Recognition of Emotion in a Realistic Dialogue Scenario

R. Huber, A. Batliner, J. Buckow, E. Nöth, V. Warnke and H. Niemann

University of Erlangen-Nuremberg
Chair for Pattern Recognition
Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
huber@informatik.uni-erlangen.de

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

Nowadays modern automatic dialogue systems are able to understand complex sentences instead of only a few commands like *Stop* or *No*. In a call-center, such a system should be able to determine in a critical phase of the dialogue if the call should be passed over to a human operator. Such a critical phase can be indicated by the customer's vocal expression. Other studies prooved that it is possible to distinguish between anger and neutral speech with prosodic features alone. Subjects in these studies were mostly people acting or simulating emotions like anger. In this paper we use data from a so-called **W**izard **o**f **Oz** (WoZ) scenario to get more realistic data instead of simulated anger. As shown below, the classification rate for the two classes "emotion" (class E) and "neutral" (class ¬E) is significantly worse for these more realistic data. Furthermore the classification results are heavily speaker dependent. Prosody alone might thus not be sufficient and has to be supplemented by the use of other knowledge sources such as the detection of repetitions, reformulations, swear words, and dialogue acts.

1. INTRODUCTION

Present automatic speech dialogue systems try to communicate with the user in a natural way. Instead of permitting only a few commands like *Stop*, *No* and *Yes* it is now possible to communicate with complete and complex sentences. For example, the sentence *"Which movies are shown in the cinema Cinestar today in the evening"* can be processed correctly: the system understands that the user wants to go to the cinema Cinestar today evening between 7 p.m. and 9 p.m. and it presents all movies which start at this time.

If the system does not understand, however, people get angry. If this happens in a call-center and the customer hangs up, the call-center might loose this customer for ever. Therefore it is important to detect automatically that the user is getting frustrated and to initialize a clarification dialogue or to refer to a human operator.

Apart from anger there are some other well-known emotions, like fear, surprise or sadness. As described in [11] they also can be detected with acoustic and prosodic features alone. For the application of the emotion detector in speech understanding systems, however, they will most probably not be relevant. We therefore concentrate our work on the detection of anger or frustration and will only distinguish between a neutral and an angry emotional state.

The data used in our research is described in Section 2. The prosodic features used for the classification of *anger* versus *non anger* are described in Section 3, experimental results are given in Section 4.

2. DATABASE

In most studies data from an *actor scenario* are used. That means that some people, sometimes actors, sometimes students, are asked to read the same sentence simulating different emotions like *fear*, *anger*, *happiness* or *sorrow*. The task in most of these studies is to distinguish between the different emotions with the help of prosodic features like fundamental frequency, energy contours or speaking rate [12, 13, 11, 6, 1, 10]. With data from such an *actor scenario* such so-called *basic emotions* can be classified automatically.

For the two classes *emotion* (class E) and *neutral* (class ¬E) with data from an *actor scenario* and the use of prosodic features and neural networks as classifier we achieved a classification rate of 87% for a test set with unknown speakers [8].

To get more *realistic* data we use data from a so-called **W**izard **o**f **Oz** (WoZ) scenario. In this scenario, several people were asked to schedule 10 appointments with an automatic dialogue system in 30 minutes. The automatic system was simulated by an operator (the wizard) sitting next door. At predefined steps in the dialogue, the wizard's behavior changed. The goal of the experiment was to provoke emotional reactions (in this scenario *angry* re-

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under the Grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

actions) of the user in a well structured schema; recurrent phrases are defined which are completely independent of the speaker's utterances and which are repeated several times throughout the dialogues such that the speakers' reactions to the same system output can be compared over time [7].

For the experiments described in this paper, 39 dialogues with 39 different speakers (20f/19m, 8.30 hours of speech) were used. All 39 dialogues are annotated as for lexical, conversational, and prosodic peculiarities [7]. The database contains 4684 utterances (turns) with 46845 word tokens and 1247 word types, i.e. different word forms. In this paper we deal only with the prosodic peculiarities which are annotated in this database. There are ten different peculiarities annotated with digits from zero to nine. Zero is chosen if there is no prosodic peculiarity in this turn, although there is maybe a lexical and/or conversational peculiarity; for detailed information about the annotation, cf. [7].

In this WoZ scenario the speaker's linguistic and prosodic behavior can completely change, although the system's response is the same. In the following examples the lexical, conversational and prosodic annotation is given as digits between @ signs (in this order; 0 means always no peculiarity). In the first example, the speaker reacts cooperatively, i.e. he shows no anger, and reformulates his proposal. He uses no lexical or prosodic peculiarities, his conversational behavior can be classified as 'using meta-language' and is annotated with 3, so the emotional annotation at the beginning of this utterance is *@030@*. In contrast, in the later reaction, the speaker uses a swear word (snore-bag) and insults the system. The swear word is marked as a lexical peculiarity (5) and the insult is annotated as a conversational peculiarity with 9. Furthermore the speaker changed his prosodic behavior and used pauses between some words (4). The annotation at the beginning of the turn thus shows *@594@*:

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 2 pm)
.....
WoZ: *ein Termin um vier Uhr morgens ist nicht möglich.* (an appointment at four am in the morning is not possible)
user: *@594@ deshalb machen wir ihn ja auch um acht, du Schnarchsack. fünfter Januar, acht bis zehn.* (that's why we make it at eight, you snore-bag. fifth of January, eight to ten.)

3. FEATURES AND CLASSIFICATION

In our experiments we are interested in the classification of whole sentences, i.e. whether an utterance is spoken angry (class E) or not (class ¬E). For the classification we use different feature sets and multi layer perceptrons (MLP) as classifiers, trained with different topologies using r-prop as the training algorithm [14]. A prosodic feature vector is used as input vector for the MLPs. The data set is divided into a training set, a validation set (both used for the training of the MLPs), a test set with turns of speakers which are used for training and validation (*test-seen*) and a test set with unknown speakers (*test-unseen*), i.e. all turns from some speakers who are neither in the training nor in the validation nor in the *test-seen* set. As prosodic features we use features which model pausing, fundamental frequency and – normalized as for their mean values across a large database – energy, speaking rate, and duration. Additionally we use as lexical features 30 *part of speech* flags (POS).

For the **word–based** feature set we calculate a forced time alignment of the spoken word chain to get a word hypotheses graph (WHG) as described in [9]. For every word in the WHG we compute altogether 121 features (91 prosodic and 30 POS flags), modeling the word itself and a context of two words to the left and to the right. Note that the POS flags cover a context of five words, thus the classifier is able to learn a simple pentagram language model. The feature vector with 121 components of every word in the WHG is used as input vector for the MLPs, and every word is classified as belonging to the class E or ¬E; in this case the MLPs will be trained on the word level. As annotation of the emotion on the word level, which we need for the training of the MLPs with word–based features, we use the following simple method: every word of the utterance is labeled as belonging to the prosodic peculiarity, which is annotated at the beginning of the utterance (cf. section 2). Furthermore we define the digit 0 as class ¬E every other label as class E.

For the **global** feature set we calculate the same prosodic features, but not the normalized and the POS features, because here we have no word boundary information. To model speaking rate and the duration of the words, we count the number of voiced and unvoiced frames and regions. Based on this information we compute a few features like the ratio of voiced frames and number of all frames. Altogether we use as global features 27 features and compute for every utterance **one** feature vector which is used as input vector of the MLPs. Thus every MLP will be trained on the sentence level (one feature vector for every sentence). As annotation of the emotion on the sentence level, we label every sentence as belonging to the prosodic peculiarity which is annotated at the beginning of the utterance (cf. section 2). Again, 0 denote class ¬E, all other labels class E. For a more detailed description of the word–based, global and POS features, cf. [5, 3, 4, 2].

Using the word–based features in the classification, every word i of the utterance is assigned a probability $P(E_i)$ and $P(\neg E_i)$ for the classes E and ¬E by the MLP. Following to [8] we calculate the costs $C(Y)$ of an utterance with n words $Y_1, Y_2, \ldots, Y_n$ with eqn.(1).

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

With eqn.(1) we get for every utterance two costs $C(E) =$

$C(E_1, E_2, \ldots, E_n)$ and $C(\neg E) = C(\neg E_1, \neg E_2, \ldots, \neg E_n)$ belonging to E and ¬E, respectively. If $C(E) \leq C(\neg E)$ is true, we classify the utterance as emotional, otherwise as neutral. Using the global features for classification, there will be only one feature vector for every utterance, and every utterance is assigned a probability $P(E)$ and $P(\neg E)$ belonging to the class E and ¬E respectively. If $P(E) \geq P(\neg E)$, the utterance is classified as emotional, otherwise as neutral.

4. EXPERIMENTAL RESULTS

We carried out some experiments with different feature sets and different data sets. In a first experiment we choose at random five speakers (i.e. dialogues) and put all utterances of this dialogues into the test set *test-unseen* (633 turns). The utterances of the remaining dialogues were divide into a *training* set (65% of the turns), a *validation* set (25%), and the second test set *test-seen* (10%, see above). The *training* and the *validation* set were used for the training of MLPs with both word–based and global features. Table 1 shows the recall and precision for the classes ¬E and E on the sentence level for *test-seen* and *test-unseen*, both for the best MLP with global and word–based feature vectors. For the evaluation on the sentence level with the word–based features, we use eqn.(1). It can be seen in

test-seen

27 global feat.			**121 word–based feat.**		
class	rec %	prec %	rec %	prec %	#all
¬E	62	67	**69**	72	193
E	73	67	**75**	73	211
∅	68	67	**72**	73	404

test-unseen

27 global feat.			**121 word–based feat.**		
class	rec %	prec %	rec %	prec %	#all
¬E	61	63	**79**	60	303
E	**67**	65	52	73	330
∅	64	64	**66**	67	633

Table 1: Recall (rec) and precision (prec) for the *test-seen* and *test-unseen* sets of the first experiment with 27 global features and 121 prosodic and POS features in per cent.

Table 1 that for *test-seen*, both recall and precision for the two classes E and ¬E are higher using word–based features instead of global features. The evaluation of *test-unseen* shows, that the average recall using word–based features is higher than the average recall using global features (66% versus 64%, see table 1), but with word–based features, only a recall of 52% for the class E can be achieved. Generally, recall of E for *test-unseen* is markedly worse then for *test-seen*. Thus we believe the prosodic marking of an emotional state like anger is strongly speaker dependent.

We therefore conducted another experiment with all 39 dialogues, a so called *leave–one–out* test (LOO). We divided at random all 39 dialogues into seven partitions with five dialogues, and one remaining partition with four. Next we used each of these eight partitions as *test-unseen*$_i$ ($i = 1, 2, \ldots, 8$), i.e. no turn of these dialogues is used for training or validation. The remainder of the dialogues is divided, like in the first experiment, into a training set (65% of the turns), a validation set (25%) and a test set with known speakers (10%, *test-seen*$_i$). For each of the eight different training and validation sets, different MLPs were trained. Only the global features are considered for these experiments, due to time constraints. For each partition i, the MLP with the highest average of the recall evaluated on the validation set, was selected (MLP$_i$). With MLP$_i$, we evaluate *test-unseen*$_i$. Altogether we obtain eight pairs of recall and precision for the class E and ¬E, respectively (for every *test-unseen*$_i$ one pair for E and one for ¬E) which are listed in table 2. For the LOO experiment the minimum of the recall both for E and, ¬E is 46% and the maximum 72%. The mean value of the recall for ¬E is 61% and the standard deviation is 7.4%. For E, the mean value of the recall is 59% and the standard deviation is 7.6%. Best result both for the recall of ¬E and E can be seen as 72%/60% (partition two). The recognition rates of *anger* versus *non anger* (recall E / recall ¬E) range between 46% (partition five) and 66% (partition two), and the mean value is 60%. Thus Table 2 shows the strong speaker dependent emotional behavior mentioned above.

	¬E		E	
partition	rec %	prec %	rec %	prec %
1	59	46	52	64
2	**72**	65	**60**	67
3	68	57	62	73
4	65	64	62	64
5	46	67	46	26
6	56	64	72	65
7	59	64	65	61
8	63	55	55	63
∅	61	60	59	60

Table 2: Recall (rec) and precision (prec) for the classes E and ¬E for every *test-unseen*$_i$, $i = 1, 2, \ldots, 8$, of the LOO–training in per cent.

Note that in this approach *every* word of an utterance is labeled with label ¬E or E, depending at the label at the beginning. But usually even in an emotional utterance not every single word is spoken emotionally, most of the time just a few.

Thus the MLPs will be trained with incorrect data, because these words are partially labeled incorrectly. There are a couple of reasons, why it is difficult to get correctly labeled words and sentences. First it is very time consuming, because you have to listen to each single word and to decide if the word is pronounced neutral or with a prosodic peculiarity. On the other hand, even with correctly labeled words, it is difficult to label and to evaluate whole sentences. For example an utterance has ten words and seven belong to the class ¬E and three to the class

E. Assuming the MLP classifies all words correctly, there are seven words classified as ¬E and three classified as E. The question is now how to classify or to label the whole sentence? Maybe the whole sentence indicates anger, even though most of the words are spoken prosodically normal. In this case the utterance is labeled and classified correctly as E. But if another utterance has ten words and all ten belong to the class ¬E, and are labeled with ¬E and the MLP classifies only seven of them as ¬E, the classification of the whole sentence as E would be wrong, although the number of words classified as ¬E and E, respectively is equal. Thus it is difficult to choose the best method for labeling and classifying whole sentences.

The results shows that it is possible to classify *anger* versus *non anger* with prosodic and lexical features alone. But for the integration of the recognition of emotion in complex speech dialogue systems the classification rates have to be increased. To increase the classification rates other knowlegde sources have to be integrated. These other knowledge sources could be the detection of repetitions and reformulations, the consideration of swear words and dialogue acts as well as mimic (if possible). The architecture of such a module, called MoUSE (**M**onitoring **o**f **U**ser **S**tate [especially of] **E**motion), is described in [3].

5. CONCLUDING REMARKS

In this paper we showed, that the classification of emotion (here = anger) and a neutral state in speech with prosodic features plus linguistic flags alone is possible, even for the more realistic data of the WoZ experiments. In contrast to the very good classification results for acted data, 86% for a test set with unknown speakers as achieved in [8], the classification performance goes down to 72% for known speaker and 66% for unknown speakers. Furthermore we could show that the recognition of anger versus neutral is very speaker dependent. A reason for these results is, that in acted speech the speakers can only express their emotional behavior using prosodic clues. In the WoZ experiments, the speakers have in addition to prosody some other facilities to express their emotion, like using swear words or meta-language. Furthermore the goal of the speakers themselves is to arrange some appointments and if this does not work they often use other linguistic strategies like reformulation or simple repetition without changing the prosody. In the experiments with actors, the task is to *sound* angry or glad, while in the WoZ experiments the task is to *arrange* some appointments, with or without prosodically marked anger. Another point is that the label method used and the classification of every word itself combined with the computation of the costs with eqn.(1) to classify whole utterances as angry or not is not the best one. We believe that the main problem in the task of classification of emotion, especially of anger, in speech is the need of realistic speech data, such as angry people in real situations. WoZ experiments are one step closer to real life scenarios than acted speech data but they still are not real life.

6. REFERENCES

1. N. Amir and S. Ron. Towards an automatic classification of emotions in speech. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, Sydney, Australia, 1998.
2. A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The prosody module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, page (to appear). Springer, New York, Berlin, 2000.
3. A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately seeking emotions: Actors, wizards, and human beings. In *Proc. ISCA Workshop on Speech and Emotion*, page (to appear), Northern Ireland, September 2000.
4. A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. .Fischer. The recognition of emotion. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, page (to appear). Springer, New York, Berlin, 2000.
5. A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 519–522, Budapest, Hungary, 1999.
6. F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, Philadelphia, USA, 1996.
7. K. Fischer. Annotating emotional language data. Verbmobil Report 236, 1999.
8. R. Huber, E. Nöth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann. You BEEP Machine — Emotion in Automatic Speech Understanding Systems. In *Proc. Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, pages 223–228, Brno, 1998. Masaryk University.
9. R. Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
10. Y. Li and Y. Zhao. Recognizing emotions in speech using short-term and long-term features. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, Sydney, Austrailien, 1998.
11. I. Murray and J. Arnott. Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. In *Journal of the Acoustic Society of America*, volume 2, pages 1097–1108, 1993.
12. I. R. Murray and J. L. Arnott. Synthetizing emotions in speech: Is it time to get excited. In *Proc. Int. Conf. on Spoken Language Processing*, pages 1816–1819, Philadelphia, USA, 1996.
13. K. R. Scherer. Emotion expression in speech and music. In J. Sundberg, L. Nord, and R. Carlson, editors, *Music, Language, Speech, and Brain.* Wenner-Gren Center International Symposium Series Macmillan, London, 1991.
14. A. Zell. *Simulation neuronaler Netze.* Addison Wesley Longmann Verlag GmbH, München, 1997.