.-A. (1998a). Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41(3-4):419–442.

Oviatt, S., MacEachern, M., and Levow, G.-A. (1998b). Predicting Hyperarticulate Speech During Human-Computer Error Resolution. *Speech Communication* 24:87–110.

Picard, R. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Pirker, H., and Loderer, G. (1999). I Said "two ti-ckets": How to Talk to a Deaf Wizard. In *Proceedings of the ESCA Workshop on Dialogue and Prosody, September 1rst - 3rd, 1999, De Koningshof, Veldhoven, The Netherlands*, 181–186.