You BEEP Machine - Emotion in Automatic Speech Understanding Systems

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

University of Erlangen-Nuremberg,
Chair for Pattern Recognition,
Martensstr. 3, D-91058 Erlangen,
huber@informatik.uni-erlangen.de,
WWW home page: http://www5.informatik.uni-erlangen.de

Abstract. In this paper we report on first experiments for the detection of emotion and the use of this information in a complex speech understanding system like VERBMOBIL. We do not look at lexical information like swear words but rather try to find emotional utterances with the use of acoustic prosodic cues. We only want to classify angry versus neutral speaking style. 20 speakers were asked to produce 50 neutral and 50 angry utterances. With this data set we created one training set and two test sets. One test set with seen speakers, but new turns, the other with unseen speakers, but seen turns. Each word of the emotional utterances was labeled as belonging to the class "emotional", each word in the neutral utterances as belonging to the class "neutral". For each word 276 prosodic features were calculated and multi layer perceptrons were trained for the two classes. We achieved a precision of 87% and a recall of 92% for the one test set and 94% respectively 84% for the other (precision respectively recall), when classifying turns as being either emotionally or neutral.

1 Introduction

Just like people kick soda vending machines, when these don't work, it is expected that users will get mad and angry at speech understanding systems, when a dialogue with such a system goes wrong. Especially in the scenario of call-center applications, it is important to detect such a situation if one does not want to loose a potential customer for ever. After the detection of such a communicative cule-de-sac, appropriate steps like referring the customer to a human operator or starting a clarification dialogue have to be taken.

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.

Even though there are many different emotions in human languages like sadness, joy or fear we are only interested in the distinction between anger and normal speaking. Other emotions will most probably not be relevant for the application of the emotion detector in speech understanding systems like the VERBMOBIL system [1]. Besides, it is more difficult to tell apart emotions like joy and anger [7].

VERBMOBIL is a system for speech-to-speech translation between the three languages German, English and Japanese. There are three application domains for the system, business appointment scheduling, travel planing and computer support talk line. For example in the domain of business appointment scheduling, a German and a Japanese want to fix a date. Everybody speaks in his own language and the system translates the spontaneous speech into the language of the partner and produces synthesized speech.

The psychological aspects of emotion and the correlation between acoustic prosodic cues and emotion are the interest of several research groups, c.f. [7], [2], but to our knowledge this is the first attempt to include emotion detection into an end-to-end speech understanding system to improve the dialogue between the user and the system. In the next section we want to show how to detect emotion in speech signals.

2 The Computation of Prosodic Features

Emotion can be expressed in at least two verbal means (in addition to non verbal cues like body language).

First by the lexical information of some words of an utterance, e.g. swear words like *bullshit*. Second by acoustic prosodic cues like strong changes of the loudness and/or the fundamental frequency (F_0) of the speech signal and changes of the duration values. All these changes do not have to be only in one utterance. They can also occur in relation to earlier utterances of a dialogue.

Furthermore emotion can be expressed by a combination of these parameters. We currently concentrate on the use of acoustic prosodic cues to find emotional utterances and do not use the lexical information of words.

Input to the emotion detector is the word hypotheses graph and the speech signal. Output is a prosodically scored word hypotheses graph [5], i.e., to each of the word hypotheses a probability is attached for it to be emotionally spoken. The computation of prosodic information is described in more detail in [4, 3], where we show how acoustic prosodic features are used for classification of phrase boundaries and phrasal accents c.f. [6].

Based on the speech signal, the F_0 and loudness contours are computed. Then for each of the word hypotheses, a time–alignment of the corresponding phonemes according to the standard pronunciation is performed. This also results in a segmentation of the speech signal into syllable segments given a specific word hypothesis. For the computation of prosodic features for each word hypothesis pointers to the optimal predecessor/successor are established using Viterbi search. Then, for each word hypothesis the following types of features are computed based on the surrounding context (±2 words as well as ±2 syllables and syllable nuclei with respect to the word final syllable): the relative duration [8]; features describing F_0 and energy contours like regression coefficients, minima, maxima, and their relative positions; the length of the pause (if any) after and before the word; the speaking rate; flags indicating word finality and lexical word accent. For an evaluation and a more detailed description of the different types of features cf. [4, 3]. Altogether 276 acoustic prosodic features are calculated.

3 Experiments and Results

20 speakers, seven female and 13 male, were asked to produce 50 neutral and 50 angry utterances in German. All speakers produced the same utterances (278 different words, 13740 tokens, 139 minutes of speech). We took three of the speakers, one female and two male, and five emotional and five neutral utterances of each of the remaining speakers as two test sets and the rest of the utterances as training set (1530 utterances as training, 300 respectively 170 utterances as test sets).

We decided to classify between emotional and neutral on the word level first for theoretical and practical reasons: The theoretical reason is that even in emotionally spoken utterances some words might be spoken as neutral. The practical is that we have almost an order of magnitude more training utterances on the word level vs. the turn level (≈10000 vs. ≈1500).

There are two ways to label the words. First one can listen to every utterance and label each word that belongs to the class *emotional* with **E**. The words that belong to the class *neutral* get the label **X**, e.g.

ich (X) habe (X) gesagt (X) Montag (E) um (E) sechs (E)
I (X) have (X) said (X) Monday (E) at (E) six (E).

This method is very time-consuming and we are not really interested in which words of an utterance are emotionally spoken. Rather we are interested to find those utterances which were spoken when the speaker was angry, no matter which words were spoken with emotion.

The second way is to label each word in the emotional utterances as belonging to the class *emotional* and each word in the neutral utterances as belonging to the class *neutral*, e.g.

ich (E) habe (E) gesagt (E) Montag (E) um (E) sechs (E)
I (E) have (E) said (E) Monday (E) at (E) six (E).

This is a very primitive method to label a data set, but the process does not need much time. We used the second method for labeling. Notice that this method of labeling still allows to incorporate the above mentioned fact that some words of an emotional utterance might be spoken neutral if we apply a bootstrap training algorithm: First we train a classifier with the coarse labeling (all word of an emotional utterance are labeled as emotional). In the second step we retrain the classifier using only those training samples, which are classified correctly by the first classifier.

Multi layer perceptrons (MLP) were trained for the two classes. For the test each word was assigned a probability for the classes neutral and angry. We used two ways to classify an utterance as angry or neutral. First we looked in every single utterance at the percentage of words that are classified as emotional. Then we defined thresholds to classify an utterance as emotional or neutral. Table 1 shows recognition results for the two test sets and different thresholds (> 50 means that more than 50% of the words have to be classified as emotional for the utterance to be classified as emotional).

Table 1. Recognition rates for two thresholds and the two test sets.

			recognized as			
			seen speaker		new speaker	
sp	threshold		emotional	neutral	emotional	neutral
o	$> 50\%$	emotional	82%	18%	79%	21%
k		neutral	7%	93%	7%	93%
e	$\geq 50\%$	emotional	89%	11%	95%	5%
n		neutral	12%	88%	22%	78%

Another way to classify an utterance as emotional or neutral is to interpret the probabilities of each word for belonging to the classes neutral or emotional as statistically independent. With eqn. (1) one can calculate the probability of n statistical independent events.

$$P(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} P(Y_i) \tag{1}$$

To avoid very small numbers we calculate instead of $P(Y_1, Y_2, \ldots, Y_n)$ the costs $C(Y_1, Y_2, \ldots, Y_n)$ with eqn. (2).

$$C(Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^{n} -\log(P(Y_i)) \tag{2}$$

With eqn. (2) we can get now two costs for each utterance with n words, $C(E_1, E_2, \ldots, E_n)$ and $C(X_1, X_2, \ldots, X_n)$, where $C(E_1, E_2, \ldots, E_n)$ is the cost that each word of the utterance is emotional and $C(X_1, X_2, \ldots, X_n)$ that each word is neutral.

If $C(E_1, E_2, \ldots, E_n) \leq C(X_1, X_2, \ldots, X_n)$ is true, we classify the utterance as emotional otherwise as neutral. Table 2 shows recognition results with the two test sets.

For the test set with the seen speakers, precision is 87% and recall is 92%. For the test set with the new speakers, precision is 94% and recall is 84%. The

Table 2. Recognition rates by using the logarithmic evaluation for the two test sets.

		recognized as			
		seen speaker		new speaker	
sp		emotional	neutral	emotional	neutral
ok	emotional	92%	8%	84%	16%
en	neutral	13%	87%	5%	95%

experiments show that emotion can be detected with very high accuracy. Figure 1 gives an example utterance (emotional) with the probabilities of each word, that the words belongs to the class emotional. The dashed line is the probability 0.50. The cost $C(E)$, that the utterance is spoken with an angry speaking style is 6.2, the cost $C(X)$ for a neutral speaking style is 13.5. Utterance is classified correct by emotional.

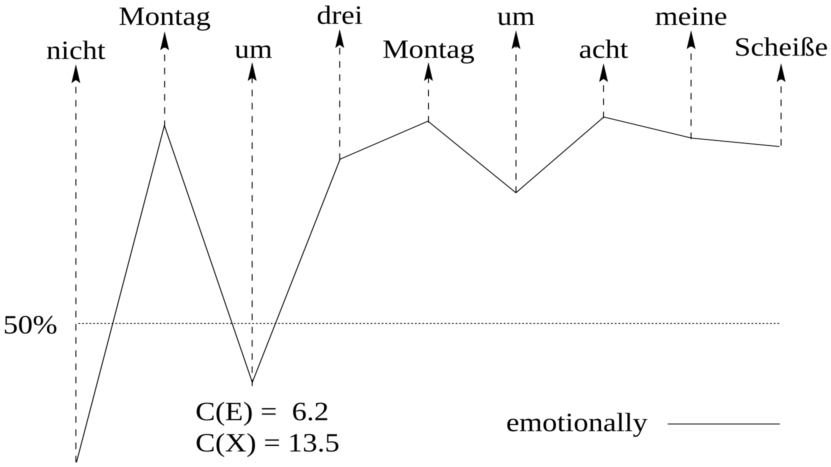

Fig. 1. An example utterance with the probabilities for emotion for every word

4 Conclusion and Future Works

Emotion has not received much attention in the context of automatic speech understanding. When going from the laboratory to real life applications we expect this to change. It is important to know, if a customer is angry, and to react appropriately.

Here we looked at the prosodic marking of anger. Of course the lexical filling of certain words like swear words has to be considered as well. On a database of neutral and angry utterances, where the anger was simulated, we show that the emotional state of the utterances can be predicted with high accuracy using only prosodic features. We achieved a precision of 94% and a recall of 84% on a test set with unseen speakers.

Our results can only be looked upon as preliminary. Many open questions remain.

1. We used a "brute force" approach by calculating as many features as possible. Therefore we currently conduct feature selection experiments in order to reduce the number of features and to find out which prosodic parameter is influenced the most by a change of a speakers emotional state.
2. Our labeling was very coarse. We currently retrain the emotion classifier in a bootstrap procedure as described above.
3. We only worked with simulated anger. Currently Wizard-Of-Oz experiments are designed in the VERBMOBIL project which have the aim of provoking emotionally charged utterances. It remains to be seen to what extent the results for the simulated data are still valid for the real data.

References

1. T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.
2. R. W. Frick. Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin*, 97:419–429, 1985.
3. A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Classification of Boundaries and Accents in Spontaneous Speech. In R. Kuhn, editor, *Proc. of the 3rd CRIM / FORWISS Workshop*, pages 104–113, Montreal, 1996.
4. Andreas Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
5. R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic Scoring of Word Hypotheses Graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.
6. Ralf Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
7. Bernd Tischer. *Die vokale Kommunikation von Gefühlen*, volume 18 of *Fortschritte der psychologischen Forschung.* Psychologie Verlags Union, Weinheim, 1993.
8. C.W. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Trans. on Speech and Audio Processing*, 2(3):469–481, 1994.